贫困与性别不平等

——来自中国高考的证据

宗一博 郑新业 韩 奕

目录

附录 I 贫困指标界定..... 1
附录 II 实证策略模型..... 2
附录 III 遗漏变量的稳健性检验..... 3
附录 IV 附表及附图..... 4
参考文献 11

附录 I 贫困指标界定

本文使用的贫困人口的数据来自《贫困人口微观追踪数据库》。该数据库由政府部门于 2014 年建立，目的是为了精准识别并记录国家级贫困县内所有贫困户的详细家庭与个人信息。根据《河南省扶贫对象精准识别及管理办法》，贫困户的筛选标准包括严格的农民人均纯收入标准[①]和统筹考虑“两不愁三保障”因素，经过初选对象、乡镇审核、县级复审的多级识别程序后，符合扶贫标准的贫困户及家庭成员的信息将录入《贫困人口微观追踪数据库》。

若本文样本中的学生记录在《贫困人口微观追踪数据库》，则其被视为贫困学生。本文样本中共有 730 名贫困生，占总样本的 9.7%。这一比例与河南省及全国的贫困发生率相接近，说明本样本具有一定的代表性。

表 I 1 贫困发生率

	样本	样本所在县	某县所在市	河南省	全国
	（1）	（2）	（3）	（4）	（5）
贫困发生率	9.7%	14.9%	10.9%	8.8%	7.2%

注：第（1）列为本文样本的贫困生占比，第（2）-（5）列分别为 2014 年本文样本所在县、市、河南省及全国的贫困发生率。

表 I 2 统计了分年份贫困生占比，逐年贫困生占比在 8.2%至 11.9%之间变化，本样本每年的学生中都存在相似比例的贫困生。其次，对于班级分布，我们的样本涵盖了 148 个班级，其中只有 2 个班级的贫困生数量为 0。这表明除这两个班级外，其他 146 个班级均有贫困生存在。因此，在当控制了年份和班级固定效应以后，贫困生变量在我们的数据中仍然保持足够的变异性，可以用来进行统计估计。

表 I 2 分年份贫困生占比

	2014	2015	2016	2017	2019	2020	2021	2022
贫困生占比	8.2	9.0	9.6	11.9	10.4	7.9	9.7	9.9
（%）	(27.5)	(28.6)	(29.5)	(32.4)	(30.6)	(27.0)	(29.7)	(29.9)
观测值	695	947	1123	1279	1199	1199	1099	955

[①] 根据国家统计局，现行贫困标准是指农村居民每人每年生活水平在 2300 元以下（2010 年不变价）。

附录 II 实证策略模型

假设一个人在高考中的表现用以下方程表示:

$$G_{i,p,g} = \omega_g + \gamma_p + Z^G + Y_i^G + x_g^G + \sigma_p^G + \delta_{p,g}^G + \epsilon_{i,g,p}^G \quad (\text{II}1)$$

这里，G表示高考表现，i表示学生个体,g表示性别,p表示贫困状态。ω_g表示代表在高考和模考两次考试中不会变化的性别特征，γ_p表示代表在中不会变化的与贫困相关的特征，Z^G表示不因性别而异的考试特征（如考试环境等），Y_i^G表示可能对两次考试产生不同影响的个体特征，x_g^G捕捉因性别而异的影响高考的因素，σ_p^G则是因贫困状态不同而影响高考的因素，$\delta_{p,g}^G$为与贫困状态和性别同时相关的因素，$\epsilon_{i,g,p}^G$是误差项。

相应的，学生在模考中的表现（M）可以表示为：

$$M_{i,p,g} = \omega_g + \gamma_p + Z^M + Y_i^M + x_g^M + \sigma_p^M + \delta_{p,g}^M + \epsilon_{i,g,p}^M \quad (\text{II}2)$$

对于这两个方程，为了消除所有学生共同面临的考试特定特征（Z）的影响，我们使用标准化考试成绩$\widetilde{G_{\iota,g,p}}$和$\widetilde{M_{\iota,g,p}}$[②]。将两个方程相减可得：

$$\widetilde{G_{\iota,g,p}} - \widetilde{M_{\iota,g,p}} = \left(\tilde{x}_g^G - \tilde{x}_g^M\right) + \left(\tilde{\sigma}_p^G - \tilde{\sigma}_p^M\right) + \left(\tilde{\delta}_{p,g}^G - \tilde{\delta}_{p,g}^M\right) + \left(\tilde{Y}_g^G - \tilde{Y}_g^M\right) + \left(\tilde{\epsilon}_{i,g}^G - \tilde{\epsilon}_{i,g}^M\right) \quad (\text{II}3)$$

我们使用表示学生i的标准化高考分数减去标准化模考分数的差值（D_i）表示$\left(\widetilde{G_{\iota,g}} - \widetilde{M_{\iota,g}}\right)$，表示女性的虚拟变量$Female_i$来捕捉$\left(\tilde{x}_g^G - \tilde{x}_g^M\right)$，表示贫困状态的虚拟变量$Poor_i$来捕捉$\left(\tilde{x}_g^G - \tilde{x}_g^M\right)$,性别变量与贫困变量的交互项$Female_i \times Poor_i$表示$\left(\tilde{\delta}_{p,g}^G - \tilde{\delta}_{p,g}^M\right)$，并在方程中包括表示理科生的虚拟变量、表示复读生的虚拟变量、表示本地户口的虚拟变量、参加高考时的年龄、班级固定效应来控制$\left(\tilde{Y}_g^G - \tilde{Y}_g^M\right)$。回归方程表示为：

$$D_{it} = \beta_0 + \beta_1 Female_i + \beta_2 Poor_i + \beta_3 Poor_i \times Female_i + x_i\eta + \gamma_t + \alpha_c + \varepsilon_{ict} \quad (\text{II}4)$$

[②] 标准化为第三章中：$\widetilde{G_{\iota pt}} = \frac{G_{ipt} - \overline{G_{pt}}}{\delta_{G_{pt}}}$和$= \frac{M_{ipt} - \overline{M_{pt}}}{\delta_{M_{pt}}}$，文中省略了表示文理科目的$p$和考试年份$t$。

附录 III 遗漏变量的稳健性检验

学生的决策质量可能受到个体及家庭因素的影响，除本文考虑的控制变量外，为考察可能存在的遗漏变量及其对实证结果的影响，本文使用 Oster(2016)提出的方法进行稳健性检验。

Oster(2016)证明，当模型可能存在不可观测的遗漏变量是，可采用估计量$\beta^* = \beta^*(R_{max}, \delta)$获得真实系数的一致估计，该估计量需要设定两个参数，R_{max}和δ。其中，R_{max}为若不可观测的遗漏变量能够被观测并包含在回归估计中，回归方程的最大拟合优度；δ为选择比例（election proportionality），衡量可观测控制变量与核心解释变量的相关关系相比于不可观测的遗漏变量与核心解释变量的相关关系的强弱。根据 Oster(2016)检验现有文献的结果，并提出的建议，本文将采取以下方法对实证结果进行稳健性检验：（1）δ取值-1，R_{max}取值 1.3 倍当前回归拟合优度，如果$\beta^*(R_{max}, \delta)$落在了估计参数的 95%置信区间，则结果通过稳健性检验；（2）R_{max}取值与方法（1）相同，计算使$\beta = 0$的δ，若δ取值大于 1，则结果通过稳健性检验。

鉴于本文主回归的拟合优度为 0.028，取R_{max}0.0364。稳健性检验的结果汇报在表III 1。结果显示，本文的结果通过了稳健性检验。

表III1 遗漏变量的稳健性检验

检验方法	判断标准	计算结果	是否通过
（1）	$\beta^*(R_{max}, \delta) \in [-0.1655, -0.0125]$	$\beta^*(R_{max}, \delta) = -0.0752$	是
（2）	$\delta > 1$	$\delta = 5.83$	是

附录 IV 附表及附图

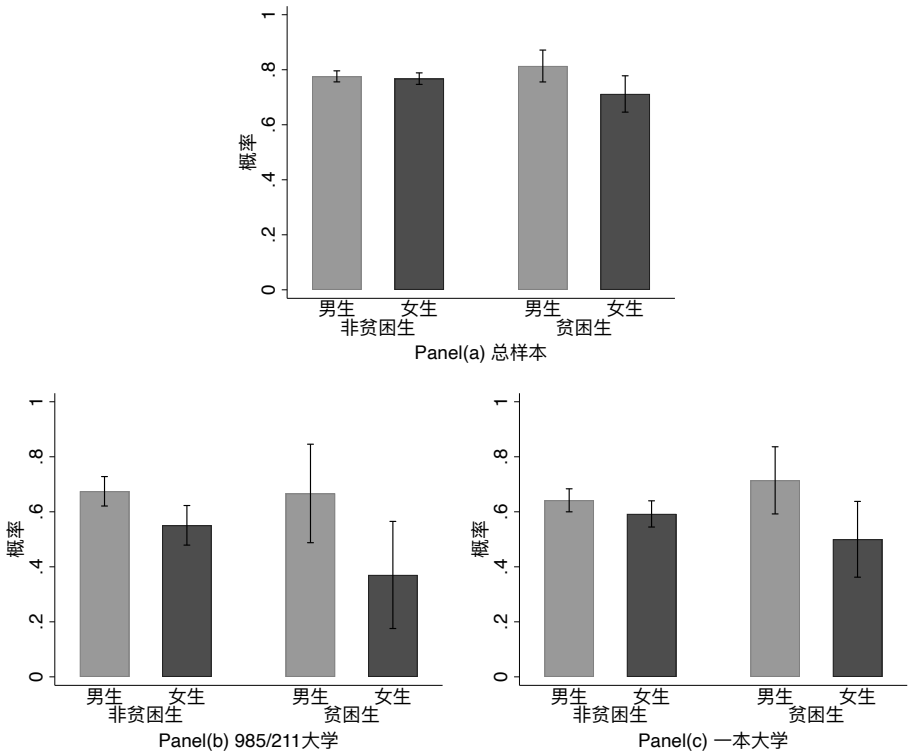

图 A1 重点大学录取率的性别差异

表 A1 河南省历年高考人数

	理科		文科		理科生占比
	最低分	考生人数	最低分	考生人数	
2014	246	334746	313	189926	63.80%
2015	288	317401	243	217345	59.36%
2016	241	352125	297	176037	66.67%
2017	256	336441	350	129270	72.24%
2019	291	370803	406	141158	72.43%
2020	339	366728	410	154177	70.40%
2021	337	374211	384	173558	68.32%
2022	405	314509	446	99663	75.94%

注：本表列出了河南省普通高考招生分数段统计表（一分一段表）中，本研究使用的样本中对应最低分及以上的考生数量，以及理科生在这些考生中的占比。

表 A2 原始分数描述性统计

	总样本	男生样本	女生样本	差值（女生-男生）
	（1）	（2）	（3）	（4）
模考分数				
总分	485.146	486.821	483.368	-3.453**
	(61.974)	(65.576)	(57.861)	(1.345)
语文	107.763	106.430	109.179	2.749***
	(8.207)	(8.097)	(8.086)	(0.176)
数学	96.938	99.488	94.230	-5.257***
	(19.634)	(19.506)	(19.409)	(0.422)
综合	173.434	177.085	169.557	-7.528***
	(35.909)	(37.519)	(33.690)	(0.775)
英语	107.023	103.841	110.401	6.560***
	(16.697)	(18.075)	(14.345)	(0.355)
高考分数				
总分	503.222	507.924	498.230	-9.694***
	(65.969)	(70.069)	(60.929)	(1.428)
语文	110.001	108.526	111.566	3.040***
	(9.542)	(9.710)	(9.104)	(0.205)
数学	99.848	101.941	97.626	-4.315***
	(18.310)	(18.436)	(17.912)	(0.395)
综合	185.531	192.746	177.872	-14.874***
	(36.404)	(37.692)	(33.322)	(0.774)
英语	107.849	104.722	111.169	6.447***
	(17.950)	(19.507)	(15.458)	(0.383)
观测值	8496	4375	4121	

表 A3 各类大学录取率(%)

	2014 年	2015 年	2016 年	2017 年
	(1)	(2)	(3)	(4)
Panel I 理科				
重点大学	20.0	17.8	15.0	13.1
普通一本院校	28.4	26.3	30.8	28.1
二本院校	44.4	47.9	47.3	49.4
高职高专	7.3	8.0	6.9	9.3
Panel II 文科				
重点大学	6.5	6.8	6.2	5.0
普通一本院校	17.4	15.1	19.8	22.5
二本院校	58.7	56.6	61.7	58.9
高职高专	17.4	21.5	12.3	13.6

注：(1)-(5)列分别列出了 2014 年-2018 年各类大学在某县的录取人数占所有大学总录取人数的比例。

表 A4 各类大学毕业生平均工资

	重点大学	普通一本大学	二本大学	高职高专大学
一年工资	6629	4724	4003	4167
三年工资	7843	5458	4551	4772
五年工资	9294	6296	5164	5486

注：本表展示了不同类型高校毕业生在毕业后第一年、第三年和第五年（2018 年）的平均薪资水平。

表 A5 贫困扩大应届生与复读生中性别差异的影响

	被解释变量：（标准化）高考总分与（标准化）模考总分的差值	
	应届生样本	复读生样本
	(1)	(2)
$Poor_i \times Female_i$	-0.107***	0.086
	(0.040)	(0.127)
$Poor_i$	0.056*	0.007
	(0.030)	(0.095)
$Female_i$	-0.023	-0.064
	(0.016)	(0.052)
个人控制变量	是	是
年份、班级固定效应	是	是
观测值	7587	909
R^2	0.026	0.083

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。复读生样本中包括参加过前一年高考、第二次参加高考的考生，应届生样本中包括第一次参加高考的考生。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量、表示是否为本地户口的虚拟变量、和参加高考时的年龄。

表 A6 分文理科贫困对扩大性别差异的影响

	被解释变量：各科目高考分数与模考分数的差值			
	语文	数学	综合	英语
	(1)	(2)	(3)	(4)
Panel I 理科样本				
$Poor_i \times Female_i$	-0.125	0.019	-0.067	-0.022
	(0.097)	(0.062)	(0.055)	(0.054)
$Poor_i$	0.018	-0.073	0.049	0.038
	(0.060)	(0.046)	(0.033)	(0.039)
$Female_i$	-0.026	0.013	-0.056***	-0.033*
	(0.030)	(0.020)	(0.017)	(0.017)
观测值	6788	6788	6788	6788
R^2	0.024	0.027	0.034	0.024
Panel II 文科样本				
$Poor_i \times Female_i$	0.093	-0.143	-0.526**	-0.255*
	(0.239)	(0.113)	(0.245)	(0.133)
$Poor_i$	-0.070	0.215**	0.402*	0.185
	(0.252)	(0.098)	(0.223)	(0.139)
$Female_i$	-0.008	0.037	-0.209***	0.002
	(0.078)	(0.044)	(0.071)	(0.061)
观测值	1708	1708	1708	1708
R^2	0.011	0.031	0.051	0.022

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。模型（1）-（4）分别是对学生的语文、数学、综合、英语的标准化分数的差值进行回归。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示复读生的虚拟变量或表示理科生的虚拟变量、表示本地户口的虚拟变量、和参加高考时的年龄。括号内为班级一级的聚类标准误。

表 A7 进入优质大学的概率(%)

	贫困女生	贫困男生	非贫困女生	非贫困男生
总样本	0.749	0.843	0.790	0.795
	(0.435)	(0.365)	(0.407)	(0.403)
985/211 大学	0.367	0.703	0.525	0.666
	(0.490)	(0.463)	(0.500)	(0.472)
一本	0.592	0.761	0.626	0.678
	(0.495)	(0.430)	(0.484)	(0.468)
二本	0.913	0.929	0.895	0.895
	(0.283)	(0.259)	(0.306)	(0.307)

注：表中分别列出了贫困女生、贫困男生、非贫困女生、非贫困男生进入优质大学的概率。第（1）行是所有学生，优质大学为优于或属于根据考生的模考成绩可能被录取的学校；第(2)行是有能力进入985，211 等重点大学的学生，优质大学定义为“985”、“211”重点大学；第（3）行是有能力进入普通一本大学的学生，优质大学为“985”、“211”等重点大学和普通一本院校；第（4）行是有能力进入二本大学的学生，优质大学为“985”、“211”等重点大学、和普通一本院校和二本院校。

表 A8 分科目 LASSO 回归结果

	被解释变量：高考各科目分数与预测分数之间的差值			
	语文	数学	综合	英语
	(1)	(2)	(3)	(4)
$Poor_i$	-0.308***	-0.028	-0.118***	-0.061*
	(0.050)	(0.040)	(0.037)	(0.032)
控制变量	是	是	是	是
年份固定效应	是	是	是	是
观测值	4121	4121	4121	4121
R^2	0.298	0.177	0.167	0.167

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。模型（1）-（4）分别是对学生的语文、数学、综合、英语的分数与预测分数之间的差值进行回归。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

表 A9 贫困和性别与复读选择的关系

	复读
	(1)
$Poor_i \times Female_i$	-0.023
	(0.027)
$Poor_i$	0.008
	(0.020)
$Female_i$	-0.033***
	(0.009)
个人控制变量	是
年份、班级固定效应	是
观测值	5512
R^2	0.141

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。被解释变量表示学生表示在 t 年高考之后选择复读的虚拟变量。由于缺乏 2018 年和 2023 年的考生样本，无法定义 2017 年和 2022 年参加高考的学生是否选择复读的状态，因此不包括 2017 年和 2022 年参加高考的学生样本。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

表 A10 年龄对决策质量的影响

	（标准化）高考总分与（标准化）模考总分的差值
	(1)
Age_i	-0.017**
	(0.008)
$Poor_i$	-0.000
	(0.020)
$Female_i$	-0.034**
	(0.015)
个人控制变量	是
年份、班级固定效应	是
观测值	7587
R^2	0.025

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

表 A11 贫困和性别与年龄的关系

	参加高考时的年龄
	(1)
$Poor_i \times Female_i$	0.095*
	(0.051)
$Poor_i$	0.196***
	(0.042)
$Female_i$	-0.088***
	(0.018)
个人控制变量	是
年份、班级固定效应	是
观测值	7587
R^2	0.232

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。被解释变量表示学生参加高考时的年龄。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

表 A12 贫困和性别与住校的关系

	住校
	(1)
$Poor_i \times Female_i$	0.022
	(0.092)
$Poor_i$	0.271***
	(0.070)
$Female_i$	-0.004
	(0.032)
个人控制变量	是
年份、班级固定效应	是
观测值	1,041
R^2	0.109

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。被解释变量表示学生高考时住校的变量。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

表 A13 住校对决策质量的影响

	（标准化）高考总分与（标准化）模考总分的差值
	(1)
$zhuxiao_i$	-0.062
	(0.045)
$Poor_i$	-0.052
	(0.046)
$Female_i$	0.033
	(0.044)
个人控制变量	是
年份、班级固定效应	是
观测值	1,040
R^2	0.039

注：***代表 p<0.01，**代表 p<0.05，*代表 p<0.1。所有回归均控制了年份固定效应、班级固定效应和个人层面的控制变量，控制变量包括表示理科生的虚拟变量，表示复读生的虚拟变量，表示本地户口的虚拟变量，和参加高考时的年龄。

参考文献

[1] Oster, E., “Unobservable selection and coefficient stability: Theory and evidence”, *Journal of Business & Economic Statistics*, 2019, 37(2), 187-204.