Advances in Social Science, Education and Humanities Research, volume 653
Proceedings of the 2022 International Conference on Social Sciences and Humanities and Arts (SSHA 2022)

Gender and Age as Factors for Likelihood of Cheating

Yingze Yu[1]*, Chengting Wu[2], Gaofanjie Wang[3], Xinyu Liang[4]

[1]*University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA95064*
[2]*University of North Carolina, 134 E Franklin St, Chapel Hill, NC 27514*
[3]*Kimball Union Academy, 7 Campus Center Dr, Meriden, NH, 03770*
[4]*Shenzhen College of International Education, Shenzhen, Guangdong 518017, China*
[*]*Corresponding author. Email: yyu106@ucsc.edu*

ABSTRACT

This paper establishes models to analyze the effect of age and gender on students' likelihood of cheating on exams. The main model used for analysis is a binary limited dependent model, which is estimated through probit and logit. We assume presence of endogeneity of variables that characterize student's (negative) behavior in school (years in which student failed school and years in which student stopped attending school), possibly due to omitted variable bias, particularly such as the omission of students' intrinsic academic motivation, engagement in studying, and peer influence, which are not available in the data set used by the current paper. We used instruments, such as parents' time spent at work and whether they live with their children, which could be potentially used to infer parents' time spent with their children, to address the endogeneity issue. By comparing the statistical significance of coefficient estimates for age and gender, as well as their estimated average partial effects, we find that gender is a relatively more significant factor on likelihood of cheating on exam than age is, even though the magnitudes of effect of both age and gender on likelihood of cheating are marginal. The deterioration of statistical significance of age and gender is evident from the results in the IV setup, yet regressors characterizing students' family background remain statistically significant.

Keywords*: cheating on exams, gender and age, average partial effects, instrumental variables*

1.INTRODUCTION

The behavior of cheating on exams is not new. The main goal of researchers who studied this behavior in the past and present is to identify sources that affect a student's propensity to cheat on exams. Many papers in the past have identified academic motivation (the desire to have better grades) as an extrinsic factor that would affect the likelihood of cheating. For example, studies conducted by Anderman and Midgley [1], Murdock et al. [2], and Tas and Tekkaya [3] conclude that the desire for better grades is a predictor of cheating. Similarly, Tchouata et al. [4], Olafson et al. [5] concluded that when students pursue academic goals mainly for the purpose of achieving high grades or obtaining a diploma rather than for learning, cheating could become an option.

Another possible factor affecting propensity for cheating investigated by researchers is the students' engagement in studying, as demonstrated by class attendance. Ellahi et al. [6] has shown that poor class attendance and little amount of studying time lead to higher likelihood of cheating. Whitley [7] has shown a relationship between partying and cheating, whereas Patrzek et al. [8] has shown a relationship between procrastinating and cheating.

As noted in many earlier studies, peer influence plays an important role in students' decision to cheat, see Bowers [9], Crittenden et al. [10], Cummings et al. [11], Christensen Hughes and McCabe [12], Ellahi et al. [6], Kisamore et al. [13], Ma et al. [14], McCabe and Trevino [15], Chan et al. [16], Rettinger and Kramer [17], and Whitley [7]. The rationale justifying peer influence on cheating behavior is that students tend to "go with the flow," which could be characterized by whether a student have knowledge of cheating done by others, see Rettinger and Kramer [17].

Given the students' intrinsic desire to cheat and development of technology, accessibility to technologies that facilitate cheating on exams is another factor worth noting when analyzing students' likelihood to cheat, as noted in Michaut [18].

This paper primarily examines the effect of the student's gender and age on the student's likelihood to cheat on a test. The remainder of this paper is organized as follows. Section 2 introduces the model and specifications, Section 3 elaborates on the estimation of average partial effects, Section 4 discusses empirical findings, Section 5 briefly concludes, and the Appendix collects the regression results.

2.MODEL

The type of question examined in this paper falls under the category of binary response model, and similar specification was used in Behrman et al. [19]. Suppose the underlying unobservable "cheating index," y_i^*, takes on the form:

$$y_i^* = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot sex_i + \beta_3 \cdot age_i \times sex_i + \beta_4, Z_i + \epsilon_i, (1)$$

where Z is a vector collecting relevant characteristics of student i, such as student's family background (e.g. parents' time at work, education level, home commodities, family income), behavior in school (e.g. skip school, class participation), academic history (e.g. years of education, stop school). What is indeed observed is whether the student has cheated on the test or not, denoted using y_i (referred to as $copia$ in the regression tables) as an indicator variable, and we further assume that

$$y_i \equiv 1(y_i^* \geqslant K) = \begin{cases} 1, & if \quad cheated \\ 0, & otherwise \end{cases} . \quad (2)$$

Without loss of generality, we normalize by setting $K = 0$. To justify the inclusion of the interaction term between age and gender, we believe potential deviation of gender effect within age due to possible difference in grouping behavior between the two genders, which affect the extent of peer influence in students' propensity to cheat.

Then, the likelihood of cheating takes on the form:

$$P(y_i = 1|age_i,\ sex_i,\ Z_i) = G(\beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot sex_i + \beta_3 \cdot age_i \times sex_i + \beta_4, Z_i).(3)$$

For probit, we assume

$$G(u) = \Phi(u) \equiv \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv, (4)$$

and for logit, we assume

$$G(u) = \Lambda(u) \equiv \frac{\exp(u)}{1+\exp(u)}. \quad (5)$$

2.1.IV Setup for Potentially Endogenous Variables

The other regressors besides age and sex included in W consist of the amount of time parents spend at work ($momwork$ and $dadwork$), parents' education ($momeduc$ and $dadeduc$), whether student live with parents ($momhome$ and $dadhome$), family income ($famincome$), homeownership ($homeown$), access to technologies ($computer$, $cellphone$, and $internet$), and school performance and attendance ($yearsfailed$ and $yearsstop$). Since some of the previously listed regressors could well be endogenous due omitted variables bias, particularly such as the omission of students' intrinsic academic motivation, engagement in studying, and peer influence, which are not available in the data set used by the current paper, others could be introduced as instruments in the IV setup. Note that the direction of bias due to omitted variables is ambiguous.

To obtain the results presented for the IV setup, we suppose the variables $yearsfailed$ and $yearsstop$, which characterize student's performance and attendance in school, to be endogenous. The instruments that we use are whether parents live with their child (student) and how much time parents are away at work. Intuitively, such choice of instruments would satisfy the relevance condition since the amount of time that parents are able to spend with their children (students) affect the students' performance and attendance at school to a great extent. To obtain the coefficient estimates in the IV setup, Newey's two-step estimator is used, see Newey [20].

3.COMPUTATION OF MARGINAL EFFECTS AND ESTIMATION OF AVERAGE PARTIAL EFFECTS

This section presents the estimation of the β's in (3), along with desired quantities such as marginal and average partial effects, and the methods used in this paper can be found in standard text like Wooldridge [21]. The likelihood function for student i can be written as

$$f_i = P(y_i = 1|age_i,\ sex_i,\ Z_i)^{y_i} \cdot (1 - P(y_i = 1|age_i,\ sex_i,\ Z_i))^{1-y_i}.(6)$$

Then, based on (6), the average log-likelihood is constructed:

$$L_N = \frac{1}{N}\sum_{i=1}^{N} \{y_i \cdot \log(P(y_i = 1|age_i,\ sex_i,\ Z_i)) + (1 - y_i) \cdot \log(P(y_i = 1|age_i,\ sex_i,\ Z_i))\}.(7)$$

Next, we use MLE to estimate β_1, β_2, and β_3 as embedded in (3). Moreover, to examine the grade level "within" group effect, (7) is applied to subsets of students within each grade level but including both genders. But before proceeding to subsequent regressions that include more and more variables, we first examine whether the interaction effect between age and sex is significant by fitting the most basic version of (1), i.e. setting the vector Z to be empty, and testing $H_0\!: \beta_3 = 0$.

In terms of interpreting the results, since the models fitted are nonlinear, the numerical values of the β estimates are not enough for us to evaluate the magnitude of effects of regressors as we do in the case of linear models. For nonlinear models, as used in the current paper, we calculate the marginal effects of gender at the mean of other regressors and the average partial effects of both gender and age. The latter is more important since it is a better characterization of the population in contrast with the former's nature as a marginal change at a particular value of the population, and the former is included only to get a feel of the data, in particular for gender.

Let $g(\cdot) \equiv G'(\cdot)$ denote the derivative of $G(\cdot)$ in (3). In general, marginal effect is obtained by taking derivative of the conditional probability in (3) with respect to the regressor of interest, given that the regressor is continuous:

$$\frac{\partial P(y_i = 1|x_i)}{\partial x_{ij}} = g(x_i, \beta)\beta_j,$$

where x_i is the vector collection of all regressors, x_{ij} is the j-th component of x_i, and β is the coefficient vector with j-th component β_j. But since sex_i, the regressor whose marginal effect is of interest, is discrete, its marginal effect at $sex_i = 0$ (female) is

$$G(\beta_0 + \beta_1 \cdot \overline{age} + \beta_2 \cdot 1 + \cdots) - G(\beta_0 + \beta_1 \cdot \overline{age} + \beta_2 \cdot 0 + \cdots), (8)$$

in which all regressors except sex_i are evaluated at their sample means, as of our interest. For calculation, the estimates $\hat{\beta}_j$'s are substituted.

Structurally and based on (1), we think of y_i as a function of x_i and ϵ_i, i.e. $y_i(x_i, \epsilon_i)$. For average partial effect, on the population scale, we first consider the expected value of dependent variable y when the regressors take on a specific value x:

$$E[y(x,\epsilon)] = G(x'\beta) = \int 1\{x'\beta \geqslant 0\} g(\epsilon)\ d\epsilon. (9)$$

Note that the randomness of y solely comes from the randomness of ϵ, and the marginal distribution of ϵ is used in (9). Hence, $E[y(x,\epsilon)]$ is in general different from the conditional expectation $E[y(x,\epsilon)|x]$, in which the conditional density of $\epsilon|x$ would be used in the integral in (9) instead. Given (9), the average partial effect (APE) can be obtained by differentiating (9) with respect to the regressor of interest given that the regressor is continuous, in our case, age:

$$APE_{age} = \frac{\partial E[y(x,\epsilon)]}{\partial\ \ age} = \beta_1 \cdot g(x'\beta). (10)$$

On the other hand, if the regressor of interest is discrete, say sex, then APE is

$$APE_{sex} = E[y(age,\ sex = 1, \dots)] - E[y(age,\ sex = 0, \dots)] (11)$$

$$= G(\beta_0 + \beta_1 \cdot age + \beta_2 + \beta_{4'} \cdot Z + \cdots) - G(\beta_0 + \beta_1 \cdot age + \beta_{4'} \cdot Z + \cdots), (12)$$

where Z is the vector consisting of all other regressors except age and sex. Note that in (12), the interaction term between age and sex is left out assuming that we fail to reject $H_0: \beta_3 = 0$, which we do as shown in Table 5 of the Appendix.

To estimate the APE's in (10) and (12), we use the following estimators:

$$\widehat{APE}_{age} = \hat{\beta}_1 \cdot \frac{1}{n}\sum_{i=1}^{n} g(x_i, \hat{\beta}) (13)$$

$$\widehat{APE}_{sex} = \frac{1}{n}\sum_{i=1}^{n} G(\hat{\beta}_0 + \hat{\beta}_1 \cdot age_i + \hat{\beta}_2 + \hat{\beta}_{4'} \cdot Z_i + \cdots) - G(\hat{\beta}_0 + \hat{\beta}_1 \cdot age_i + \hat{\beta}_{4'} \cdot Z_i + \cdots). (14)$$

4.RESULTS

First, we fit a basic model with regressors consisting of only age and sex for all grade levels, see Tables 1 and 2 for probit and logit, respectively. All coefficient estimates exhibit significance differing from zero at the 5% level. Note that for grade level 9, the sex indicator is automatically dropped due to perfect collinearity with the dependent variable, as detected by the computation software.

Next, on top of the basic model, an interaction term between age and sex is included, see Table 3. Results in Table 3 show that the coefficient estimates for the interaction term are not statistically different from zero at 5% significance level. That is, we conclude from Table 3 failure to reject $H_0: \beta_3 = 0$ for all grade levels (except grade level 9, since variables are dropped due to collinearity). Hence from this point on, we exclude the interaction term from all subsequent model fittings.

Tables 4 through 11 contain model fitting results for extended models, which gradually include four major categories of regressors in the following order: parents' time spent with children (characterized by parents' work hours), family background (characterized by parents' education level, family income, and homeownership), access to technology (computer, cellphone, and internet), and students' (negative) behavior in school (years that student has failed/stopped school), for grade levels 8, 9, 10, and 13.

Table1: Basic model within each grade level

	(1)	(2)	(3)	(4)	(5)	(6)
	copia08_ind	copia09_ind	copia10_ind	copia11_ind	copia12_ind	copia13_ind
main						
sexo08_ind	0.0654^*** ***					
R_AGE08	0.0549^*** ***					
sexo09_ind		0				
R_AGE09		0.0502^*** ***				
sexo10_ind			0.227^*** ***			

R_AGE10			0.0457^*** ***			
sexo11_ind				0.229 ^*** ***		
R_AGE11				0.0949^*** ***		
sexo12_ind					0.103 ^*** ***	
r_age12					0.0636^*** ***	
sexo13_ind						0.116 ^*** ***
r_age13						0.0621^*** ***
Constant	-2.077^*** ***	-2.403^*** ***	-2.546^*** ***	-3.422^*** ***	-2.630^*** ***	-2.835^*** ***
Observations	104056	104554	99037	89678	82410	82118
Adjusted R^2						
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$						

Table2: Basic model within each grade level - logit

	(1)	(2)	(3)	(4)	(5)	(6)
	copia08_ind	copia09_ind	copia10_ind	copia11_ind	copia12_ind	copia13_ind
main						
sexo08_ind	0.135 ^*** ***					
R_AGE08	0.110 ^*** ***					
sexo09_ind		0				
R_AGE09		0.112 ^*** ***				
sexo10_ind			0.530 ^*** ***			
R_AGE10			0.108 ^*** ***			
sexo11_ind				0.568 ^*** ***		
R_AGE11				0.231 ^*** ***		
sexo12_ind					0.227 ^*** ***	
r_age12					0.141 ^*** ***	
sexo13_ind						0.268 ^*** ***
r_age13						0.142 ^*** ***
Constant	-3.791 ^*** ***	-4.641 ^*** ***	-5.053 ^*** ***	-7.234 ^*** ***	-5.121 ^*** ***	-5.660 ^*** ***
Observations	104056	104554	99037	89678	82410	82118
Adjusted R^2						
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$						

Table 3: Basic model with interaction term

	(1)	(2)	(3)	(4)	(5)	(6)
	copia08_ind	copia09_ind	copia10_ind	copia11_ind	copia12_ind	copia13_ind
main						
sexo08_ind	-0.180					
R_AGE08	0.0426 ^*** ***					
sexo_age08	0.0240					
sexo09_ind		0				
R_AGE09		0.0502 ^*** ***				
sexo_age09		0				

sexo10_ind			-0.297			
R_AGE10			0.0275			
sexo_age10			0.0431			
sexo11_ind				-0.163		
R_AGE11				0.0814^*** ***		
sexo_age11				0.0298		
sexo12_ind					0.297	
r_age12					0.0704^*** ***	
sexo_age12					-0.0139	
sexo13_ind						0.845 ^* *
r_age13						0.0859 ^*** ***
sexo_age13						-0.0488
Constant	-1.950 ^*** ***	-2.403 ^*** ***	-2.324^*** ***	-3.243 ^*** ***	-2.724 ^*** ***	-3.190 ^*** ***
Observations	104056	104554	99037	89678	82410	82118
Adjusted R^2						
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$						

Table 4: Extended model for grade level 8

	(1)	(2)	(3)	(4)
	copia08_ind	copia08_ind	copia08_ind	copia08_ind
copia08_ind				
sexo08_ind	0.0705 ^*** ***	0.0657 ^*** ***	0.0649 ^*** ***	0.0697 ^*** ***
R_AGE08	0.0521 ^*** ***	0.00881	0.00693	0.00473
MomHome_Student08	-0.0762 ^** **	-0.0238	-0.0306	-0.0316
DadHome_Student08	0.0334	0.0397 ^* *	0.0451 ^* *	0.0423 ^* *
MomWork08	-0.0383 ^*** ***	-0.00569	-0.00596	-0.00701
DadWork08	0.0312 ^*** ***	0.0170 ^** **	0.0182 ^** **	0.0167 ^** **
MomEduc_Parent08		-0.0389 ^*** ***	-0.0360 ^*** ***	-0.0366 ^*** ***
DadEduc_Parent08		-0.0191 ^*** ***	-0.0144 ^** **	-0.0111 ^* *
FamIncome08		-0.0685 ^*** ***	-0.0609 ^*** ***	-0.0629 ^*** ***
HomeOwn08		-0.0160 ^*** ***	-0.0162 ^*** ***	-0.0181 ^*** ***
Computer_Student08			-0.0602 ^** **	-0.0609 ^** **
CellPhone_Student08			-0.00974	-0.00597
Internet_Student08			-0.0390	-0.0421
YearsFailed08				-0.000185
YearsStop08				0.0572 ^*** ***
Constant	-2.009 ^*** ***	-1.235 ^*** ***	-1.224 ^*** ***	-1.265 ^*** ***
Observations	86223	81072	78211	74751
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 5: Extended model for grade level 8 - logit

	(1)	(2)	(3)	(4)
	copia08_ind	copia08_ind	copia08_ind	copia08_ind
copia08_ind				
sexo08_ind	0.147 ^*** ***	0.137 ^*** ***	0.135 ^*** ***	0.145 ^*** ***
R_AGE08	0.107 ^*** ***	0.0143	0.0108	0.00825
MomHome_Student08	-0.155 ^** **	-0.0470	-0.0616	-0.0645
DadHome_Student08	0.0695	0.0814 ^* *	0.0926 ^* *	0.0870 ^* *
MomWork08	-0.0810 ^*** ***	-0.0111	-0.0118	-0.0138
DadWork08	0.0631 ^*** ***	0.0334 ^** **	0.0360 ^** **	0.0331 ^** **
MomEduc_Parent08		-0.0825 ^*** ***	-0.0765 ^*** ***	-0.0781 ^*** ***
DadEduc_Parent08		-0.0397 ^*** ***	-0.0302 ^** **	-0.0235 ^* *
FamIncome08		-0.143 ^*** ***	-0.127 ^*** ***	-0.130 ^*** ***
HomeOwn08		-0.0327 ^*** ***	-0.0331 ^*** ***	-0.0370 ^*** ***
Computer_Student08			-0.126 ^** **	-0.127 ^** **
CellPhone_Student08			-0.0194	-0.0119
Internet_Student08			-0.0793	-0.0872
YearsFailed08				-0.00390
YearsStop08				0.110 ^*** ***
Constant	-3.676 ^*** ***	-2.048 ^*** ***	-2.029 ^*** ***	-2.120 ^*** ***
Observations	86223	81072	78211	74751
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 6: Extended model for grade level 9

	(1)	(2)	(3)	(4)
	copia09_ind	copia09_ind	copia09_ind	copia09_ind
copia09_ind				
sexo09_ind	0	0	0	0
R_AGE09	0.134 ^* *	0.0712	0.0707	0.131
MomHome_Student09	0.135	0.222	0.226	0.327
DadHome_Student09	0.168	0.210	0.198	0.140
MomWork09	0.0275	0.0230	0.0225	0.0123
DadWork09	0.0187	0.0147	0.0108	0.00803
MomEduc_Parent08		0.0231	0.0167	0.0154
DadEduc_Parent08		-0.0531	-0.0584	-0.0454
FamIncome09		-0.0281	-0.0375	-0.0537
HomeOwn09		0.0563 ^* *	0.0573 ^* *	0.0737 ^** **
Computer_Student09			0.109	0.123
CellPhone_Student09			0.0550	0.101
Internet_Student09			-0.0189	-0.0188

YearsFailed09				-0.136
YearsStop09				-0.131
Constant	-3.560 ^*** ***	-2.946 ^*** ***	-2.936 ^*** ***	-3.429 ^*** ***
Observations	3819	3200	3143	2976
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 7: Extended model for grade level 9 - logit

	(1)	(2)	(3)	(4)
	copia09_ind	copia09_ind	copia09_ind	copia09_ind
copia09_ind				
sexo09_ind	0	0	0	0
R_AGE09	0.282 ^** **	0.162	0.161	0.282
MomHome_Student09	0.309	0.503	0.522	0.790
DadHome_Student09	0.380	0.495	0.471	0.329
MomWork09	0.0634	0.0499	0.0471	0.0269
DadWork09	0.0474	0.0419	0.0348	0.0284
MomEduc_Parent08		0.0544	0.0381	0.0363
DadEduc_Parent08		-0.122	-0.135	-0.103
FamIncome09		-0.0561	-0.0748	-0.111
HomeOwn09		0.119 ^* *	0.123 ^* *	0.157 ^* *
Computer_Student09			0.283	0.317
CellPhone_Student09			0.124	0.232
Internet_Student09			-0.0698	-0.0801
YearsFailed09				-0.270
YearsStop09				-0.271
Constant	-7.061^*** ***	-5.931 ^*** ***	-5.929 ^*** ***	-6.986 ^*** ***
Observations	3819	3200	3143	2976
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 8: Extended model for grade level 10

	(1)	(2)	(3)	(4)
	copia10_ind	copia10_ind	copia10_ind	copia10_ind
copia10_ind				
sexo10_ind	0.343 ^*** ***	0.321 ^*** ***	0.325 ^*** ***	0.314 ^*** ***
R_AGE10	0.0690	0.0111	0.0184	0.0851
MomHome_Student10	-0.00960	-0.0351	0.0524	0.00744
DadHome_Student10	-0.0229	0.00195	0.0154	0.0517
MomWork10	-0.0453	-0.0531	-0.0498	-0.0536
DadWork10	0.0462	0.0592	0.0604	0.0556
MomEduc_Parent08		-0.00651	-0.0109	-0.0136
DadEduc_Parent08		0.0102	0.00881	0.0147

FamIncome10		-0.0567	-0.0626	-0.0699
HomeOwn10		-0.0132	-0.00911	-0.00617
Computer_Student10			0.0616	0.0470
CellPhone_Student10			0.110	0.136
Internet_Student10			-0.0506	-0.0858
YearsFailed10				-0.0921
YearsStop10				-0.199
Constant	-2.839^*** ***	-1.976 ^** *	-2.239 ^** *	-2.699 ^** *
Observations	4066	3372	3339	3200
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 9: Extended model for grade level 10 - logit

	(1)	(2)	(3)	(4)
	copia10_ind	copia10_ind	copia10_ind	copia10_ind
copia10_ind				
sexo10_ind	0.802^*** ***	0.755^*** ***	0.766^*** ***	0.736^*** ***
R_AGE10	0.151	0.0323	0.0490	0.191
MomHome_Student10	0.00397	-0.0729	0.119	0.0125
DadHome_Student10	0.0392	0.0165	0.0449	0.122
MomWork10	-0.110	-0.130	-0.123	-0.131
DadWork10	0.108	0.137 ^* *	0.141 ^* *	0.129
MomEduc_Parent08		-0.0196	-0.0294	-0.0325
DadEduc_Parent08		0.0288	0.0244	0.0405
FamIncome10		-0.125	-0.138	-0.156
HomeOwn10		0.0275	-0.0193	-0.0131
Computer_Student10			0.158	0.116
CellPhone_Student10			0.242	0.290
Internet_Student10			-0.124	-0.193
YearsFailed10				-0.186
YearsStop10				-0.467
Constant	-5.632 ^** **	-3.830	-4.415 ^* *	-5.357 ^* *
Observations	4066	3372	3339	3200
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 10: Extended model for grade level 13

	(1)	(2)	(3)	(4)
	copia13_ind	copia13_ind	copia13_ind	copia13_ind
copia13_ind				
sexo13_ind	0.306 ^*** ***	0.288 ^*** ***	0.286 ^*** ***	0.292 ^*** ***
r_age13	0.0728	0.0321	0.0369	0.0776
MomHome_Student13	-0.0865	-0.0589	-0.0694	-0.0870

DadHome_Student13	0.221	0.284 ^* *	0.285 ^* *	0.273 ^* *
MomWork13	0.0158	0.0256	0.0259	0.0325
DadWork13	0.0557 ^* *	0.0664 ^* *	0.0679 ^* *	0.0680 ^* *
MomEduc_Parent08		-0.00657	-0.00819	-0.0121
DadEduc_Parent08		0.0320	0.0280	0.0467
FamIncome13		0.0774	-0.0783	-0.0835
HomeOwn13		-0.0859 ^* *	-0.0859 ^* *	-0.0826 ^* *
Computer_Student13			0.103	0.0919
CellPhone_Student13			-0.0802	-0.101
Internet_Student13			-0.0458	-0.0625
YearsFailed13				-0.110
YearsStop13				-0.00982
Constant	-3.364 ^*** ***	-2.613 ^* *	-2.617 ^* *	-3.099 ^* *
Observations	4335	3768	3734	3537
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Table 11: Extended model for grade level 13 - logit

	(1)	(2)	(3)	(4)
	copia13_ind	copia13_ind	copia13_ind	copia13_ind
copia13_ind				
sexo13_ind	0.718 ^*** ***	0.681 ^*** ***	0.677 ^*** ***	0.700 ^*** ***
r_age13	0.170	0.0694	0.0791	0.173
MomHome_Student13	-0.180	-0.130	-0.159	-0.176
DadHome_Student13	0.504	0.674 ^* *	0.678 ^* *	0.653 ^* *
MomWork13	0.0357	0.0575	0.0598	0.0759
DadWork13	0.130 ^* *	0.155 ^* *	0.158 ^* *	0.158 ^* *
MomEduc_Parent08		-0.00435	-0.00829	-0.0175
DadEduc_Parent08		0.0650	0.0544	0.0942
FamIncome13		-0.192 ^* *	-0.193 ^* *	-0.205 ^* *
HomeOwn13		-0.214 ^* *	-0.213 ^* *	-0.204 ^* *
Computer_Student13			0.214	0.190
CellPhone_Student13			-0.209	-0.255
Internet_Student13			-0.0771	-0.113
YearsFailed13				-0.280
YearsStop13				-0.00764
Constant	-6.963 ^** **	-5.095 ^* *	-5.058	-6.176 ^* *
Observations	4335	3768	3734	3537
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

Another occurrence worths noting is that for grade level 8, the addition of regressors belonging to the family background category (see second columns of Tables 4 and 5) yields negative and statistically significant coefficient estimates for regressors from that category. Recall that the higher the value of regressors from the family background category, the better off the family is. Hence, this could imply that students whose parents are more educated and students coming from more wealthy families are less likely to cheat on exams. But such negative and statistically significant coefficient estimates for regressors from the family background category only occur within grade level 8. With this regard, we note a significant reduction in sample sizes in model fitting for grade levels 9, 10, and 13 due to missing values, which could adversely affect the statistical significance of the coefficient estimates. Similar reduction in statistical significance also happens for the "years the student stopped attending school" regressor in the student behavior category.

4.1. Magnitude of Marginal Effect and Average Partial Effects

Table 12: Marginal effects of sex (at mean) for grade levels 8, 9, 10, 13

Grade Level	8	9	10	13
Marginal Effect	0.0086353	—	0.0219685	0.0188782

Table 13: Average partial effects of sex and age for grade levels 8, 9, 10, 13

Grade Level	8	9	10	13
APE_{age}	0.0006008	0.0105437	0.006321	0.0054001
APE_{sex}	0.0088559	—	0.0232997	0.0203541

Table 12 contains the marginal effect of sex when $sex = 0$ (female) and all other regressors equal to their sample mean, calculated using equation (8) and $G(\cdot) = \Phi(\cdot)$, the standard normal cdf. Table 13 contains the calculated average partial effects for age and sex, calculated using (13) and (14) and $G(\cdot) = \Phi(\cdot)$. Note that the results on gender partial effects are dropped for grade level 9 due to collinearity, and also that these partial effects are only calculated for (1) with all regressors included (the fourth columns of Tables 4 through 11) but not the IV setup.

The coefficient estimates alone do not give the magnitude of the partial effects. The values in Tables 1 and 2 are probability/likelihood given on a scale from 0 to 1, and indeed the magnitudes of APE of age are small in the sense that a unit increase in age would only result in a tiny fraction increase ($< 1\%$) in the likelihood of cheating on exams, which is coherent with the statistical insignificance of coefficient estimate for the age regressor shown in Tables 6 through 13. On the other hand, the APE of gender being male is a roughly 2% increase in the likelihood of cheating on exams, which is small but significant in comparison to the magnitude of APE of age and is coherent with the statistical significance shown in the regression tables.

Table 14: IV model for grade levels 8,9,10,13

	(1)	(2)	(3)	(4)
	copia08_ind	copia09_ind	copia10_ind	copia13_ind
YearsFailed08	-0.223			
YearsStop08	0.315			
sexo08_ind	0.0632 ^* *			
R_AGE08	0.0699			
MomEduc_Parent08	-0.0373 ^*** ***	-0.0579	0.00718	-0.0120
DadEduc_Parent08	-0.0127	-0.0144	0.0584	0.0627
FamIncome08	-0.0659 ^*** ***			
HomeOwn08	-0.0205 ^** **			
Computer_Student08	-0.0601 ^** **			
CellPhone_Student08	-0.00442			
Internet_Student08	-0.0427			
YearsFailed09		-4.549		
YearsStop09		2.743		
sexo09_ind		0		

R_AGE09		1.622		
FamIncome09		-0.0767		
HomeOwn09		-0.0125		
Computer_Student09		0.129		
CellPhone_Student09		0.0565		
Internet_Student09		0.0333		
YearsFailed10			2.425	
YearsStop10			-2.185	
sexo10_ind			0.522	
R_AGE10			-0.716	
FamIncome10			-0.0544	
HomeOwn10			-0.00963	
Computer_Student10			0.0997	
CellPhone_Student10			0.0294	
Internet_Student10			-0.163	
YearsFailed13				-2.314
YearsStop13				4.243
sexo13_ind				0.308 ^* *
r_age13				0.416
FamIncome13				-0.107
HomeOwn13				-0.121 ^* *
Computer_Student13				0.198
CellPhone_Student13				-0.148
Internet_Student13				-0.139
Constant	-1.905	-17.13	5.925	-9.857
Observations	74751	2976	3200	3537
Adjusted R^2				
^* * $p < 0.05$, ^** ** $p < 0.01$, ^*** *** $p < 0.001$				

4.2. IV Setup Results

In principle, we do not expect that age and gender would intrinsically cause and directly account for whether a person cheat on an exam. Based on Table 14, we see reduction in statistical significance of coefficient estimates for age and gender for all grade levels in the IV setup, in which we use regressors in the “parents’ time spent with children" category as instruments for regressors in the “students’ behavior" category. Yet, the regressors from the “family background" category remain to be negative and statistically significant, in particular for grade level 8 with abundant sample size. This could imply that the variables of interest regarding effects on likelihood of cheating may indeed be regressors from the “family background" category.

5. CONCLUSION

In this empirical study, we attempt to answer to what extent do age and gender affect a student’s likelihood of cheating on exams. We establish a limited dependent variable model with binary response using age and gender as the main regressors predicting the likelihood of cheating, with probit and logit specification for the choice of cdf of the structural error. We introduce four major categories of regressors as extension to the basic model, namely, “parents’ time spent with children," “family background," “access to technology," and “behavior at school." Given these model extensions, we calculate the average partial effects of age and gender based on the fully extended model (all four categories of regressors being included) and examine the magnitude of those average partial effects in conjunction with the statistical significance of coefficient estimates for age and gender from the regression tables. Lastly, to account for endogeneity of regressors from the “behavior at school" category due to possible omitted variable bias, we take regressors from the “parents’ time spent with children" category as instruments for the possible endogenous regressors. Our conclusion is that gender is a relatively more significant factor on likelihood of cheating on exam than age is, even though the magnitudes of effect of both age and gender on likelihood of cheating are marginal. The deterioration of statistical significance of age and gender is evident from the results in the IV setup, yet regressors characterizing students’ family background remain statistically significant with signs indicating that students’ with more educated parents and from wealthier families are less likely to cheat on exams.

REFERENCES

[1] Eric M. Anderman and Carol Midgley. Changes in self-reported academic cheating across the transition from middle school to high school. Contemporary Educational Psychology,29(4):499–517, 2004.

[2] T. B. Murdock, A. D. Miller, and A. Goetzinger. Effects of classroom context on university students’ judgments about cheating: mediating and moderating processes. Social Psychology of Education, 10(2):141–169, 2007.

[3] Y. Tas and C. Tekkaya. Personal and contextual factors associated with students’ cheating in science. Journal of Experimental Education - J EXP EDUC, 78:440–463, 06 2010.

[4] C. Tchouata, M. F. Lamago, and C. R. Singo Njabo. Fraude aux examens et formation des enseignants : le cas de l’ecole normale superieure de yaounde. Form Prof, 22(3):48–62, 2014.

[5] Lori Olafson, Gregory Schraw, Louis Nadelson, Sandie Nadelson, and Nicolas Kehrwald. Exploring the judgment-action gap: College students and academic dishonesty. Ethics and Behavior - ETHICS BEHAV, 23, 01 2012.

[6] Abida Ellahi, Rabia Mushtaq, and Mohammed Bashir Khan. Multi campus investigation of academic dishonesty in higher education of pakistan. 27(6):647–666, 2021/09/16 2013.

[7] Bernard E. Whitley. Factors associated with cheating among college students: A review. Research in Higher Education, 39(3):235–274, 1998.

[8] Justine Patrzek, Sebastian Sattler, Floris van Veen, Carola Grunschel, and Stefan Fries. Investigating the effect of academic procrastination on the frequency and variety of academic misconduct: a panel study. Studies in Higher Education, 40:1–16, 03 2014.

[9] W.J.Bowers. Student dishonesty and its control in college. Bureau of Applied Social Research, Columbia University, New York, 1964.

[10] Victoria L. Crittenden, Richard C. Hanna, and Robert A. Peterson. The cheating culture: A global societal phenomenon. Business Horizons, 52(4):337–346, 2009.

[11] Rhoda Cummings, Cleborne D. Maddux, Steve Harlow, and Lynn Dyas. Academic misconduct in undergraduate teacher education students and its relationship to their principled moral reasoning. Journal of Instructional Psychology, 29:286+, 2021/9/16/ 2002.

[12] Julia Christensen Hughes and Donald Mccabe.Understanding academic misconduct. Canadian Journal of Higher Education, 36, 06 2006.

[13] Jennifer Kisamore, Thomas Stone, and I. Jawahar. Academic integrity: The relationship between individual and situational factors on misconduct contemplations. Journal of Business Ethics, 75:381–394, 02 2007.

[14] Hongyan Ma, Guofang Wan, and Eric Lu. Digital cheating and plagiarism in schools. Theory Into Practice - THEORY PRACT, 47:197–203, 07 2008.

[15] Donald L.McCabe and Linda Klebe Trevino. Individualandcontextualinfluencesonacademic dishonesty: A multicampus investigation. Research in Higher Education, 38(3):379–396, June 1997.

[16] Cynthia Chan, Jamilah Othman, Jeffrey Dsilva, and Zoharah Omar. Influence of neutralization attitude in academic dishonesty among undergraduates. International Education Studies, 7, 05 2014.

[17] David Rettinger and Yair Kramer. Situational and personal causes of student cheating. Research in Higher Education, 50:293–313, 05 2009.

[18] Christophe Michaut. Les nouveaux outils de la tricherie scolaire au lyc´ee. Recherches en ´education, pages 131–142, 06 2013.

[19] J. R. Behrman, S. W. Parker, P. E. Todd, and W. Zhang. Prospering through prospera: Cct impacts on educational attainment and achievement in mexico, 2021.

[20] Whitney K. Newey. Efficient estimation of limited dependent variable models with endogenous explanatory variables. Journal of Econometrics, 36(3):231–250, 1987.

[21] Jeffrey M. Wooldridge. Econometric analysis of cross section and panel data. MIT Press, Cambridge, Mass., 2002.