Online Appendices to:

Undergraduate Internships, Degree Completion, and the Matthew Effect

Samuel Neylon & Paul Attewell. *Journal of Postsecondary Student Success (JPSS)*, January 2026.

https://doi.org/10.33009/fsop_jpss139836

Online Appendix

Online Appendix A: Descriptive Statistics

Table A1

Descriptive Statistics

	Associate's		Baccalaureate	
	N	*% of Total*	*N*	*% of Total*
Predictors				
Internship				
No	74,668	99.32%	60,694	96.39%
Yes	514	0.68%	2,272	3.61%
Race-Gender				
White-M	7,700	10.24%	10,368	16.47%
White-F	8,854	11.78%	10,984	17.44%
Black-M	9,696	12.90%	4,175	6.63%
Black-F	13,644	18.15%	6,595	10.47%
Hispanic-M	9,844	13.09%	6,897	10.95%
Hispanic-F	12,556	16.70%	10,144	16.11%
Asian-M	6,573	8.74%	6,904	10.96%
Asian-F	6,315	8.40%	6,899	10.96%
College				
College 1	33,289	44.28%	-	-
College 2	-	-	13,061	20.74%
College 3	-	-	12,925	20.53%
College 4	18,527	24.64%	-	-
College 5	-	-	6,845	10.87%
College 6	3,639	4.84%	1,633	2.59%
College 7	13,854	18.43%	3,215	5.11%
College 8	-	-	14,958	23.76%
College 9	5,873	7.81%	10,329	16.40%
Year of Entry				
2008	6,885	9.16%	4,559	7.24%
2009	8,230	10.95%	6,808	10.81%
2010	8,706	11.58%	7,517	11.94%
2011	10,269	13.66%	8,299	13.18%
2012	10,220	13.59%	8,798	13.97%
2013	10,068	13.39%	9,036	14.35%
2014	10,565	14.05%	9,144	14.52%
2015	10,239	13.62%	8,805	13.98%
Entry Semester				
Fall	55,290	73.54%	54,551	86.64%

Spring	19,892	26.46%	8,415	13.36%
Low-Income Student				
No	41,822	55.63%	40,734	64.69%
Yes	33,360	44.37%	22,232	35.31%
Age at Entry				
17-19	40,266	53.56%	41,152	65.36%
19-21	20,586	27.38%	12,784	20.30%
21-25	14,330	19.06%	9,030	14.34%
Credits after 1st Year				
1-15	25,612	34.07%	3,183	5.06%
16-30	40,616	54.02%	33,263	52.83%
>=31	8,954	11.91%	26,520	42.12%
GPA after 1st Year				
2-2.5	20,231	26.91%	10,918	17.34%
2.5-3	22,023	29.29%	17,269	27.43%
3-3.5	19,651	26.14%	19,546	31.04%
3.5-4	13,277	17.66%	15,233	24.19%
Wages in 1st Year				
None	32,644	43.42%	27,725	44.03%
0 - 2K	9,587	12.75%	9,078	14.42%
2K - 10K	17,765	23.63%	15,709	24.95%
> 10K	15,186	20.20%	10,454	16.60%
Proficiency in Math AND Literacy				
Yes	22,508	29.94%	-	-
No	52,674	70.06%	-	-
Outcomes				
Degree within 16 sems?				
No	36,730	48.85%	20,454	32.48%
Yes	38,452	51.15%	42,512	67.52%
Dropout within 16 sems?				
No	43,906	58.40%	45,858	72.83%
Yes	31,276	41.60%	17,108	27.17%
	Mean	*SD*	*Mean*	*SD*
Credits after 16 sems	72.76	44.03	106.03	36.87
N	*75,182*		*62,966*	

Online Appendix B: Predicting Internship Participation

Table B1

Predicting Internship Participation of Associate's Students

	Odds Ratios	
Predictor	Race-Gender	Full Model
Race-Gender		
White M	(Reference)	(Reference)
White F	1.862*	1.466
Black M	2.497***	2.636***
Black F	3.752***	3.339***
Hispanic M	1.530	1.435
Hispanic F	2.663***	2.208**
Asian M	3.014***	2.076**
Asian F	3.669***	2.186**
College		
College 1		(Reference)
College 4		0.347***
College 6		1.681**
College 7		1.205
College 9		1.080
Year of Entry		
2008		0.749
2009		(Reference)
2010		1.212
2011		3.834***
2012		9.072***
2013		9.394***
2014		6.663***
2015		4.600***
2016		3.438**
Low-Income Student		
No		(Reference)
Yes		1.590***
Entry Semester		
Fall		(Reference)
Spring		0.698**
Age at Entry		
17-19		(Reference)
19-21		1.116

21-25		0.974
Credits after 1st Year		
1-15		0.398***
16-30		(Reference)
>=31		1.727***
GPA after 1st Year		
2-2.5		0.221***
2.5-3		0.552***
3-3.5		(Reference)
3.5-4		1.578***
Wages in 1st Year		
None		(Reference)
0 - 2K		1.576***
2K - 10K		1.294*
> 10K		0.592***
Proficiency in Math AND Literacy		
Yes		(Reference)
No		0.887
N	*84,782*	*84,782*

*p<.05. **p<.01. ***p<.001.

Table B2

Predicting Internship Participation of Baccalaureate Students

	Odds Ratios	
Predictor	Race-Gender	Full Model
Race-Gender		
White M	(Reference)	(Reference)
White F	1.570***	1.449***
Black M	3.297***	2.712***
Black F	4.539***	3.816***
Hispanic M	2.036***	1.651***
Hispanic F	3.039***	2.409***
Asian M	1.888***	1.549***
Asian F	3.468***	2.727***
College		
College 2		0.505***
College 3		1.454***
College 5		1.258**
College 6		1.953***
College 7		2.989***
College 8		(Reference)
College 9		1.002
Year of Entry		
2008		0.505**
2009		(Reference)
2010		1.655***
2011		2.493***
2012		3.059***
2013		3.093***
2014		2.268***
2015		1.773***
2016		0.761*
Entry Semester		
Fall		(Reference)
Spring		0.937
Low-Income Student		
No		(Reference)
Yes		1.681***
Age at Entry		
17-19		(Reference)

19-21		0.895
21-25		0.562***
Credits after 1st Year		
1-15		1.127
16-30		(Reference)
>=31		0.969
GPA after 1st Year		
2-2.5		0.422***
2.5-3		0.677***
3-3.5		(Reference)
3.5-4		1.396***
Wages in 1st Year		
None		(Reference)
0 - 2K		1.153*
2K - 10K		1.052
> 10K		0.616***
N	*71,521*	*71,521*

*p<.05. **p<.01. ***p<.001.

Online Appendix C: Conditional Average Treatment Effects (CATTs)

Table C1

Bayesian Estimates of Conditional Average Treatment Effect on the Treated (CATT) by Race-Gender Groups

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
Race-Gender Group	**CATT** [95% Cred. Int.]	**CATT** [95% Cred. Int.]	**CATT** [95% Cred. Int.]
Associate's students			
Black-M	0.199 [0.148, 0.259]	-0.176 [-0.222, -0.135]	19.827 [15.805, 25.404]
Hispanic-M	0.188 [0.136, 0.236]	-0.175 [-0.220, -0.127]	18.300 [10.763, 23.095]
White-M	0.168 [0.119, 0.210]	-0.158 [-0.194, -0.114]	18.610 [13.484, 24.180]
Asian-M	0.160 [0.100, 0.202]	-0.143 [-0.176, -0.096]	18.791 [13.737, 23.009]
Black-F	0.163 [0.122, 0.201]	-0.141 [-0.170, -0.108]	18.408 [14.798, 22.131]
Hispanic-F	0.148 [0.114, 0.181]	-0.127 [-0.153, -0.096]	17.505 [13.548, 21.537]
White-F	0.138 [0.086, 0.173]	-0.117 [-0.144, -0.076]	18.092 [13.650, 22.769]
Asian-F	0.113 [0.079, 0.139]	-0.098 [-0.119, -0.069]	14.228 [5.441, 20.027]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Black-M	0.238 [0.207, 0.265]	-0.195 [-0.223, -0.169]	21.315 [18.838, 26.164]
Hispanic-M	0.231 [0.202, 0.275]	-0.181 [-0.215, -0.158]	18.186 [16.399, 21.050]
White-M	0.154 [0.102, 0.181]	-0.144 [-0.164, -0.123]	17.967 [15.515, 21.144]
Asian-M	0.180 [0.147, 0.203]	-0.141 [-0.159, -0.110]	17.738 [14.781, 19.707]
Black-F	0.181 [0.157, 0.202]	-0.142 [-0.159, -0.124]	18.045 [16.307, 19.709]
Hispanic-F	0.175 [0.155, 0.203]	-0.134 [-0.154, -0.119]	16.570 [14.797, 18.357]

White-F	0.120	-0.097	16.897
	[0.089, 0.139]	[-0.112, -0.068]	[14.048, 18.923]
Asian-F	0.103	-0.089	15.309
	[0.066, 0.135]	[-0.107, -0.054]	[12.582, 17.176]
N	*54,949*	*62,937*	*62,621*

Note. Conditional Average Treatment Effect on the Treated (CATT) is a Bayesian estimate of the treatment effect, conditional on all the covariates. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of an effect. Degree and Dropout CATTs represent percentage point changes, e.g., 0.199 means the treated group (interns) had a 19.9 percentage points greater probability of the outcome, relative to similar students in the untreated group.

Table C2

Bayesian Estimates of Conditional Average Treatment Effect on the Treated (CATT) by Gender Groups

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	CATT	**CATT**	**CATT**
Gender Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
Male	0.181	-0.164	19.027
	[0.139, 0.222]	[-0.197, -0.128]	[15.361, 22.612]
Female	0.147	-0.126	17.370
	[0.115, 0.176]	[-0.150, -0.099]	[14.076, 20.691]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Male	0.204	-0.167	18.848
	[0.184, 0.224]	[-0.184, -0.149]	[17.240, 20.561]
Female	0.151	-0.120	16.757
	[0.136, 0.166]	[-0.131, -0.107]	[15.319, 18.196]
N	*54,949*	*62,937*	*62,621*

Note. Conditional Average Treatment Effect on the Treated (CATT) is a Bayesian estimate of the treatment effect, conditional on all the covariates. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of an effect. Degree and Dropout CATTs represent percentage point changes, e.g., 0.181 means the treated group (interns) had an 18.1 percentage points greater probability of the outcome, relative to similar students in the untreated group.

Table C3

Bayesian Estimates of Conditional Average Treatment Effect on the Treated (CATT) by Income Groups

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	CATT	**CATT**	**CATT**
Income Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
Not Low-Income	0.173	-0.162	20.862
	[0.126, 0.219]	[-0.205, -0.122]	[15.151, 25.981]
Low-Income	0.151	-0.126	16.215
	[0.115, 0.184]	[-0.151, -0.095]	[12.108, 20.389]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Not Low-Income	0.159	-0.138	19.848
	[0.142, 0.177]	[-0.153, -0.123]	[17.910, 21.921]
Low-Income	0.178	-0.134	15.474
	[0.159, 0.195]	[-0.147, -0.119]	[13.611, 17.341]
N	*54,949*	*62,937*	*62,621*

Note. Conditional Average Treatment Effect on the Treated (CATT) is a Bayesian estimate of the treatment effect, conditional on all the covariates. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of an effect. Degree and Dropout CATTs represent percentage point changes, e.g., 0.173 means the treated group (interns) had a 17.3 percentage points greater probability of the outcome, relative to similar students in the untreated group.

Table C4

Bayesian Estimates of Conditional Average Treatment Effect on the Treated (CATT) by GPA Groups

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	CATT	**CATT**	**CATT**
GPA Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
2-2.5	0.217	-0.200	20.645
	[0.159, 0.277]	[-0.252, -0.152]	[16.705, 25.351]
2.5-3	0.200	-0.180	19.063
	[0.152, 0.249]	[-0.222, -0.140]	[15.564, 23.006]
3-3.5	0.154	-0.138	17.582
	[0.119, 0.187]	[-0.166, -0.108]	[14.127, 21.062]
3.5-4	0.124	-0.102	17.051
	[0.096, 0.150]	[-0.123, -0.077]	[13.085, 20.617]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
2-2.5	0.315	-0.246	31.010
	[0.273, 0.364]	[-0.286, -0.214]	[25.054, 35.903]
2.5-3	0.235	-0.184	21.363
	[0.204, 0.266]	[-0.208, -0.163]	[17.584, 24.975]
3-3.5	0.161	-0.129	15.222
	[0.135, 0.182]	[-0.144, -0.113]	[13.340, 17.516]
3.5-4	0.082	-0.079	13.503
	[0.058, 0.104]	[-0.092, -0.062]	[11.335, 15.342]
N	*54,949*	*62,937*	*62,621*

Note. Conditional Average Treatment Effect on the Treated (CATT) is a Bayesian estimate of the treatment effect, conditional on all the covariates. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of an effect. Degree and Dropout CATTs represent percentage point changes, e.g., 0.217 means the treated group (interns) had a 21.7 percentage points greater probability of the outcome, relative to similar students in the untreated group.

Table C5

Bayesian Estimates of Conditional Average Treatment Effect on the Treated (CATT) by Probability of Completion Groups

		Degree by 16 semesters?	Dropout by 16 semesters?
		CATT	**CATT**
Probability of Completion		[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
	Low	0.286	-0.249
		[0.200, 0.372]	[-0.323, -0.181]
	Med	0.219	-0.187
		[0.171, 0.267]	[-0.224, -0.147]
	High	0.115	-0.099
		[0.085, 0.143]	[-0.118, -0.076]
	N	*38,973*	*46,852*
Baccalaureate students			
	Low	0.332	-0.274
		[0.300, 0.367]	[-0.300, -0.245]
	Med	0.192	-0.152
		[0.174, 0.210]	[-0.166, -0.137]
	High	0.074	-0.064
		[0.059, 0.087]	[-0.075, -0.052]
	N	*54,949*	*62,937*

Note. Conditional Average Treatment Effect on the Treated (CATT) is a Bayesian estimate of the treatment effect, conditional on all the covariates. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of an effect. Degree and Dropout CATTs represent percentage point changes, e.g., 0.286 means the treated group (interns) had a 28.6 percentage points greater probability of the outcome, relative to similar students in the untreated group.

Online Appendix D: Conditional Average Treatment Effect (CATT) Contrasts

Table D1

Grand Mean Contrasts of Bayesian Conditional Average Treatment Effects (CATTs) by Race-Gender Group

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	Grand Mean Contrast	**Grand Mean Contrast**	**Grand Mean Contrast**
Race-Gender Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
Black-M	0.041	-0.037	1.867
	[0.017, 0.088]	[-0.071, -0.020]	[-0.168, 6.981]
Hispanic-M	0.029	-0.036	0.340
	[-0.002, 0.058]	[-0.063, -0.003]	[-6.686, 3.573]
White-M	0.009	-0.018	0.650
	[-0.029, 0.036]	[-0.042, 0.012]	[-3.342, 5.433]
Asian-M	0.002	-0.004	0.831
	[-0.057, 0.025]	[-0.022, 0.036]	[-3.899, 3.748]
Black-F	0.005	-0.001	0.448
	[-0.021, 0.019]	[-0.013, 0.013]	[-1.278, 2.360]
Hispanic-F	-0.010	0.013	-0.455
	[-0.025, 0.006]	[0.001, 0.032]	[-3.338, 1.750]
White-F	-0.021	0.023	0.132
	[-0.069, 0.001]	[0.005, 0.059]	[-2.769, 3.231]
Asian-F	-0.045	0.041	-3.732
	[-0.076, -0.027]	[0.027, 0.063]	[-12.133, 0.242]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Black-M	0.069	-0.060	3.851
	[0.045, 0.088]	[-0.082, -0.040]	[2.102, 8.499]
Hispanic-M	0.062	-0.046	0.721
	[0.040, 0.105]	[-0.080, -0.030]	[-0.401, 3.501]
White-M	-0.015	-0.009	0.502
	[-0.068, 0.007]	[-0.023, 0.008]	[-1.761, 3.649]
Asian-M	0.010	-0.005	0.273
	[-0.020, 0.029]	[-0.018, 0.024]	[-2.609, 1.703]
Black-F	0.012	-0.006	0.580
	[-0.006, 0.026]	[-0.017, 0.007]	[-0.418, 1.567]

Hispanic-F	0.006	0.001	-0.894
	[-0.009, 0.034]	[-0.016, 0.010]	[-2.002, 0.263]
White-F	-0.050	0.039	-0.568
	[-0.080, -0.033]	[0.026, 0.068]	[-3.382, 0.889]
Asian-F	-0.066	0.047	-2.155
	[-0.100, -0.038]	[0.031, 0.079]	[-4.629, -0.928]
N	*54,949*	*62,937*	*62,621*

Note. To determine whether there is a difference between the CATTs of race-gender groups and the average treatment effect for all students, we subtract the distribution of CATTs for each race-gender group from the grand mean distribution of CATTs. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of a difference between the group CATTs and the grand mean CATT. Degree and Dropout contrasts represent percentage point differences, e.g., 0.041 means students in this race-gender group had a 4.1 percentage points higher CATT than the grand mean treatment effect for all students.

Table D2

Contrasts of Bayesian Conditional Average Treatment Effects (CATTs) by Gender Group

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	Group Contrast	**Group Contrast**	**Group Contrast**
Gender Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
Female - Male	-0.035 [-0.062, -0.011]	0.038 [0.020, 0.056]	-1.657 [-4.580, 1.022]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Female - Male	-0.053 [-0.072, -0.035]	0.048 [0.033, 0.063]	-2.092 [-3.746, -0.842]
N	*54,949*	*62,937*	*62,621*

Note. To determine whether there is a difference between the CATTs of male and female students, we subtract the distribution of male student CATTs from the distribution for female students. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of a difference between the group CATTs. Degree and Dropout contrasts represent percentage point differences, e.g., -0.035 means female students had a 3.5 percentage points lower CATT than male students.

Table D3

Contrasts of Bayesian Conditional Average Treatment Effects (CATTs) by Income Groups

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	Group Contrast	**Group Contrast**	**Group Contrast**
Income Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
Low-Income - Non-Low-Income	-0.022	0.035	-4.647
	[-0.073, 0.009]	[0.010, 0.084]	[-10.943, 2.711]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
Low-Income - Non-Low-Income	0.019	0.004	-4.375
	[-0.006, 0.036]	[-0.009, 0.022]	[-7.192, -1.797]
N	*54,949*	*62,937*	*62,621*

Note. To determine whether there is a difference between the CATTs of income groups, we subtract the distribution of non-low-income student CATTs from the distribution for low-income students. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of a difference between the group CATTs. Degree and Dropout contrasts represent percentage point differences, e.g., -0.022 means low-income students had a 2.2 percentage points lower CATT than non-low-income students.

Table D4

Grand Mean Contrasts of Bayesian Conditional Average Treatment Effects (CATTs) by GPA Group

	Degree by 16 semesters?	Dropout by 16 semesters?	Credits by 16 semesters?
	Grand Mean Contrast	**Grand Mean Contrast**	**Grand Mean Contrast**
GPA Group	[95% Cred. Int.]	[95% Cred. Int.]	[95% Cred. Int.]
Associate's students			
2-2.5	0.059	-0.061	2.685
	[0.026, 0.105]	[-0.099, -0.037]	[0.722, 6.669]
2.5-3	0.041	-0.041	1.103
	[0.023, 0.073]	[-0.064, -0.026]	[-0.210, 3.725]
3-3.5	-0.005	0.001	-0.378
	[-0.022, 0.004]	[-0.007, 0.013]	[-1.635, 1.175]
3.5-4	-0.035	0.038	-0.909
	[-0.052, -0.018]	[0.025, 0.055]	[-3.690, 0.369]
N	*38,973*	*46,852*	*73,600*
Baccalaureate students			
2-2.5	0.146	-0.110	13.545
	[0.111, 0.192]	[-0.147, -0.087]	[7.833, 18.116]
2.5-3	0.066	-0.048	3.899
	[0.041, 0.091]	[-0.068, -0.037]	[0.417, 7.365]
3-3.5	-0.008	0.007	-2.243
	[-0.030, 0.008]	[-0.002, 0.019]	[-3.430, -0.024]
3.5-4	-0.087	0.056	-3.962
	[-0.107, -0.066]	[0.046, 0.071]	[-5.686, -2.738]
N	*54,949*	*62,937*	*62,621*

Note. To determine whether there is a difference between the CATTs of GPA groups and the average treatment effect for all students, we subtract the distribution of CATTs for each GPA group from the grand mean distribution of CATTs. If the credibility interval does not overlap zero, this means there is a greater than 95% probability of a difference between the group CATTs and the grand mean CATT. Degree and Dropout contrasts represent percentage point differences, e.g., 0.059 means students in this GPA group had a 5.9 percentage points higher CATT than the grand mean treatment effect for all students.

Online Appendix E: Testing BART Effectiveness

Bayesian Additive Regression Trees (BART) has been compared with other sophisticated machine learning methods for causal inference at a series of contests hosted at Atlantic Causal Inference Conferences (Dorie et al., 2019; Thal & Finucane, 2023). Dorie et al. (2019) detail the results of the 2016 contest. The contest was designed to test machine learning methods for causal inference, through testing methods submitted by competitors on a simulated dataset. The simulated data were constructed so that there were complex nonlinear relationships between the confounders and both treatment and outcome. Simulation was an effort to replicate the complex relationships which are often found in real-world applications, and which stymie widely-used but simpler methods for causal inference.

Methods were scored on how biased their estimates of treatment effects were, their measures of uncertainty, and estimation of heterogeneous treatment effects. Methods using BART consistently dominated the top spots (Dorie et al., 2019; Thal & Finucane, 2023). Most of the top methods combine BART with other algorithms, but standalone "vanilla" BART was consistently one of the best approaches on multiple measures. It is striking that vanilla BART performs so well, even though it only models the outcome, and not the treatment assignment. However, for the present paper, we used a method where BART is first used to flexibly model the treatment variable and to estimate propensity scores. These p-scores are then added as a covariate to the BART outcome model. This combined method scored above vanilla BART in the ACIC contests.

Online Appendix F: Sensitivity Analysis

Sensitivity Analysis Methods

All three of the methods used here to control for selection bias make a strong assumption, referred to as conditional independence or ignorability, which is our belief that we have controlled for all relevant confounders. If there are confounding variables—those which are correlated with both outcome and treatment—which we have not measured, this means that we are not controlling for selection bias, and have not properly identified the treatment effect. In our case, there may be unmeasured confounders, such as intrinsic motivation or "grit" (Duckworth et al., 2007) which are correlated with both a student's propensity to pursue the internship program and their probability of college completion. Sensitivity analysis is an important tool to help us understand how robust our findings may be, as it allows us to estimate how large the effect of one of these unobserved confounders would have to be in order to nullify our estimates of the treatment effect.

We have run two types of sensitivity analysis, "Robustness of an Inference to Replacement" (RIR; Frank et al., 2013) and "Impact Threshold of a Confounding Variable" (ITCV; Frank, 2000), using the 'konfound' package in R (Narvaiz et al., 2024). These tests are designed to see how large an unknown, and unobserved, confounder would need to be to invalidate our inferences. RIR measures how many of our observations would have to have, in fact, a null effect (rather than the effect we have estimated for them), to invalidate our estimated treatment effect. The ITCV test takes another approach, estimating how correlated an unmeasured confounder would have to be, with both our treatment and outcome, to invalidate our inference.

In order to estimate the RIR and ITCV, the 'pkonfound' command in the konfound package takes several parameters as input: the estimated treatment effect, the standard error, the

number of observations, and the number of covariates. In order to estimate these parameters from our BART model, we use the population average treatment effect on the treated (PATT) estimates reported in (main text) Table 1. These "population" treatment effects are based on the mean and standard deviation of the posterior predictive distribution of the BART model. This Bayesian standard deviation is generally equivalent to a frequentist standard error (Rockova, 2020). The posterior predictive is used to make a claim about population effects by adding in the additional noise (error) which is estimated during the model-fitting process. We then feed these means and standard errors, along with the number of observations and confounders used in the model, into the 'pkonfound' command. As you can see in Table 1, in our case, additional noise from the posterior predictive leads to larger standard errors for our BART models than the two frequentist models. By taking this larger range of uncertainty into account in sensitivity analysis, we get a more stringent test of our results. While these are frequentist methods applied to Bayesian estimates, the similarity of our Bayesian point estimates and standard errors with our frequentist IPWRA and CEM estimates means that RIR and ITCV findings for all ATT's would be similar.

Sensitivity Analysis Results

Using the 'pkonfound' command from the 'konfound' package in R (Narvaiz et al., 2024), we test the ATT results from our BART model. For the BART results in (main text) Table 1, we used the "population" estimates (PATTs), which come from the mean and standard deviation of the posterior predictive distributions of the model. As you can see in Table 1, in our case, the additional noise used in the PATTs leads to larger standard errors for our BART models than the two frequentist models. In fact, our BART standard errors are close to *twice* as large as those estimated by the IPWRA models. By taking this larger range of uncertainty into account in sensitivity analysis, we get a more stringent test of our results. While we do not conduct

sensitivity analysis on all our heterogeneous treatment effect findings, the robustness of our main ATT results to sensitivity analysis, combined with the measures of uncertainty we have already reported, gives us confidence in these results as well.

First, we report the results from ITCV analysis (Frank, 2000), on our six BART models, in (online appendix) Table F1. ITCV analysis finds how much an unobserved confounder would need to "impact" our treatment effect in order to nullify this effect. ITCV finds how much an unobserved confounder would need to be correlated with the outcome, and with the treatment, and then multiplies these to get the impact threshold an unobserved confounder would have to exceed to nullify our result. In Table F1, we show the correlations with outcome and treatment (which are always the same number) and the impact threshold, for the ITCV from each model. In order to get some intuition for what these results mean, below each ITCV result we report the covariate from each model which had the largest impact, which is calculated by taking this covariate's correlation with the outcome and the treatment, and multiplying them. Coincidentally, the largest covariate impact from all six models was an ordinal categorical variable, either the student's GPA after the first year (binned), or the student's number of credits after the first year (binned). The correlations between these ordinal variables and outcome and treatment were estimated using Pearson's *r*.

As we can see, the impact of these covariates (confounders) is always lower than the ITCV impact threshold. The implication of this is that there would have to be an unobserved confounder with a higher correlation with outcome and treatment than any we have observed. For example, if we were concerned that "grit" (Duckworth et al., 2007) was an unobserved confounder correlated with both academic outcomes and selection into treatment, then the correlation of grit would have to be, for Baccalaureate students, more than *twice* the correlation

of GPA, which is a strong confounder. The ITCV estimates for Associate's students are closer to the largest covariate impacts, but still imply a large, completely unobserved, confounder effect. Because we think this is unlikely, the ITCV test gives us greater confidence that we have controlled for the appropriate confounders and properly identified our result.

Second, we conduct RIR (Frank et al., 2013) tests of our ATT results. These RIR results are presented in Table F1 as percentages. The RIR tests what percentage of observations would need to be replaced by observations which had zero effect from the treatment to nullify our results. It is conceivable that a small percentage of the students in our sample had, in reality, zero effect from the treatment, when we found that they did have an effect from the treatment. However, it is much less likely that a large percentage of students had, in reality, a zero treatment effect. Our results in Table F1 fit the latter hypothetical, as the RIR tests found that between 60.82% (for the model estimating degree outcomes for Associate's students) and 86.73% (for the credits outcomes for Baccalaureate students) would have had to, in reality, had a zero treatment effect for our estimate of the ATT to be nullified. Because it is unlikely that these large proportions of students had zero treatment effects, the RIR test makes us more confident that we have measured all important confounders, and properly estimated the true effect of the internship.

Table F1

Sensitivity Analysis using 'Konfound'

	ITCV			RIR
	Corr. Outcome ×	Corr. Treat =	Impact	% Null
Associate's students				
Degree within 16 terms?	0.125	0.125	0.016	60.82%
Largest Covariate Impact:				
*GPA after 1st year**	*0.206*	*0.051*	*0.011*	
Dropout within 16 terms?	-0.12	0.12	-0.014	61.18%
Largest Covariate Impact:				
*GPA after 1st year**	*-0.193*	*0.053*	*-0.01*	
Credits after 16 terms	0.128	0.128	0.016	69.28%
Largest Covariate Impact:				
*Credits after 1st year**	*0.382*	*0.039*	*0.015*	
Baccalaureate students				
Degree within 16 terms?	0.211	0.211	0.045	84.13%
Largest Covariate Impact:				
*GPA after 1st year**	*0.251*	*0.053*	*0.013*	
Dropout within 16 terms?	-0.195	0.195	-0.038	82.90%
Largest Covariate Impact:				
*GPA after 1st year**	*-0.211*	*0.056*	*-0.012*	
Credits after 16 terms	0.227	0.227	0.051	86.73%
Largest Covariate Impact:				
*GPA after 1st year**	*0.222*	*0.056*	*0.012*	

* Pearson's *r* used to calculate correlation for ordinal categorical variable.

Online Appendix References

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43–68. https://doi.org/10.1214/18-STS667

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147–194. https://doi.org/10.1177/0049124100029002001

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460. https://doi.org/10.3102/0162373713493129

Narvaiz, S., Lin, Q., Rosenberg, J. M., Frank, K. A., Maroulis, S. J., Wang, W., & Xu, R. (2024). konfound: An R sensitivity analysis package to quantify the robustness of causal inferences. *Journal of Open Source Software*, *9*(95), 5779. https://doi.org/10.21105/joss.05779

Rockova, V. (2020). On semi-parametric Inference for BART. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 8137–8146. https://proceedings.mlr.press/v119/rockova20a.html

Thal, D. R. C., & Finucane, M. M. (2023). Causal methods madness: Lessons learned from the 2022 ACIC competition to estimate health policy impacts. *Observational Studies*, *9*(3), 3–27. https://doi.org/10.1353/obs.2023.0023