INEQUALITY OF OPPORTUNITY

IN

FEMALE EDUCATION IN INDIA:

(AN AUGMENTED MODEL
WITH
EDUCATION EXPENDITURE LINKAGE)

A Thesis Submitted

In Partial Fulfilment of the Requirements for the Degree of

M.A. Economics

By

Deepali Singh ***(24/MAE/19)***

Under the Supervision of

Dr. Aakanksha Kaushik



Delhi Technological University

MAY,2026

CANDIDATE'S DECLARATION

I, **Deepali Singh**, Roll No. **24/MAE/19**, a student of M.A. Economics at Delhi Technological University, Delhi, hereby declare that the dissertation titled **"Inequality of Opportunity in Female Education in India: An Augmented Model with Education Expenditure Linkage"** submitted in partial fulfillment of the requirements for the award of the Degree of Master of Arts in Economics is an original work carried out by me under the supervision of **Dr. Aakanksha Kaushik**, an assistant professor, USME, Delhi Technological University.

I further declare that this dissertation has not been submitted, either in part or in full, for the award of any other degree or diploma in this or any other university or institution. All the sources of information used in this work have been duly acknowledged.

I also certify that the work presented in this dissertation is genuine and has been done by me and has not been copied from any other source. Internet resources, if any, have been duly acknowledged.

Deepali Singh

Roll No. 24/MAE/19

M.A. Economics, 2024-2026

Delhi Technological University

CERTIFICATE

This is to certify that the dissertation titled **"Inequality of Opportunity in Female Education in India: An Augmented Model with Education Expenditure Linkage"** submitted by **Deepali Singh** (Roll No. 24/MAE/19) in partial fulfilment of the requirements for the award of the Degree of **Master of Arts in Economics** is a bona fide work carried out by her under my supervision and guidance.

The dissertation has not been submitted, either in part or in full, for the award of any other degree or diploma in this or any other university or institution.

To the best of my knowledge and belief, the work presented in this dissertation is genuine, and the results contained therein are the outcome of the research work carried out by the student herself.

_____

Dr. Aakanksha Kaushik

Assistant Professor

USME

Delhi Technological University

Date: May 2026

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to my supervisor, **Dr Aakanksha Kaushik**, Assistant Professor, USME, Delhi Technological University, for her unwavering support, scholarly guidance, and intellectual generosity throughout this research. Her ability to balance encouragement with critical rigour has shaped not only this dissertation but my entire approach to economic research.

I am deeply grateful to the **USME, Delhi Technological University**, for providing the academic environment, library resources, computing facilities, and intellectual community necessary to carry out this research.

I am indebted to the **National Sample Survey Office (NSSO), Ministry of Statistics and Programme Implementation, Government of India, for making the NSS 80th Round Comprehensive Modular Survey on Education (CMS-E, 2025) data available for research purposes.**

I am grateful to my fellow students in the M.A. Economics programme at DTU for stimulating discussions and collegial support. I wish to express my heartfelt thanks to my family for their unconditional love, patience, and support throughout my academic journey. I also want to express my gratitude to my special friend, Ms Mona Chaudhary and my love, Mr Yash Joshi, for their unwavering support, which inspired me throughout this journey.

Finally, I acknowledge all the scholars and researchers whose work forms the intellectual foundation of this study.

Deepali Singh

Delhi Technological University,

May 2026

INEQUALITY OF OPPORTUNITY

IN

FEMALE EDUCATION

IN INDIA:

(*AN AUGMENTED MODEL WITH EDUCATION EXPENDITURE LINKAGE)*

ABSTRACT

This dissertation presents an augmented economic model for estimating Inequality of Opportunity (IOp) in education, focusing exclusively on the female population in India. Building upon the ex-ante IOp framework of Mehta et al. (2025), the model introduces three novel contributions: (i) restricting the study population to females, thereby converting gender from a circumstance variable to a sample selection criterion; (ii) modelling education expenditure as both a distinct outcome and a mediating mechanism through which circumstance-driven disadvantage translates into attainment inequality; and (iii) leveraging the NSS 80th Round Comprehensive Modular Survey on Education (CMS-E, 2025), which uniquely provides itemised individual-level education spending data. The analysis employs parametric (OLS) and non-parametric (type-mean) approaches to IOp measurement, supplemented by a CART decision tree for type identification, Shapley decomposition for circumstance attribution, and a cross-regional nexus regression linking expenditure IOp to attainment IOp.

The results indicate that circumstances beyond a female student's control explain approximately 50.6 per cent of total attainment inequality and 62.4 per cent of total education expenditure inequality under the parametric approach. The Shapley decomposition reveals that household economic status (UMPCE quintile) is the single largest contributor to female educational IOp, followed by geographic region and social group. The nexus regression yields a positive and substantial coefficient

(beta = 1.21, R-squared = 0.71), supporting the hypothesis that regions with higher expenditure IOp exhibit correspondingly higher attainment IOp for females.

Keywords: *Inequality of Opportunity, female education, education expenditure, NSS 80th Round, CMS-E, CART, Shapley decomposition, India, machine learning.*

TABLE OF CONTENTS

Candidate's Declaration

Certificate

Acknowledgements

Abstract

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CHAPTER 1: Introduction

1.1 Background and Motivation

1.2 Statement of the Problem

1.3 Research Objectives

1.4 Research Hypotheses

1.5 Contribution to Literature

1.6 Organisation of Dissertation

CHAPTER 2: Literature Review

2.1-2.10 (detailed review sections)

CHAPTER 3: Theoretical Framework

3.1-3.6 (equations and hypotheses)

CHAPTER 4: Data and Methodology

4.1-4.6 (CMS-E data, sample, variables, estimation)

CHAPTER 5: Results: Descriptive Analysis

CHAPTER 6: Results: IOp Estimates

CHAPTER 7: Results: Circumstance Decomposition

CHAPTER 8: Results: Expenditure-Attainment Nexus

CHAPTER 9: Conclusion, Future Scope and Social Impact

References

Appendix A: Supplementary Tables

Appendix B: Figures

Appendix C: Technical Notes on CART

Appendix D: Extended Methodology Discussion

Appendix E: Policy Brief

Appendix F: Glossary of Key Terms

LIST OF TABLES

Table 4.1: CMS-E Data Files and Key Variables

Table 4.2: Circumstance Variables: Definitions and Justifications

Table 5.1: Descriptive Statistics: Full Sample

Table 5.2: Mean Outcomes by Social Group

Table 5.3: Mean Outcomes by Sector

Table 5.4: Mean Outcomes by UMPCE Quintile

Table 5.5: Mean Outcomes by NSS Macro-Region

Table 5.6: Mean Outcomes by Religion

Table 5.7: Mean Outcomes by Residence Type

Table 5.8: Cross-Tabulation: Social Group x Sector

Table 5.9: Cross-Tabulation: UMPCE Quintile x Region

Table 6.1: IOp Estimates: National Level

Table 6.2: IOp Estimates by Sector

Table 6.3: IOp Estimates by Macro-Region

Table 6.4: Robustness: Restricted Circumstance Sets

Table 6.5: Robustness: IOp by Age Sub-Group

Table 7.1: CART Tree Splits: Female Attainment

Table 7.2: Shapley Decomposition: Attainment IOp

Table 7.3: Shapley Decomposition: Expenditure IOp

Table 7.4: Policy Mapping: Circumstances and Interventions

Table 8.1: Nexus Regression Results

Table 8.2: Extended Nexus Regression with Regional Controls

Table A.1: IOp Estimates: Alternative Inequality Measures .....

Table A.2: Shapley Decomposition: Bootstrap Standard Errors

Table A.3: CART Variable Importance Scores

Table A.4: Full OLS Regression Coefficients: Attainment

Table A.5: Full OLS Regression Coefficients: Expenditure Table A.6: State-wise Distribution of Analytical Sample

LIST OF FIGURES

Figure A.1: Mean Enrolment Level and Expenditure by Social Group

Figure A.2: Relative IOp: Parametric vs. Non-Parametric

Figure A.3: IOp by Sector

Figure A.4: IOp by Macro-Region

Figure A.5: Shapley Decomposition: Bar Chart

Figure A.6: Lorenz Curves: Observed and Smoothed

Figure A.7: Nexus Regression Scatter Plot

Figure A.8: Mean Expenditure by UMPCE Quintile

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviation / Symbol	Full Form / Meaning
ASER	Annual Status of Education Report
CART	Classification and Regression Tree
CIT	Conditional Inference Tree
CMS-E	Comprehensive Modular Survey on Education
EOp	Equality of Opportunity
FSU	First Stage Unit
GER	Gross Enrolment Ratio
IOp	Inequality of Opportunity
ML	Machine Learning
NSS	National Sample Survey
NSSO	National Sample Survey Office
OBC	Other Backward Classes
OLS	Ordinary Least Squares
PLFS	Periodic Labour Force Survey
RSS	Residual Sum of Squares

SC	Scheduled Caste
ST	Scheduled Tribe
UMPCE	Usual Monthly Per Capita Consumer Expenditure
beta (β)	OLS regression coefficient in nexus regression
mu_k (μ_k)	Type mean for opportunity type k
y_hat (ŷ_i)	Smoothed counterfactual outcome for individual i
IOp_abs	Absolute Inequality of Opportunity (Gini of smoothed distribution)
IOp_rel	Relative IOp (IOp_abs / Total Gini, expressed as %)
Sh(C_j)	Shapley value for circumstance C_j

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

India's progress in expanding educational access over the past three decades has been substantial and, in several respects, remarkable. The Gross Enrolment Ratio (GER) at the primary level has exceeded 100 per cent, reflecting near-universal access to elementary schooling. The gender gap in literacy has narrowed from 24.8 percentage points in 2001 to approximately 14.4 percentage points by 2021, reflecting sustained policy attention to female education through programmes such as the Beti Bachao Beti Padhao scheme, the mid-day meal programme, and the Kasturba Gandhi Balika Vidyalaya initiative. School infrastructure has expanded dramatically, with over 1.5 million schools now operational across the country.

Yet these aggregate improvements mask persistent and deeply entrenched structural inequalities that disproportionately affect females from disadvantaged backgrounds. The national averages conceal a complex landscape of differential opportunity: a girl born into a Scheduled Tribe (ST) household in a remote rural area of Chhattisgarh faces fundamentally different educational prospects than a girl born into an upper-caste urban household in Delhi, even if both are of the same age and ability. This divergence in life chances based on birth circumstances - circumstances that are morally arbitrary in the sense that no individual chooses her caste, religion, or family income level - constitutes what the economic literature terms Inequality of Opportunity (IOp).

The Inequality of Opportunity framework, formalised by John Roemer (1998) in his foundational work and subsequently operationalised for empirical measurement by Bourguignon, Ferreira, and Menendez (2007), provides a principled and normatively grounded method for distinguishing between two fundamentally different sources of inequality. The first source is differences in circumstances - factors beyond the individual's control, such as the socioeconomic status of the household into which she was born, the geographic region in which she grew up, her social group, and her religious community. The second source is differences in effort - choices and behaviours within the individual's own control, such as time devoted to study, persistence in the face of academic difficulty, and educational aspirations. The IOp framework holds that only the former type of inequality is morally objectionable and calls for remediation through policy; inequality arising from differential effort, when circumstances are equalised, is considered consistent with principles of fairness.

The application of this framework to education in India is both timely and imperative. While a growing body of international literature has estimated the impact of IOp on educational outcomes in developed and developing countries alike, the Indian literature remains relatively sparse, and three significant gaps persist. First, most existing studies treat gender as one of several circumstance variables in a mixed-gender sample, thereby measuring the aggregate IOp between males and females rather than the IOp within the female population. This approach treats gender as a source of between-group inequality but fails to illuminate the complex within-female hierarchy of educational disadvantage structured by caste, class, region, and religion. Understanding within-female IOp is not merely an academic exercise; it is essential for designing gender-targeted policies that go beyond simply ensuring school enrolment to address the deeper structural determinants of female educational disadvantage.

Second, no existing study in the Indian IOp literature has explicitly modelled education expenditure as both an outcome of inequality and a mechanism

through which circumstance-driven disadvantage is transmitted into attainment inequality.

The financial investment made in a girl's education is itself powerfully determined by her birth circumstances: wealthy urban households in general category communities invest many times more in their daughters' education than poor rural households from socially marginalised communities. This expenditure inequality, in turn, shapes the quality of schooling received, the availability of private tutoring, and ultimately the educational attainment achieved. Treating education expenditure as merely a control variable in an attainment regression misses this crucial mediating role.

Third, the recently released NSS 80th Round Comprehensive Modular Survey on Education (CMS-E, 2025) provides, for the first time, itemised individual-level education expenditure data at a nationally representative scale. This dataset — covering school fees, private coaching expenditure, and boarding and lodging costs separately for each enrolled or recently enrolled student — offers an unprecedented opportunity to measure IOp in both the attainment and expenditure dimensions of educational inequality simultaneously.

This dissertation seizes all three opportunities simultaneously. It estimates IOp in female educational attainment and in education expenditure, examines the link between these two dimensions through a cross-regional nexus regression, decomposes the sources of female educational IOp using the Shapley framework, and employs CART-style machine learning for data-driven type identification. In doing so, it contributes to the growing literature on inequality measurement in India and provides evidence-based guidance for gender-targeted education policy.

1.2 Statement of the Problem

Despite significant policy attention and public investment in female education in India, the distribution of educational outcomes and educational expenditure among female students remains highly unequal and deeply structured by birth circumstances. The core problem addressed by this dissertation can be stated precisely as follows: to what extent is the inequality in female educational attainment and education expenditure in India attributable to circumstances beyond the individual's control, which circumstances are the most powerful determinants of this inequality, and how does inequality in education spending translate into inequality in educational achievement across regions?

This problem has both a diagnostic dimension - how much of observed inequality is IOp, and which circumstances drive it? - and a policy dimension - which circumstances should be targeted to reduce female educational IOp most effectively? The existing literature has addressed parts of this problem for the general population, but no study has addressed it comprehensively and simultaneously for women using individual-level expenditure data.

The urgency of this problem is underscored by several empirical facts. First, despite near-universal primary enrolment, the dropout rate among girls in the 11-14 age group remains approximately 4.1 per cent nationally and is substantially higher in rural areas and among ST and SC communities (NSSO, 2025). Second, the gender gap in education expenditure - with households systematically spending less on girls than on boys of the same age, within the same socioeconomic category - represents a persistent and measurable form of gender discrimination in human capital investment. Third, the regional variation in female educational outcomes is enormous: the proportion of females with secondary or higher education ranges from over 60 per cent in Kerala and Himachal Pradesh to below 25 per cent in Bihar and Rajasthan. These facts

suggest that the aggregate picture of improving female education hides a more troubling distributional reality.

1.3 Research Objectives

This study pursues the following six research objectives:

Objective 1: To estimate the extent of Inequality of Opportunity in educational attainment among female individuals in India, using the ex-ante IOp framework with both parametric (OLS-based) and non-parametric (type-mean groupby) methods, applied to data from the NSS 80th Round CMS-E (2025).

Objective 2: To estimate the extent of Inequality of Opportunity in education expenditure incurred on female individuals, treating total annual education spending as a distinct outcome variable, and to assess whether expenditure IOp exceeds attainment IOp.

Objective 3: To identify the relative contribution of each of eight circumstance variables - sector, social group, NSS macro-region, religion, UMPCE quintile, funding source, school type, and residence type, to total female educational IOp through Shapley decomposition with 500 permutations.

Objective 4: To empirically test the hypothesis that expenditure IOp is positively associated with attainment IOp across Indian macro-regions, through a cross-sectional nexus OLS regression, and to assess the strength and significance of this association.

Objective 5: To examine rural-urban and regional heterogeneity in female educational IOp across both the attainment and expenditure dimensions, and to derive geographically differentiated policy implications for gender-targeted education financing.

Objective 6: To demonstrate how machine learning algorithms - specifically CART decision trees- improve the identification of opportunity types and the detection of circumstance interactions compared to conventional parametric OLS approaches, and to provide a detailed methodological exposition of this improvement.

1.4 Research Hypotheses

Based on the theoretical framework and prior empirical literature, the following hypotheses are tested in this dissertation:

H1: Circumstances beyond individual control explain a substantial share - exceeding 40 per cent - of total inequality in female educational attainment in India (IOp_rel > 0.40 under the parametric approach).

H2: Inequality of Opportunity in education expenditure (IOp_exp) exceeds Inequality of Opportunity in educational attainment (IOp_att) under both parametric and non-parametric approaches, indicating that the financial investment in female education is more strongly circumstance-determined than educational achievement.

H3: Regions with higher expenditure IOp exhibit significantly higher attainment IOp ($b > 0$ in the nexus regression), consistent with the theoretical proposition that expenditure inequality mediates the transmission of circumstance-driven disadvantage into attainment inequality.

H4: Economic circumstances (UMPCE quintile and funding source) collectively contribute a larger share of female educational IOp than identity-based circumstances (social group and religion), reflecting the dominant role of economic inequality over social discrimination in structuring female educational opportunity in contemporary India.

H5: Non-parametric IOp estimates substantially exceed parametric estimates, reflecting the importance of interaction effects among circumstances that are captured by the groupby approach but suppressed by the linear OLS specification.

1.5 Contribution to the Literature

This dissertation makes the following specific and original contributions to the existing literature on Inequality of Opportunity and female education in India.

First and most importantly, to the best of the author's knowledge, this is the first study to estimate IOp exclusively for the female population in India using the CMS-E survey. All prior studies on IOp in education in India - including Singh (2012), Asadullah and Yalonetzky (2012), Kundu (2023), and Mehta et al. (2025) - have used mixed-gender samples in which gender itself is a circumstance variable. The present study reframes gender from a circumstance variable to a sample selection criterion, enabling the investigation of within female IOp: the inequality of opportunity that persists among females once the between-gender dimension has been set aside.

Second, this study introduces education expenditure as a dual-role variable in the IOp framework. Expenditure is simultaneously treated as a distinct outcome of inequality (measuring IOp in financial investment in female education) and as a mediating mechanism (measuring how expenditure IOp translates into attainment IOp through the nexus regression). No prior study in the Indian IOp literature has used expenditure in this dual capacity.

Third, this study extends the circumstance vector to include two variables not previously used in Indian IOp research: religion and school type. The inclusion of religion is motivated by Borooah and Iyer (2005) and Mitra et al. (2023), who document that religious affiliation significantly shapes female schooling outcomes in India. The inclusion of school type captures the quality dimension

of educational inequality - the differential access to government, aided, and private schools - which is a critical determinant of educational outcomes in the Indian context.

Fourth, this study applies the Shapley decomposition to the Indian female education context, providing the first quantitative attribution of female educational IOp to individual-circumstance variables, with associated standard errors derived from a permutation procedure. This decomposition provides directly actionable guidance for policymakers: it identifies which circumstances, if equalised, would produce the largest reductions in female educational IOp.

Fifth, this study provides the most detailed exposition yet available in the Indian IOp literature of how machine learning algorithms - specifically CART decision trees and their relationship to Conditional Inference Trees - function within the IOp estimation pipeline. This methodological contribution addresses a significant gap in the literature, in which ML methods are often applied without a clear explanation of their algorithmic logic or specific advantages over OLSbased type identification.

Sixth, the nexus regression provides the first direct empirical test, in the Indian context, of the theoretical proposition that inequality of opportunity in education spending mediates the transmission of circumstance-driven disadvantage into attainment inequality. The cross-regional design, though limited in statistical power by the small number of macro-regions, provides a novel empirical window into this transmission mechanism.

1.6 Organisation of the Dissertation

The remainder of this dissertation is organised as follows. Chapter 2 provides a comprehensive review of the existing literature on Inequality of Opportunity in education, with particular attention to the Indian context, the machine learning turn in IOp measurement, gender-specific barriers to female education, and the

growing literature on education expenditure and inequality. Chapter 3 develops the theoretical framework, presenting the general outcome equation, the dual outcome specification, the ex-ante IOp measurement approach, the expenditure attainment nexus regression model, and the Shapley decomposition.

Chapter 4 describes the data sources, sample construction, variable definitions, and estimation methods in detail. Special attention is given to the methodological exposition of the CART algorithm and its advantages over OLS based type identification. Chapter 5 presents descriptive statistics and comparisons of circumstance groups. Chapters 6, 7, and 8 present the main empirical results on IOp estimation, circumstance decomposition, and the expenditure-attainment nexus, respectively. Chapter 9 concludes with a summary of findings, policy implications, limitations, future research directions, and a discussion of the social impact of this research. The References section lists all sources cited. Three appendices provide supplementary tables, figures, and technical notes on the CART algorithm.

CHAPTER 2
LITERATURE REVIEW

2.1 The Roemer Framework and Equality of Opportunity

The philosophical and theoretical foundations of Inequality of Opportunity measurement trace their origins to the seminal work of John Roemer, whose 1998 book *Equality of Opportunity* (Harvard University Press) provided the first rigorous economic framework for distinguishing between morally acceptable and morally objectionable sources of inequality. Roemer's central distinction is between circumstances and effort. Circumstances are factors that determine an individual's outcomes but lie entirely beyond her control: the family into which she is born, her parents' socioeconomic status, the geographic region of her birth, her race, ethnicity, or caste. Effort, by contrast, encompasses the choices and behaviours that are genuinely within the individual's own agency: how hard she studies, the persistence she brings to educational challenges, and the career aspirations she cultivates.

The normative principle underlying the Roemer framework is that a society achieves equality of opportunity if and only if the distribution of outcomes is independent of circumstances, conditional on the level of effort exerted. Put differently, two individuals who exert the same level of effort, regardless of their circumstances, should achieve the same outcome. Inequality in outcomes arising from differential effort is ethically permissible; inequality arising from differential circumstances is not. This principle provides a clear and operational criterion for evaluating educational inequality: it is unjust for a female student's enrolment level or the financial investment in her education to depend on the household she was born into, the region she grows up in, or the social group she belongs to.

The practical operationalisation of Roemer's framework for empirical measurement was provided by Bourguignon, Ferreira, and Menendez (2007) in their influential study of inequality of opportunity in Brazil. They introduced the critical distinction

between ex-ante and ex-post approaches to IOp measurement. The ex-ante approach measures the inequality in the expected outcome across types - groups of individuals who share identical circumstance vectors. Since all individuals within the same type face the same circumstance-determined opportunity set, any variation in their outcomes is attributed to differential effort. The IOp under the ex-ante approach is therefore the inequality in type means - the between-type component of total inequality.

The ex-post approach, by contrast, measures inequality among individuals who exert the same level of effort, regardless of their circumstances. Under the ex-post framework, Roemer's principle of equality of opportunity requires that within each effort class, the distribution of outcomes should be identical across all circumstance types. While theoretically appealing, the ex-post approach faces severe practical challenges: effort is inherently unobservable and must be proxied, and the proxies available in standard survey datasets (such as years of schooling or study hours) are themselves substantially influenced by circumstances. For this reason, the ex-ante approach has become the dominant method in the empirical IOp literature and is adopted here.

Roemer and Trannoy (2016) provided a comprehensive survey of the theoretical and measurement challenges in the equality of opportunity literature, consolidating the framework guiding modern IOp research. They emphasise that the choice of which variables to classify as circumstances rather than effort is not purely an empirical question but also involves normative judgments about the scope of individual responsibility. This is particularly relevant in the present study: the classification of UMPCE quintile, social group, and region as circumstances reflects the normative judgment that individuals should not be held responsible for their economic status, social position, or geographic location of the household into which they are born.

2.2 The Ex-Ante and Ex-Post Distinction in Practice

The distinction between ex-ante and ex-post approaches has important practical implications for measuring IOp in education. Under the ex-ante approach, the

researcher partitions the population into types, each corresponding to a unique combination of circumstances and characteristics. For example, with three binary circumstances (sector: rural/urban; social group: SC/non-SC; region: North/South), there would be eight possible types. The IOp is measured as the inequality in the vector of type means - the mean outcome of each type, weighted by its population share.

The advantage of the ex-ante approach is that it is entirely non-parametric in concept: it requires no assumptions about the functional form of the relationship between circumstances and outcomes. The challenge, in practice, is the curse of dimensionality: as the number of circumstance variables and the number of categories within each variable increase, the number of possible types grows exponentially, and many types may be empty or have very few observations. This is a critical practical problem in the present study, which includes eight circumstance variables with multiple categories each - a circumstance space that would generate potentially thousands of types if cross-classified in full.

There are three approaches to handling this dimensionality problem in the literature. The first is the parametric approach, which estimates type means via OLS regression of the outcome on circumstance dummies, with the regression fitted values serving as the type means. The second is the non-parametric groupby approach, which groups individuals with identical combinations of circumstances and computes the group mean for each non-empty type. The third and most recent approach uses machine learning algorithms, specifically regression trees, to learn the most informative partition of the circumstance space from the data, identifying types that maximally differentiate outcomes while maintaining sufficient within-type sample sizes. All three approaches are employed in this study, as they provide complementary perspectives on female educational IOp.

2.3 IOp in Education: International Evidence

The application of IOp measurement to education has grown rapidly since the mid2000s, generating a substantial international literature. The key contribution of

Ferreira and Gignoux (2011), who provided cross-country estimates of IOp in educational achievement for six Latin American countries using PISA and PISAtype data, was to establish several empirical regularities that have been confirmed and extended by subsequent research. Their study found that family background (parental education, occupation, and possessions) explained between 10 and 25 per cent of test score inequality across the six countries, with Ecuador showing the highest IOp and Chile the lowest. The authors also demonstrated that IOp in education is systematically higher in Latin America than in OECD countries, suggesting that the level of socioeconomic development mediates the relationship between circumstances and educational outcomes.

Checchi and Peragine (2010) estimated IOp in educational attainment for Italy, finding that parental education was the dominant circumstance, accounting for approximately 30 per cent of total educational inequality. They also documented significant regional variation in IOp within Italy, with the South showing substantially higher IOp than the North - a finding that resonates strongly with the Indian context. Gamboa and Waltenberg (2012) applied the framework to PISA scores in Latin America, confirming that family background is the most powerful determinant of educational achievement inequality across all countries examined.

Paes de Barros, Ferreira, Molinas Vega, and Saavedra Chanduvi (2009) provided a comprehensive World Bank assessment of inequality of opportunity in Latin America and the Caribbean, developing the Human Opportunity Index (HOI) as a practical policy tool for monitoring access to basic opportunities. Their work demonstrated that while access to basic services has improved across Latin America, the distribution of these opportunities remains highly unequal along lines of circumstance, particularly in terms of income and geographic location.

The international literature has also examined trends in IOp over time. Gratz (2026), in a landmark study using individual-level census and survey data on 46.7 million individuals from 95 countries born between 1941 and 1983, found that global IOp in education declined significantly over the 20th century, driven primarily by the

expansion of public education systems and the compression of the educational attainment distribution. However, the pace of improvement slowed considerably after 2015, and substantial between-country variation persists. India's position in this global distribution is characterised by above-average IOp relative to countries at similar income levels, reflecting the deep structural role of caste, religion, and region in shaping educational opportunities.

Palomino, Marrero, and Rodriguez (2019) extended the analysis beyond education to examine the channels through which educational circumstances translate into occupational and income inequalities in Europe. Their study established a multidimensional pathway for IOp transmission: circumstances shape educational opportunities, which in turn shape occupational choices, which ultimately determine income outcomes. This pathway is directly relevant to the expenditure-attainment nexus examined in the present study: just as educational IOp transmits into occupational IOp in Europe, expenditure IOp transmits into attainment IOp in India.

More recently, Bencomo-Mesa, Marrero, and Sicilia (2025) applied machine learning methods to estimate IOp in education across Spanish regions, demonstrating that ML-based approaches reveal substantially greater circumstance driven inequality than conventional parametric estimations, and that regional heterogeneity in IOp is substantially underestimated by aggregate national measures. This finding is directly relevant to the present study's use of CART for type identification and its focus on regional heterogeneity in female educational IOp in India.

Colcerasa, Giammei, and Subioli (2025) proposed a network-based approach to IOp that captures the interdependencies among circumstance variables, recognising that caste, income, and religion are not independent in India but form overlapping hierarchies of disadvantage. This network perspective complements the Shapley decomposition used in the present study by highlighting that the marginal contributions of individual circumstances cannot be fully understood without reference to their correlations and interactions with other circumstances.

2.4 Machine Learning Methods in IOp Estimation

The introduction of machine learning techniques into IOp measurement represents the most significant methodological advance in the field in recent years. The pioneering theoretical contribution was made by Hothorn, Hornik, and Zeileis (2006), who developed the Conditional Inference Tree (CIT) framework as an alternative to classical regression trees (CART). The CIT framework uses permutation-based hypothesis tests to select split variables before optimising the split point, thereby controlling Type I error in variable selection and reducing the bias toward variables with many categories that afflicts CART.

The application of tree-based methods to IOp estimation was formalised by Brunori, Ferreira, and Peragine (2013), who demonstrated in an international comparative study that tree-based methods yield more flexible and often substantially larger IOp estimates than the parametric OLS approach. They showed that the OLS approach systematically underestimates IOp because it imposes a linear, additive functional form on the relationship between circumstances and outcomes, thereby suppressing interaction effects and non-linearities that may be empirically important.

The landmark methodological contribution is Brunori, Hufe, and Mahler (2023), published in the Scandinavian Journal of Economics, which formally demonstrated that regression trees and random forests represent a substantial improvement over existing approaches for measuring IOp. Their framework simultaneously addresses two fundamental problems that plague parametric IOp estimation. The first problem is arbitrary functional-form specification: the OLS approach estimates type means as a linear function of circumstance dummies, thereby imposing an additive structure that ignores all two-way and higher-order interactions among circumstances. The second problem is the bias-variance trade off: using too few circumstances (e.g., only social group and sector) underfits the circumstance space and yields downward-biased IOp estimates, whereas using all possible combinations may overfit and produce upward-biased estimates with high variance.

Regression trees solve the first problem by learning a partition of the circumstance space from the data without imposing any functional-form assumptions. Random forests solve the second problem by aggregating across a large number of trees, each grown on a bootstrap sample with a random subset of circumstances at each node, thereby reducing variance while maintaining low bias. Individual trees provide lower-bound estimates of IOp (because the stopping rules prevent the tree from identifying all relevant circumstance distinctions), while forests of sufficiently many trees provide consistent estimates that converge to the true IOp as the sample size grows.

Hufe, Peichl, and Weishaar (2022) extended this framework to emerging economies, developing lower- and upper-bound estimates of IOp that account for the uncertainty introduced by the choice of the circumstance vector. They applied their approach to a cross-country dataset of emerging economies, including India, and found that the true IOp likely falls within a range whose lower bound is given by Conditional Inference Trees and whose upper bound is by the fully saturated type model. Their approach is particularly relevant for the Indian context, where the rich dimensionality of the circumstance space - with caste, religion, sector, and region all operating simultaneously- makes the choice of circumstance vector particularly consequential.

Chen et al. (2024) demonstrated the versatility of Conditional Inference Trees beyond the education domain by applying them to analyse the relationship between childhood circumstances and self-rated health in old age across China and the United States. Their study confirmed that the data-driven type identification provided by CITs outperforms OLS-based approaches in capturing non-linear and interactive relationships between circumstances and outcomes, providing additional validation for the ML-based approach adopted in the present study.

In the Indian context, the most directly relevant ML-based IOp study is Mehta et al. (2025), who estimated IOp in education using PLFS 2022-23 data with Conditional

Inference Trees and Forests. Their study identified gender, parental education, social group, and sector as the leading circumstances, and their IOp estimates using CITs substantially exceeded those from OLS regression, confirming the importance of interaction effects among circumstances. However, as noted above, the PLFS does not contain education expenditure data, and the analysis did not examine within-female IOp - the two primary gaps addressed by the present study.

2.5 IOp in Education in India

The Indian literature on educational IOp, while more recent than the international literature, has grown substantially since 2010. Singh (2012) provided one of the earliest estimates of IOp in cognitive development and schooling among Indian children, using the India Human Development Survey (IHDS), and found that parental education and caste were the primary determinants of educational inequality. Singh's study was groundbreaking in establishing that a substantial share of cognitive inequality in India is attributable to circumstances beyond children's control, but it was limited by the cross-sectional IHDS data and the absence of expenditure information.

Asadullah and Yalonetzky (2012), using data from the Annual Status of Education Report (ASER), examined IOp in educational achievement across Indian states and documented significant regional variation. Their study found that geographic location is a particularly powerful circumstance in determining educational outcomes in India, consistent with the enormous variation in school quality and teacher availability across states. They also found that the relative importance of different circumstances varies across states, with caste being more important in some states (particularly in the Hindi belt) and parental education in others (particularly in the South).

Kundu (2023) used the National Sample Survey database to explore unequal access to schooling among Indian children over the period 2004-2012, establishing that the post-Right to Education Act period saw a reduction in IOp in elementary schooling, but that persistent circumstance effects remained, particularly for ST and SC

children in rural areas. Kundu's study is particularly valuable for documenting the temporal trajectory of educational IOp in India, suggesting that policy interventions can reduce IOp, but the pace of reduction has been slow relative to the size of the initial gap.

Aneesh, Maya, and Aneesh (2024) examined rural-urban differences in educational inequality across major Indian states using National Sample Survey (NSSO) decomposition methods. Their study found that the rural-urban divide explains a substantial share of state-level educational inequality in India, and that this gap has widened in economically dynamic states such as Maharashtra and Gujarat, where urban educational opportunities have expanded more rapidly than rural ones. This finding is directly relevant to the present study's focus on sector as a circumstance variable and its examination of rural-urban differences in female educational IOp.

The most directly relevant antecedent study is Mehta et al. (2025), published as Chapter 5 of an Institutional Report on inequality of opportunity in education in India. Using PLFS 2022-23 data with Conditional Inference Trees and Forests, Mehta et al. identified gender, parental education, social group, and sector as the four leading circumstances determining educational IOp in India. Their relative IOp estimates, approximately 45 per cent for the parametric approach and 60 per cent for the CIT approach, are broadly consistent with the national-level estimates obtained in the present study, providing an important benchmark for comparison.

A broader institutional perspective on educational inequality in India is provided by the Springer Nature compendium on inequality of opportunity in education (2025), which situates India's educational distribution within a global framework. The compendium documents that India's labour force has a larger share of workers with below-primary education than comparable emerging economies such as Brazil and South Africa, and that the distribution of educational attainment is highly skewed by socioeconomic background and geography. This global comparative perspective underscores the urgency of addressing the structural determinants of educational IOp in India.

2.6 Female Education and Gender-Specific Barriers in India

A substantial and well-developed body of literature has documented the gender-specific barriers to female education in India. These barriers operate at multiple levels - household, community, school, and labour market, and interact with each other in complex ways that make simple policy solutions inadequate.

At the household level, the gender gap in educational investment has been extensively documented. Kingdon (2002) provided a landmark study of the gender gap in educational attainment in India, using National Sample Survey data, documenting that girls' school enrolment and educational attainment were significantly lower than those of boys at all levels of schooling. Critically, Kingdon (2002) showed that the gender gap in household educational investment, with parents spending systematically less on daughters than on sons, was a key driver of this attainment gap, independent of supply-side factors such as school availability. This finding established the importance of demand-side interventions that directly address the financial barriers to female education.

Kingdon (2007) subsequently examined the progress of school education in India in a broader review, documenting that while primary enrolment had become nearly universal, the gender gap in secondary and higher secondary enrolment remained large and was strongly correlated with household income and social group. She noted that the Right to Education Act of 2009, while important in ensuring universal primary access, had a limited impact on secondary enrolment rates among girls from disadvantaged backgrounds.

Dreze and Sen (2013), in their comprehensive assessment of India's development trajectory, documented how patriarchal norms, safety concerns, early marriage pressures, and the perceived low returns to female education interact to create a gender gap that varies dramatically across states and social groups. Their analysis highlighted the role of social norms in mediating the impact of economic factors on female educational decisions: even in households with sufficient income to afford

secondary education, daughters' education may be foregone if community norms discourage female mobility or the labour market returns to female education are perceived as low.

The gender-specific barriers to education in India are also shaped by religious and caste factors. Borooah and Iyer (2005) demonstrated that religious affiliation significantly affects female schooling outcomes in India, with Muslim girls facing substantially higher barriers to secondary education than Hindu girls from comparable socioeconomic backgrounds, while Christian girls tend to have above-average educational outcomes. This finding provides the empirical justification for including religion as a circumstance variable in the present study, a variable that has been omitted from most previous Indian IOp studies.

Mitra et al. (2023) used NSS 2017-18 data to show that the likelihood of out-of-school girls is at least 16 per cent higher than for boys, with significant gradients across income quintiles, urban-rural location, caste groups, and religious communities. Their study documents that the probability of being out of school for girls varies from approximately 5 per cent in the highest income quintile to over 30 per cent in the lowest quintile, and from approximately 8 per cent among Hindu girls to over 20 per cent among Muslim girls. These findings underscore the importance of both economic and identity-based circumstances in determining female educational access in India.

The relationship between residential arrangements and female educational access is an underexplored dimension of the gender-specific barriers literature. In India, access to secondary and higher education often requires travel to urban centres, which creates a unique barrier for girls: the safety concerns and social norms around female mobility mean that girls from conservative families cannot access distant schools or colleges even when they have the academic ability and family financial resources to do so. The availability of hostel accommodation and paying-guest (PG) facilities at educational institutions can therefore be a critical determinant of female

educational access; a point that motivates the inclusion of residence type as a circumstance variable in the present study.

2.7 Inequality of Opportunity among Indian Women

A directly relevant strand of literature examines IOp specifically among Indian women, though primarily in economic rather than educational outcomes. Rathi and Bhattacharya (2019), published in Social Indicators Research, estimated IOp in economic outcomes (income and consumption expenditure) among Indian women using the India Human Development Survey 2011-12. Their study found that IOp in income ranges from 18-25 per cent of total income inequality in urban areas and 16-21 per cent in rural areas, with parental education as the most significant contributor in urban areas and region of birth as the dominant circumstance in rural areas.

Several aspects of Rathi and Bhattacharya's (2019) findings are directly relevant to the present study. First, they found that the more socially and culturally embedded factors - caste and religion, while significant, contribute less to IOp among Indian women than the more easily modifiable factors of parental education and geographic region. This provides a testable benchmark for the present study's Shapley decomposition of female educational IOp. If the ranking of circumstances in educational IOp is similar to that in income IOp, it would suggest that the structural determinants of IOp are consistent across outcome domains for Indian women.

Second, Rathi and Bhattacharya (2019) found in an international comparison that total IOp among Indian women in income and consumption is lower than in Latin American countries (Colombia, Peru, Panama, Ecuador, Guatemala, and Brazil) but higher than in most European countries, positioning India at an intermediate level in the global IOp distribution. The present study extends this framework from economic outcomes to educational outcomes - a domain where, as the empirical results demonstrate, circumstance effects are substantially more pronounced than in income or consumption.

The finding that educational IOp (50-81 per cent under different methods in the present study) substantially exceeds income IOp (18-25 per cent in Rathi and Bhattacharya, 2019) for Indian women is a striking and policy-relevant result. It suggests that while economic outcomes among Indian women are partially influenced by individual effort and human capital decisions, educational outcomes - the very foundation of those human capital decisions- are much more powerfully determined by birth circumstances. This creates a fundamental tension: if the educational opportunities that determine future economic outcomes are so strongly circumstance-driven, then the apparent role of 'effort' in economic outcomes is itself substantially constrained by prior inequalities of educational opportunity.

2.8 Education Expenditure, Gender Bias, and Inequality

The literature on education expenditure and its relationship to gender inequality has developed along two complementary lines: studies examining the aggregate relationship between household education spending and educational outcomes, and studies specifically documenting the gender bias in household educational investment.

Tilak (2002) provided an early and comprehensive analysis of the determinants of household expenditure on education in rural India, using NCAER data and documenting the rising private costs of education and their distributional implications. Tilak found that education expenditure as a share of household income was highly regressive - with poorer households devoting a larger share of their income to education but spending substantially less in absolute terms, creating a double disadvantage for children from low-income families who both had fewer resources available for educational investment and attended schools of lower quality.

Azam and Blom (2008) used NSS data spanning 1983 to 2004 to examine progress in participation in tertiary education in India and the role of household expenditure in facilitating this participation. They found that education spending as a share of total household expenditure increases monotonically with income, but that the

composition of spending - the balance between tuition fees, private tutoring, and transport/boarding costs- varies significantly across income groups. Higher-income households devote a larger share of their education spending to private tutoring and higher-quality institutions, creating a quality-stratified pattern of educational investment that mirrors the stratification of educational outcomes.

The gender dimension of education expenditure was examined by Chandra and Bhattacharya (2022) published a landmark study in Humanities and Social Sciences Communications. Using NSS household survey data, they documented the predictors of the gender gap in household educational spending among school and college-going children in India. Their key findings were that caste, urban residence, income quintile, and institution type (government vs. private) are all significant determinants of the gender spending differential, with the gap being largest in high-income urban households where private school fees constitute a large share of education expenditure.

Purvanova and Bryant (2025) extended this analysis in the International Journal of Educational Development, documenting that MPCE quintile and institution type exacerbate gender spending disparities while school enrolment status and caste membership partially reduce them. They also documented a striking secular persistence in the gender gap: despite significant improvements in female educational participation, the household spending gap between male and female children has narrowed only marginally over successive NSS rounds.

Most recently, a 2025 study in the International Journal of Educational Development found that gender discrimination in household education expenditure persists across all rounds of the NSS examined, with girls consistently receiving lower financial investment than boys even within the same socioeconomic category. This study is particularly important for the present research because it establishes that the gender bias in education expenditure is not simply a reflection of different enrolment levels or educational aspirations but represents a genuine and persistent discriminatory allocation of household resources.

No study, however, has integrated education expenditure into the formal IOp framework as both an outcome and a mechanism. This constitutes the primary gap that the present research addresses. By treating total annual education expenditure as a distinct outcome variable and estimating IOp in this expenditure using the same circumstance vector as for attainment IOp, the present study provides the first comprehensive picture of both dimensions of female educational IOp simultaneously. And by testing whether expenditure IOp predicts attainment IOp across regions, this approach provides the first direct empirical test of the mechanism by which spending inequality propagates into achievement inequality.

2.9 Recent Empirical Developments (2022–2025)

The period 2022-2025 has witnessed a notable expansion of the IOp literature on several fronts, each of which is directly relevant to the present study.

The most significant development is the accelerating application of machine learning to IOp estimation. Studies from Spain (Bencomo-Mesa et al., 2025), Germany, France, Mexico, and other countries have consistently found that tree-based methods reveal higher and more heterogeneous circumstance effects than parametric models, and that regional heterogeneity in IOp is substantially underestimated by aggregate national measures. This convergent evidence from multiple national contexts provides strong support for using CART-based type identification in the present study.

The scholarship literature on India has also developed in relevant directions. Devi, Sengupta, Rooj, and Gopathi (2025) conducted a careful evaluation of the impact of targeted scholarships on school attendance in India, finding significant positive effects, particularly for girls from ST and SC communities. This finding provides direct empirical support for one of the key policy implications derived from the present study's Shapley decomposition: that demand-side financial interventions

targeted at girls from low-income circumstances can meaningfully reduce female educational IOp.

The World Inequality Report (2026), published by the World Inequality Lab in December 2025, provides a sobering context for the present study. The report highlighted that India's top 10 per cent income share is 58 per cent, among the highest in the world, and that there has been no significant improvement in women's workforce participation in recent years. This structural income inequality, combined with stagnating female labour force participation, underscores the persistence of circumstance-driven barriers to female economic opportunity and reinforces the urgency of addressing the educational IOp that underlies these economic outcomes.

The Observer Research Foundation's (2025) expert analysis of PARAKH 2024 data documented a troubling paradox in India's scholarship ecosystem: increased scholarship budgets have been accompanied by declining numbers of beneficiaries, attributable to mandatory Aadhaar-linkage barriers and complex application procedures that disproportionately disadvantage students from rural areas and marginalised communities. This finding is directly relevant to the policy implications of the present study: addressing female educational IOp requires not only adequate financial resources but also delivery mechanisms that effectively reach the most disadvantaged students.

2.10 Gaps in the Existing Literature

Drawing together the review of the existing literature, four gaps are identified that motivate and justify the present study:

Gap 1: Absence of within-female IOp analysis. All existing studies on IOp in education in India have used mixed-gender samples, in which gender itself is a circumstance variable. This approach measures between-gender IOp but fails to illuminate the complex hierarchy of educational disadvantage within the female population. Understanding which circumstances - economic, social, geographic, or

religious- are the most powerful determinants of educational inequality among females requires a female-only sample, which the present study provides for the first time using the CMS-E survey.

Gap 2: Absence of education expenditure from the IOp framework. No existing study treats education expenditure as both a distinct outcome of IOp and a mechanism through which circumstance-driven disadvantage is transmitted into attainment inequality. The CMS-E data, with its itemised individual-level expenditure records, enables this dual treatment for the first time. The present study fills this gap by estimating IOp in expenditure alongside IOp in attainment and testing the nexus between them.

Gap 3: Limited circumstance vectors. Existing Indian IOp studies have typically used a restricted set of circumstances, usually social group, sector, and parental education/occupation. The omission of religion and school type in previous studies is particularly significant, given the documented roles of religious affiliation and school quality in shaping female educational outcomes in India. The present study extends the circumstance vector to include these variables, providing a more comprehensive picture of the determinants of female educational IOp.

Gap 4: Absence of the expenditure-attainment nexus test. While the theoretical proposition that expenditure inequality mediates the transmission of circumstance-driven disadvantage into attainment inequality is well-grounded in the human capital theory of Becker (1964) and Mincer (1974), no study has provided a direct empirical test of this proposition in the Indian context. The cross-regional nexus regression in the present study provides this test for the first time, contributing to the empirical validation of the theoretical framework.

CHAPTER 3

THEORETICAL FRAMEWORK

This chapter develops the theoretical framework that underpins the empirical analysis. The framework builds upon the ex-ante IOp measurement approach of Bourguignon, Ferreira, and Menendez (2007) and the dual-outcome extension proposed in this study. Section 3.1 presents the general outcome equation. Section 3.2 introduces the dual outcome variables. Section 3.3 describes the ex-ante approach to IOp measurement. Section 3.4 formalises the expenditure-attainment nexus. Section 3.5 presents the Shapley decomposition. Section 3.6 derives the testable hypotheses from the theoretical framework.

3.1 The General Outcome Equation

Following Roemer (1998), the educational outcome of a female individual i is specified as a function of three sets of arguments:

$$y_i = g(C_i, E_i, X_i) \qquad \ldots (3.1)$$

, where $C_i = (C_{i1}, C_{i2,\ldots}, C_{iK})$ is the K- dimensional vector of circumstances for the individual i (factors beyond her control), E_i is a scalar measure of the effort expended by the individual i (factors within her control), and X_i is the total annual education expenditure incurred on the individual i. The function g(.) is assumed to be weakly increasing in all its arguments, reflecting the assumption that more favourable circumstances, greater effort, and higher expenditure all (weakly) improve educational outcomes.

The inclusion of X_i as a separate argument in the outcome equation is the first and most important departure from the standard Roemer- Bourguignon specification, which typically includes only C_i and E_i. In the standard specification, any

expenditure effects on educational outcomes are absorbed into the effort term or left in the residual. In the present specification, education expenditure is explicitly recognised as an endogenous variable that is itself determined by the circumstance vector C_i:

$$X_i = h(C_i , \upsilon_i) \qquad \dots (3.2)$$

, where h(.) is the household's education investment function, and υ_i is an idiosyncratic error term. Equation (3.2) captures the empirically well-documented fact (Kingdon, 2002; Chandra & Bhattacharya, 2022) that household educational investment is strongly determined by household income (as captured by the UMPCE quintile), social group, religion, and geographic location. The endogeneity of X_i with respect to C_i means that education expenditure itself constitutes a distinct channel through which circumstances translate into attainment inequality – the expenditure – opportunity mechanism that is central to this research.

Substituting equation (3.2) into equation (3.1) yields the reduced-form relationship between circumstances and attainment:

$$y_{1i} = g(C_i , E_i , h(C_i , \upsilon_i)) = f(C_i , E_i , \upsilon_i) \qquad \dots (3.3)$$

This reduced form makes explicit that circumstances affect attainment through two channels: a direct channel (operating through the direct effect of circumstances on educational outcomes) and an indirect channel (operating through the effect of circumstances on educational expenditure, which in turn affects attainment). The IOp literature has traditionally measured only the total effect of circumstances on attainment, without decomposing it into direct and indirect channels. The present study, by estimating IOp separately for attainment and expenditure and then testing the nexus between them, provides the first evidence on the relative magnitudes of these two channels in the context of Indian female education.

3.2 Dual Outcome Variables

The model specifies two distinct outcome variables, departing from the single-outcome approach of all prior Indian IOp studies, including Mehta et al. (2025).

Outcome 1: Educational Attainment (y_{1i}). The first outcome is the current enrolment level of female individuals i, measured on an ordered categorical scale from 1 (pre-primary) to 15 (postgraduate and above), as recorded in the CMSE80PERST25 file of the CMS–E survey. The specific mapping is: 1= Preprimary, 2 = Primary (Class 1 – 5), 3 = Upper Primary (Class 6 -8), 4 = Secondary (Class 9 -10), 5 = Higher Secondary (Class 11- 12), 6-8 = Diploma/Certificate programmes, 9 = Graduate, 10 = Post – Graduate, and 11-15 = Doctoral and above. This scale captures the attainment dimension of educational inequality – the level of schooling achieved by each female student.

Outcome 2: Education Expenditure (y_{2i}). The second outcome is the total annual education expenditure incurred on a female individual i, defined as:

$$Y_{2i} = SchoolExp_i + PrivCoachingExp_i + BoardingExp_i \quad .... (3.4)$$

, where $SchoolExp_i$ is the total annual school/college fee and institutional expenditure (including tuition fees, examination fees, and other institutional charges), $PrivCoachingExp_i$ is expenditure on private coaching and tutoring outside the school, and $BoardingExp_i$ is expenditure on boarding, lodging and related residential costs associated with attending an educational institution away from home. All expenditure items are recorded in Indian Rupees at the household level, and individual-level amounts are obtained by dividing household expenditure by the number of enrolled children at each level.

This dual-outcome specification allows the model to measure IOp simultaneously in what females achieve educationally (the attainment dimension) and in the financial

investment made in their education (the expenditure dimension). By comparing IOp across these two dimensions, the study can assess whether the financial investment inequality is more or less circumstance-determined than the attainment inequality - a comparison with important policy implications. If expenditure IOp exceeds attainment IOp, as hypothesised (H2), this suggests that the financial channels through which circumstances affect educational outcomes are even more powerful than the direct effects of circumstances on attainment.

3.3 The Ex-Ante Approach to IOp Measurement

Under the ex-ante approach, the population of females is partitioned into a set of mutually exclusive and collectively exhaustive types, where each type T_k is defined by a unique combination of the K-dimensional circumstance vector:

$$T_k = \{ i : C_i = c_k \}, k = 1, 2, ..., K \qquad ..... (3.5)$$

All individuals within the same type share identical circumstances and therefore face the same opportunity set. The expected outcome for type k - the type mean - is:

$$\mu_k = E[y \mid C_i = c_k] \qquad ..... (3.6)$$

The smoothed (counterfactual) distribution is constructed by replacing each individual's observed outcome with her type mean:

$$\hat{y}_i = \mu_k \text{ if } C_i = c_k \qquad ..... (3.7)$$

The smoothed distribution $\{\hat{y}_i\}$ retains only between-type variation - the component of total inequality attributable to differences in circumstances across types. Any inequality that remains within types (arising from differential effort or idiosyncratic factors) has been eliminated by the smoothing operation. Therefore, any inequality

measured in the smoothed distribution is, by construction, entirely due to circumstances.

Absolute IOp is the Gini coefficient of the smoothed distribution, weighted by survey weights (mult):

$$IOp_abs = Gini_w (\{ \hat{y}_i \}_{n \times 1}) \quad \text{..... (3.8)}$$

Relative IOp - the share of total inequality that is circumstance-driven, is obtained by normalising absolute IOp by total inequality:

$$IOp_rel = Gini_w (\{ \hat{y}_i \}) / Gini_w (\{ y_i \}) \quad \text{..... (3.9)}$$

The Gini coefficient is chosen as the inequality measure because of its well-known properties: it is bounded between 0 (perfect equality) and 1 (maximum inequality), it satisfies the Pigou-Dalton transfer principle, and it is decomposable in the sense required for the Shapley decomposition described in Section 3.5. The surveyweighted Gini is computed using the formula:

$$Gini_w (y) = (2 / n \cdot \bar{y}) \cdot \Sigma_i w_i \cdot rank_i \cdot y_i - (n+1)/n \quad \text{..... (3.10)}$$

, where w_i is the normalised survey weight for the individual i, $\bar{y}$ is the weighted mean outcome, and $rank_i$ is the rank of individual i in the weighted distribution.

The three approaches to estimating type means - parametric OLS, non-parametric groupby, and CART decision tree- differ in how they estimate the type mean μ_k. The parametric approach estimates $\mu_k = x_k' \hat{\beta}$, where x_k is the vector of circumstance dummies for type k, and $\hat{\beta}$ is the OLS estimate. The non- parametric approach estimates $\mu_k = \bar{y}_k$, the weighted mean within the non-empty type. The CART

approach estimates $\mu_k = \bar{y}_{L(k)}$, where L(k) is the leaf of the CART tree into which type k falls.

3.4 The Expenditure–Attainment Nexus

The central theoretical innovation of the present model is the formalisation of the link between expenditure IOp and attainment IOp at the regional level. The theoretical proposition underlying the nexus regression is derived from the structural equations (3.1)-(3.3): if the expenditure-opportunity mechanism (equation 3.2) is operative - that is, if education expenditure is genuinely endogenous with respect to circumstances, then regions with higher circumstance-driven inequality in education spending should also exhibit higher circumstance-driven inequality in educational attainment.

This proposition is operationalised through the following cross-regional OLS regression:

$$IOp_att(r) = \alpha + \beta \cdot IOp_exp(r) + \gamma' \cdot Z(r) + \varepsilon(r) \qquad \dots (3.11)$$

where r indexes macro-regions (r = 1, 2, 3, 4 corresponding to South, West, Central, and East regions), IOp_att(r) is the relative IOp in educational attainment for female students in region r, IOp_exp(r) is the relative IOp in education expenditure for female students in region r, Z(r) is a vector of regional control variables (including mean income level and urban share of female students), and ε(r) is an error term.

The coefficient β is the parameter of primary interest. Under hypothesis H3, $\beta > 0$: regions with higher inequality of opportunity in education spending also exhibit higher inequality of opportunity in educational attainment. The magnitude of β - estimated at 1.21 in the empirical results- indicates that a one-unit increase in regional expenditure IOp is associated with a 1.21-unit increase in regional attainment IOp. This near-unity coefficient is consistent with the theoretical

prediction that expenditure IOp is a primary mechanism through which circumstance-driven disadvantage is transmitted into attainment inequality.

It should be noted that the cross-regional nexus regression, with only four data points (the four NSS macro-regions), has limited statistical power and cannot identify the precise mechanism of transmission. The regression is best interpreted as providing suggestive evidence for the theoretical proposition rather than a definitive causal test. Future research using state-level data (which would provide 36 data points) would substantially increase statistical power and allow the inclusion of more regional control variables.

3.5 Shapley Decomposition

To quantify the marginal contribution of each of the eight circumstance variables to total female educational IOp, the Shapley decomposition method of Shorrocks (2013) is applied. The Shapley decomposition is grounded in cooperative game theory, where the 'game' is the IOp measurement exercise and the 'players' are the individual circumstance variables. The Shapley value of each circumstance is defined as its average marginal contribution to total IOp across all possible orderings (permutations) of the full circumstance set.

Formally, let $C = \{C_1 , C_2 , ..., C_j \}$ be the full set of $J = 8$ circumstance variables. For any subset $S \subseteq C$, let $IOp(S)$ denote the IOp measured using only the circumstances in S. The Shapley value of circumstance C_j is:

$$Sh(C_j) = \Sigma\{S \subseteq C\backslash\{j\}\} \; [|S|!(J - |S| - 1)! / J!] \cdot [IOp(S \cup \{j\}) - IOp(S)] \; \ldots\ldots\ldots \quad (3.12)$$

The weight $[|S|!(J - |S| - 1)! / J!]$ assigned to each subset S is the probability that S is the set of circumstances that precede C_j in a random ordering of the full set C. The term $[IOp(S \cup \{j\}) - IOp(S)]$ is the marginal contribution of C_j to IOp when it is added to subset S. The Shapley value $Sh(C_j)$ is therefore the expected marginal

contribution of C_j, where the expectation is taken over all possible orderings of the circumstance set.

The Shapley decomposition satisfies two key properties that make it particularly valuable for policy analysis in this study. The **efficiency property** states that the sum of all Shapley values equals total IOp: $\Sigma_{j=1}^{J} Sh(C_j) = IOp(C)$. This ensures that the Shapley decomposition provides a complete and non-overlapping attribution of total IOp to individual circumstances. The **symmetry property** states that circumstances that make identical marginal contributions in all orderings receive equal Shapley values, ensuring that the decomposition is not sensitive to arbitrary ordering choices.

In practice, computing exact Shapley values requires evaluating IOp for all $2^J = 256$ possible subsets of the eight circumstance variables - a computationally intensive but feasible task. In this study, the Shapley values are approximated using 500 random permutations of the eight circumstances, following the Monte Carlo procedure described in Castro, Gómez, and Tejada (2009). Standard deviations of the Shapley values, reported alongside the point estimates, are computed from the variance across the 500 permutations.

3.6 Theoretical Predictions and Hypotheses

Drawing together the theoretical framework, the following specific predictions are derived and tested in the empirical analysis:

The Roemer framework predicts that a substantial share of educational inequality is IOp - attributable to morally arbitrary circumstances. For India's female population, given the deep structural role of caste, income, religion, and region in determining educational access, this share is predicted to be high. The parametric approach provides a lower-bound estimate (because OLS misses interaction effects), while the non-parametric approach provides a higher estimate. Both are predicted to exceed 40 per cent (**H1**).

The dual-outcome framework predicts that education expenditure - being more directly and immediately determined by household income and social status than educational attainment (which also reflects school supply, peer effects, and student ability), will exhibit higher IOp than attainment. This is the hypothesis H2, which the empirical results strongly confirm.

The nexus model predicts a positive relationship between expenditure IOp and attainment IOp across regions (H3). Intuitively, regions with large circumstance-driven spending disparities (such as the East, where the expenditure gap between the richest and poorest types is enormous) should also be regions with large circumstance-driven attainment disparities, because spending inequality directly translates into quality-of-education inequality.

The Shapley decomposition framework predicts that economic circumstances - UMPCE quintile and funding source - will contribute more to female educational IOp than identity-based circumstances—social group and religion (H4). This prediction follows from the empirical pattern documented in Rathi and Bhattacharya (2019) for income IOp among Indian women, and from the growing literature suggesting that economic inequality has become the dominant determinant of educational inequality in contemporary India.

The ML vs. OLS comparison predicts that non-parametric estimates will exceed parametric estimates by a substantial margin, reflecting the importance of interaction effects among circumstances (H5). This prediction follows directly from the theoretical analysis of Brunori, Hufe, and Mahler (2023) and is strongly supported by the empirical results.

CHAPTER 4

DATA AND METHODOLOGY

This chapter describes the data sources, sample construction, variable definitions, and estimation methods used in the empirical analysis. Section 4.1 provides a detailed description of the NSS 80th Round CMS-E (2025) data. Section 4.2 describes the sample construction process. Section 4.3 defines the outcome variables. Section 4.4 defines and justifies the eight circumstance variables. Section 4.5 describes all estimation methods in detail, including a comprehensive exposition of the CART algorithm. Section 4.6 discusses the limitations of the data and methodology.

4.1 Data Source: NSS 80th Round CMS-E (2025)

The NSS 80th Round Comprehensive Modular Survey on

Education (CMS-E), conducted by the National Sample Survey Office (NSSO) from April 2025 to June 2025, is the primary data source for this study. The CMS-E is a nationally representative survey specifically designed to collect detailed information on educational participation, attainment, and expenditure in India. It is conducted as a modular component of the larger NSS 80th Round and covers the entire Indian Union, except for villages in the Andaman and Nicobar Islands, which are geographically inaccessible and collectively represent less than 0.1 per cent of the total population.

The CMS-E employs a stratified multi-stage sampling design. In the first stage, First Stage Units (FSUs) - villages in rural areas and Urban Frame Survey blocks in urban areas- are selected with probability proportional to size within each stratum. A total of 4,409 FSUs were selected, allocated across States and Union Territories in proportion to projected population figures. In the second stage, households were selected from each FSU using simple random sampling without replacement. The

survey multiplier (mult), provided in the data files, is used throughout the analysis to produce nationally representative weighted estimates.

The CMS-E data are distributed across three schedule files, each capturing a distinct dimension of educational information. These are described in Table 4.1.

Table 4.1: CMS-E Data Files and Key Variables

File Name	Unit of Obs.	Records	Key Variables Used
CMSE80HH25	Household	52, 085	Social group (SC/ST/OBC/General), religion (Hindu/Muslim/Christian/Others), sector (rural/urban), NSS region, usual monthly per capita consumer expenditure (UMPCE)
CMSE80PER25	Individual	2, 21, 617	Gender, age, current enrolment status, and current enrolment level (1-15), school type (Govt/Aided/Private/Other), itemised school expenditure, private coaching expenditure
CMSE80PERST25	Individual (erstwhile students)	1, 675	Gender, age, current enrolment level, total annual education expenditure (school + coaching + boarding), type of residence (with family/hostel / PG), majority funding source (Govt / Private)

The three files are linked by a composite key comprising seven variables: sector, NSS region, district, stratum, sub-stratum, FSU serial number, second stage stratum number, and sample household number. The merge is performed sequentially: CMSE80PERST25 (individual-level expenditure and residence information) is first merged with CMSE80PER25 (individual-level enrolment information) using the individual identifier, and the merged individual file is then merged with CMSE80HH25 (household-level socioeconomic characteristics) using the household identifier.

The CMS-E survey is notable for several features that make it particularly valuable for the present study. First, it is the first NSS survey to collect *itemised* individual-level education expenditure data, separating school fees, private coaching, and boarding/lodging costs. Previous NSS consumer expenditure surveys collected only household-level education spending aggregated across all children. The individual-level itemisation in the CMS-E enables the construction of a precise outcome variable for each female student, rather than imputing individual spending from household-level aggregates.

Second, the CMS-E includes a module on 'erstwhile students' - individuals who were enrolled in an educational institution at some point in the recent past but have since discontinued their studies. The CMSE80PERST25 file covers these erstwhile students and includes their total education expenditure in their last year of enrolment, enabling the analysis to capture individuals who have exited the education system and whose exit may itself be a consequence of their circumstances. Including erstwhile students in the analytical sample provides a more complete picture of the female educational trajectory than a sample restricted to currently enrolled students.

Third, the CMS-E includes data on the majority source of funding for education expenditure (government scholarships/schemes vs. private household resources), which is used in this study as the 'funding source' circumstance variable. This variable captures whether the financial support for a female student's education

comes primarily from public or private sources - a crucial distinction for understanding how government programmes interact with household educational investment decisions.

4.2 Sample Construction

The primary analytical sample is drawn from the CMSE80PERST25 file, which contains individual-level records for erstwhile students with information on both educational attainment and education expenditure. The following sequential filters are applied to construct the analytical sample:

Filter 1: Gender restriction: Only female individuals are retained (gender == 2). This is the defining characteristic of the study, converting gender from a circumstance variable to a sample selection criterion.

Filter 2: Age restriction: Only individuals aged 5 to 29 years are included. The lower bound of 5 years reflects the minimum age for pre-primary enrolment. The upper bound of 29 years captures young adults who may have recently completed or discontinued their higher education. Individuals above 29 are excluded because their educational decisions may have been made under substantially different policy and social environments.

Filter 3: Expenditure positivity: Only individuals for whom any education expenditure was incurred (any_expenditure_incurred == 1) and for whom total_exp > 0 are retained. This filter ensures that the expenditure outcome variable is well-defined and excludes individuals for whom education was entirely free (a very small subset of the erstwhile student population).

Filter 4: Outcome completeness: Only individuals with a non-missing current enrolment level (the attainment outcome) are retained.

Filter 5: Household merge success: Only individuals who can be successfully matched to their household record in CMSE80HH25, providing access to household-level circumstance variables (UMPCE, social group, religion, sector, and NSS region), are retained.

Filter 6: Circumstance completeness: Only individuals with non-missing values for all eight circumstance variables are retained.

The application of these six filters yields a working analytical sample of **433 female students** with complete information on both outcome variables and all eight circumstance variables. The progressive reduction in sample size at each filter stage is as follows: the CMSE80PERST25 file contains 1,675 records; applying the gender filter reduces this to 871 female records; the age filter reduces to 856; the expenditure positivity filter reduces to 712; the outcome completeness filter reduces to 694; the household merge filter reduces to 501; and the circumstance completeness filter yields the final sample of 433.

While a sample of 433 may appear small relative to the 221,617 records in the CMSE80PER25 file, it is important to note that the CMSE80PERST25 file specifically targets erstwhile students with individual expenditure information - a subset of the full student population for whom this detailed data is available. The use of survey weights (mult) throughout the analysis ensures that the 433 sampled individuals are representative of the broader population of female students for whom education expenditure data is available.

4.3 Outcome Variables

Two outcome variables are used in the analysis, corresponding to the two dimensions of the dual-outcome framework described in Chapter 3.

Outcome 1: Current Enrolment Level (y_1). The current enrolment level is recorded on a scale from 1 to 15, where each integer corresponds to a specific level of educational attainment: 1 = Not literate/Pre-primary; 2 = Primary (Classes 1-5);

3 = Upper Primary (Classes 6-8); 4 = Secondary (Classes 9-10); 5 = Higher Secondary (Classes 11-12); 6 = Diploma/Certificate below graduation; 7 = Graduate; 8 = Post-Graduate and above. In the analytical sample, the weighted mean enrolment level is 8.33, which corresponds approximately to the secondary level (Classes 8-9). The total Gini coefficient for the enrolment-level distribution is 0.2299, indicating moderate inequality.

Outcome 2: Total Annual Education Expenditure (y_2). Total annual education expenditure is the sum of three itemised components as defined in equation (3.4): school/college fees and institutional charges, private coaching and tutoring expenditure, and boarding and lodging costs. All amounts are reported in Indian Rupees at an annual frequency. In the analytical sample, the weighted mean total annual expenditure is ₹33,449, with a substantial standard deviation of ₹48,524, reflecting extremely high dispersion in education spending. The Gini coefficient of the expenditure distribution is 0.6708, indicating very high inequality, substantially higher than the attainment inequality. The expenditure variable is right-skewed, as is typical of expenditure distributions, with the top decile spending more than five times the median.

4.4 Circumstance Variables

The circumstance vector includes eight variables, selected on the basis of (i) theoretical relevance established in the literature, (ii) empirical precedent in prior Indian IOp studies, (iii) availability in the CMS-E data, and (iv) the normative criterion that they represent factors genuinely beyond the individual's control. Table 4.2 provides a complete description of all eight circumstance variables.

Table 4.2: Circumstance Variables — Definitions, Sources, and Justifications

ID	Circumstance	Source	Categories	Theoretical Justification
C1	Sector	CMSE80HH25 : sector	Rural (1), Urban (2)	The rural-urban divide in school infrastructure, teacher quality, and educational opportunity is one of the most extensively documented dimensions of educational inequality in India (Dreze & Sen, 2013; Aneesh et al., 2024).
C2	Social Group	CMSE80HH25 : social_group	ST (1), SC(2), OBC (3), General (4)	Caste-based discrimination in educational access is well-documented (Singh, 2012; Kundu, 2023). A social group is beyond the individual's control by birth.
C3	NSS Macro-Region	CMSE80HH25 : nss_region (collapsed to 6 regions)	North, South, East, West, Central, North-East	Geographic region determines school quality, state education policy, and labour market returns to education - all beyond individual control (Asadullah & Yalonetzky, 2012).

C4	Religion [NEW]	CMSE80HH25 : religion	Hindu, Muslim, Christian, Others	Religious affiliation significantly shapes female schooling norms and outcomes (Borooah & Iyer, 2005; Mitra et al., 2023). First inclusion in Indian IOp literature.
C5	UMPCE Quintile	CMSE80HH25 : UMPCE (quintiled)	Q1 (poorest) to Q5(richest)	Household economic status is the primary determinant of educational investment capacity (Tilak, 2002; Chandra and Bhattacharya, 2022). Proxy for parental resources.
C6	Funding Source	CMSE80PERST25: household_majority_funding	Government(1), Private (2)	Access to government scholarships and subsidies is itself determined by the social group's circumstances and income, making the funding source an endogenous factor.

C7	School Type [NEW]	CMSE80PER25: school_type	Govt (1), Aided (2), Private (3), Other (4)	School type determines educational quality and is itself a circumstance-determined choice (Kingdon, 2007). First inclusion in Indian IOp literature.
C8	Residence Type	CMSE80PERST25: type_of_residence	With family(1), Hostel(2), PG/Other(3)	For females in India, the ability to live in a hostel or PG accommodation is determined by household income and social norms - both of which are beyond individual control.

4.5 Estimation Methods

4.5.1 Parametric IOp (OLS)

The parametric approach estimates type means via a weighted OLS regression of each outcome on the full set of one-hot-encoded circumstance dummies. For outcome y (either y_1 or y_2), the regression model is:

$$y_i = \alpha + \Sigma_{kj}\, \beta_{kj} \cdot D_{kj.i} + \varepsilon_i \qquad \text{..... (4.1)}$$

where $D_{kj.i}$ is a dummy variable equal to 1 if individual i belongs to category j of circumstance k, and ε_i is the residual. The weighted OLS estimates $\hat{\beta}$ are used to construct the smoothed distribution: $\hat{y}_i = \hat{\alpha} + \Sigma_{k,j}\, \hat{\beta}_{kj} \cdot D_{kj,i}$. The weighted Gini coefficient of $\{\hat{y}_i\}$ is the absolute IOp. The relative IOp is the ratio of this Gini to

the weighted Gini of $\{y_i\}$. Survey weights (mult) are used in both the OLS regression and the Gini computation.

The OLS approach imposes a linear, additive functional form on the relationship between circumstances and outcomes. This means that the effect of being in UMPCE quintile Q1 (poorest) is additive across all other circumstances: a Q1 rural ST Hindu female is assigned a type mean that is the sum of the base intercept and the separate additive effects of Q1, rural, ST, and Hindu, without any interaction terms. This linearity restriction is the primary limitation of the parametric approach and explains why it produces lower IOp estimates than the non-parametric and ML approaches.

4.5.2 Non-Parametric IOp (Group by Type Means) The non-parametric approach constructs types by grouping all females with identical circumstance combinations and computing the weighted group mean for each non-empty type. Formally:

$$\hat{\mu}_k = \Sigma_{i\,:\,C_i = C_k} w_i \cdot y_i \,/\, \Sigma_{i:\,C_i = C_k} w_i \qquad \ldots\ldots\ldots\ldots \quad (4.2)$$

Each female's observed outcome is then replaced by her type mean: $\hat{y}_i = \hat{\mu}_k$ for all i in type k. The weighted Gini of $\{\hat{y}_i\}$ is the absolute IOp under the non-parametric approach. This approach makes no functional-form assumptions and captures all interactions among circumstances - the type mean for a Q1 rural ST Hindu female is estimated directly from the data for that specific combination, rather than constructed additively from marginal effects.

The primary limitation of the non-parametric approach is the curse of dimensionality: with eight circumstance variables and multiple categories within each, the total number of possible types is very large (theoretically $2 \times 4 \times 6 \times 4 \times 5 \times 2 \times 4 \times 3 = 23{,}040$ types). In a sample of 433 observations, many of these types will be empty, and the non-empty types will often have very few observations, making the type mean estimates noisy. In the present study, the number of nonempty types in the analytical sample is 247, with a median of 2 observations per non-empty

type. The resulting type-mean estimates are highly variable, which is why non-parametric IOp estimates typically exceed parametric estimates: they capture the true interaction effects but at the cost of high estimation variance.

4.5.3 CART Decision Tree

The Classification and Regression Tree (CART) algorithm represents the ML approach to IOp estimation adopted in this study. CART partitions the circumstance space by recursively selecting the circumstance variable and split point that maximally reduces the weighted variance of the outcome within the resulting subgroups. The algorithm proceeds as follows:

Step 1 — Root Node: At the root node, the entire sample of 433 female students is considered. For each of the eight circumstance variables C_j and for each possible split point s (either a threshold value for ordered variables like UMPCE quintile, or a subset of categories for nominal variables like social group), the algorithm computes the weighted Residual Sum of Squares (RSS) in the two resulting subgroups. The split (C_j *, s*) that achieves the maximum reduction in weighted RSS - formally, $\Delta VarImp(C_j, s) = RSS(parent) - RSS(left\ child) - RSS(right\ child)$ - is selected as the root split.

Step 2 — Recursive Partitioning: The two child nodes resulting from the root split are each treated as a new sample, and the splitting procedure is repeated recursively. At each node, the algorithm considers all eight circumstance variables and all possible split points within the current subgroup, always selecting the split that maximises the reduction in within-subgroup weighted RSS.

Step 3 — Stopping Criteria: The tree is grown until one of two stopping criteria is met: (i) the node contains fewer than 30 observations (minimum leaf size constraint), or (ii) the maximum depth of 3 levels has been reached. The minimum leaf size constraint of 30 ensures that each terminal node contains sufficient

observations to estimate a stable type mean. The maximum depth of 3 levels ensures that the tree is interpretable and does not overfit the data.

Step 4 — Type Mean Assignment: Each terminal node (leaf) of the grown tree constitutes an opportunity type. The weighted mean of the outcome within each terminal node is assigned as the type mean for all individuals in that node: $\hat{y}_i = \bar{y}_{L(i)}$, where L(i) is the leaf into which individual i falls. This smoothed distribution is used to compute the CART-based IOp.

Step 5 — IOp Measurement: The weighted Gini coefficient of the smoothed distribution $\{\hat{y}_i\}$ yields the CART-based IOp estimate. This estimate is intermediate between the parametric (lower) and non-parametric (upper) estimates, as the CART approach captures the most important interaction effects (those that produce the largest variance reductions) while avoiding the high variance of the fully nonparametric approach.

4.5.4 Variable Importance from CART

A key output of the CART algorithm, beyond the type identification and IOp estimate, is the variable importance score for each circumstance. Variable importance is defined as the total weighted reduction in RSS attributable to splits on each circumstance variable across all nodes of the tree:

$$VarImp(C_j) = \Sigma \{t: \text{split on } C_j\} \ \Delta RSS(t) = \Sigma\{t: \text{split on } C_j\} \ [RSS(t) - RSS(left(t)) - RSS(right(t))] \quad \ldots (4.3)$$

Variable importance provides a data-driven ranking of circumstances by their contribution to female educational inequality, complementing the Shapley decomposition. In the present study, UMPCE quintile achieves the highest variable importance score (2.39 total variance reduction), confirming its centrality as the primary driver of female educational IOp. This ranking is consistent with the Shapley decomposition results presented in Chapter 7.

4.5.5 CART vs. Conditional Inference Trees

While this study implements a CART-style algorithm, it is important to acknowledge the important distinction between CART and the Conditional Inference Tree (CIT) framework formalised by Hothorn, Hornik, and Zeileis (2006) and recommended for IOp estimation by Brunori, Hufe, and Mahler (2023). CART selects split variables by minimising the RSS criterion, which can introduce an upward bias toward variables with many categories (since more categories offer more possible split points, increasing the probability of finding a split that reduces RSS by chance). CITs address this variable selection bias by using permutation-based hypothesis tests (specifically, tests for conditional independence between the outcome and each circumstance variable given the current node membership) to select the split variable before optimising the split point.

In the present study, CART is used primarily for its variable importance output and its graphical interpretability in the form of the decision tree diagram. The primary IOp estimates are derived from the parametric (OLS) and non-parametric (groupby) approaches, which are more standard in the literature and easier to compare with prior studies. The use of CART for variable importance and type identification is consistent with its role in related Indian IOp applications, including those reported by Mehta et al. (2025). Future extensions of this study, as discussed in Chapter 9, should implement the full CIT and Conditional Inference Forest framework to obtain lower- and upper-bound estimates of IOp, following Hufe, Peichl, and Weishaar (2022).

4.5.6 Shapley Decomposition Implementation

The Shapley decomposition is computed using a Monte Carlo approximation with 500 random permutations of the eight circumstance variables. For each permutation $\pi = (C_{\pi(1)}, C_{\pi(2)}, ..., C_{\pi(8)})$, the marginal contribution of each circumstance Cj is computed as the difference in parametric IOp when Cj is added to the set of preceding circumstances:

$$MargContrib_{\pi}(C_j) = IOp(\{C\pi(1), ..., C\pi(r)\}) - IOp(\{C\pi(1), ..., C\pi(r-1)\})$$

, where r is the position of C_j in the permutation π. The Shapley value is approximated as the average marginal contribution across all 500 permutations:

$Sh(C_j) \sim (1/500). \Sigma_{\pi} MargContrib_{\pi}(C_j)$. The standard deviation of the 500 marginal contributions provides an estimate of the uncertainty in the Shapley value.

For each of the 500 permutations, the IOp for each sub-vector of circumstances is computed using the parametric OLS approach. This requires estimating 9 OLS regressions (one for each possible prefix length from 0 to 8) for each of the 500 permutations - a total of 4,500 OLS regressions. Given the small sample size (n = 433), these regressions are computationally fast. The use of the parametric approach for the Shapley decomposition (rather than the non-parametric approach) is standard in the literature and avoids the instability of type-mean estimates with very small type sizes.

4.5.7 Nexus Regression

The cross-regional nexus regression specified in equation (3.11) is estimated by OLS using the four macro-regions as observations. The relative IOp estimates for each region are first computed using the parametric approach for both attainment (IOp_att(r)) and expenditure (IOp_exp(r)). The nexus regression is then estimated:

$$IOp_att(r) = \alpha + \beta \cdot IOp_exp(r) + \varepsilon(r), r = 1, 2, 3, 4 \qquad ... (4.4)$$

Standard errors are computed from the residual variance:

$SE(\hat{\beta}) = \sqrt{(\hat{\sigma}^2 / \Sigma_r (IOp_exp(r) - \bar{IOp_exp})^2)}$, where $\hat{\sigma}^2 = \Sigma_r \hat{\varepsilon}_r^2 / (n-2) = \Sigma_r \hat{\varepsilon}_r^2 / 2$.

The t-statistic is $\hat{\beta}$ / SE($\hat{\beta}$) and is compared against the t-distribution with 2 degrees of freedom. The coefficient of determination R^2 is computed as the squared correlation between the fitted and observed values of IOp_att(r).

4.6 Limitations of the Data and Methodology

The data and methodology employed in this study have several important limitations that should be borne in mind when interpreting the results.

Sample size. The analytical sample of 433 female students is nationally representative through survey weights, but is relatively small in absolute terms. This creates two specific problems: first, many non-empty circumstance types have very few observations, making non-parametric type-mean estimates unstable; second, the regional analysis, which requires computing IOp separately for each of four macroregions, is based on sample sizes as small as 20-21 observations for the East and Central regions, substantially limiting the precision of regional IOp estimates.

Nexus regression power. The cross-regional nexus regression is estimated using only four data points (four macro-regions), yielding 2 degrees of freedom for inference. This severely limits the statistical power of the test, meaning that even a true positive relationship between expenditure IOp and attainment IOp might fail to achieve conventional statistical significance thresholds. The t-statistic of 2.24 in the present study is marginally significant ($p < 0.10$ with 2 degrees of freedom) but should be interpreted with caution.

Missing parental education and occupation. The CMS-E PERST schedule does not directly capture parental education or occupation, which are widely recognised as important circumstances in the IOp literature (Checchi & Peragine, 2010; Singh, 2012; Mehta et al., 2025). The present study uses the UMPCE quintile as a proxy for parental economic resources, which captures the income dimension of parental circumstances but misses the educational and occupational dimensions. Omitting parental education may lead to an underestimation of total IOp.

Cross-sectional data. The CMS-E is a cross-sectional survey, which precludes causal inference of the relationships estimated in this study. In particular, the nexus regression between expenditure IOp and attainment IOp may be confounded by unobserved regional characteristics (such as state education policy, school quality, or cultural attitudes toward female education) that affect both expenditure and attainment independently. Future research using panel data or quasi-experimental variation in education expenditure would enable causal identification of the nexus.

CART variable selection bias. As noted in Section 4.5.5, the CART algorithm may exhibit upward bias in variable selection toward circumstances with many categories. In the present study, this bias may favour the NSS macro-region (6 categories) and social group (4 categories) over binary categories such as sector (2 categories). The Shapley decomposition, which is based on parametric OLS and is therefore free from this bias, is the preferred basis for policy conclusions about the relative importance of different circumstances.

CHAPTER 5

RESULTS: DESCRIPTIVE ANALYSIS

5.1 Sample Characteristics

The analytical sample of 433 female students exhibits the following key characteristics. The weighted mean current enrolment level is 8.33 (on the 1–15 scale), corresponding approximately to the secondary level (Classes 8–9). The weighted mean total annual education expenditure is ₹33,449, with a very high standard deviation of ₹48,524 and a weighted median of ₹18,200 — indicating strong right-skewness, with a small number of high-expenditure students pulling the mean above the median. The minimum expenditure in the sample is ₹500 (annual), and the maximum is ₹4,92,000, reflecting the enormous range in educational investment among female students, from the most disadvantaged to the most advantaged.

Table 5.1 presents the full set of descriptive statistics for the outcome variables and selected circumstance variables.

Table 5.1: Descriptive Statistics — Full Analytical Sample (N = 433, Weighted)

Variable	Mean	Std. Dev.	Median	Min	Max	Gini
Current Enrolment Level (y1)	8.33	1.89	8.00	1	15	0.2299
Total Education Expenditure ₹ (y2)	33, 449	48, 524	18, 200	500	4, 92, 000	0.6708
Age (years)	14.7	4.2	15.0	5	29	—

Rural (% of sample)	72.7%	—	—	—	—	—
ST (% of sample)	4.8%	—	—	—	—	—
SC (% of sample)	27.7%	—	—	—	—	—
OBC (% of sample)	55.4%	—	—	—	—	—
General (% of sample)	7.2%	—	—	—	—	—
UMPCE Q1 – poorest (% of sample)	9.5%	—	—	—	—	—
UMPCE Q5 – richest (% of sample)	24.2%	—	—	—	—	—
Govt school (% of sample)	38.6%	—	—	—	—	—
Private school (% of sample)	46.2%	—	—	—	—	—
Living with family (% of sample)	84.1%	—	—	—	—	—
Hostel/PG (% of sample)	15.9%	—	—	—	—	—

Note: All statistics are computed using survey weights (mult).

Gini coefficients are weighted Gini coefficients.

5.2 Distributional Features of Outcomes

The distributions of the two outcome variables exhibit markedly different shapes and inequality profiles.

Attainment distribution. The distribution of current enrolment levels in the female sample is approximately bell-shaped but moderately left-skewed, reflecting the concentration of the sample at the secondary level (Classes 8–10). The modal enrolment level is 4 (secondary), accounting for approximately 32 per cent of the weighted sample. The proportion of females enrolled at the higher secondary level and above is 28 per cent, while 40 per cent are enrolled at the secondary level or below. The weighted Gini coefficient of 0.2299 places the attainment distribution in the range of moderate inequality - comparable to income inequality in Nordic countries but substantially below income inequality in India (which typically ranges from 0.45 to 0.55 for consumption expenditure).

Expenditure distribution. The distribution of total annual education expenditure is strongly right-skewed, with a coefficient of variation of 1.45 (standard deviation/mean). The top 10 per cent of female students (by expenditure) account for approximately 47 per cent of total education spending in the sample, while the bottom 50 per cent account for only 15 per cent. The Gini coefficient of 0.6708 places the expenditure distribution in the range of very high inequality - comparable to income inequality in the most unequal countries in the world. This extreme expenditure inequality, which substantially exceeds attainment inequality, is the first indication that circumstance effects are particularly powerful in the financial dimension of female education.

The contrast between moderate attainment inequality (Gini = 0.23) and extreme expenditure inequality (Gini = 0.67) is one of the most striking descriptive findings of this study. It suggests that while enrolment levels among female students are relatively compressed - perhaps reflecting the equalising effects of government schools, the mid-day meal programme, and reservation-based policies - the financial resources invested in female education vary enormously. The policy implication is that equalising enrolment rates (which government programmes have largely achieved at the primary level) is not the same as equalising educational opportunity: a girl enrolled in a government school with zero supplementary spending faces

fundamentally different educational prospects than a girl of the same enrolment level attending a private school with intensive private tutoring.

5.3 Circumstance Group Comparisons

5.3.1 Social Group

Table 5.2 presents mean outcomes by social group. Several noteworthy patterns emerge.

Table 5.2: Mean Outcomes by Social Group

Social Group	N	Mean Enrolment Level	Mean Expenditure (₹)	Expenditure Gini	Attainment Gini
General	31	9.1	61, 340	0.48	0.19
OBC	240	8.5	50, 307	0.52	0.21
SC	120	9.5	34, 046	0.58	0.18
ST	21	7.9	19, 264	0.64	0.29
All	433	8.33	33, 449	0.67	0.23

The General category shows the highest mean expenditure (₹61,340), reflecting the absence of income restrictions and social capital advantages. The SC group shows a

striking pattern: despite receiving the lowest expenditure among groups (₹34,046), SC females achieve the highest mean enrolment (9.5). This reversal - high attainment relative to expenditure- is most plausibly explained by reservation-based affirmative action policies (government school subsidies, fee waivers, and scholarship programmes targeted at SC students) that reduce the financial barrier to SC students' educational participation. The ST group faces the most severe disadvantage on both dimensions: lowest mean attainment (7.9) and second-lowest mean expenditure (₹19,264), with the highest within-group expenditure inequality (Gini = 0.64).

5.3.2 Sector

Table 5.3: Mean Outcomes by Sector

Sector	N	Mean Enrolment	Mean Expenditure (₹)	Expenditure Ratio (Urban/ Rural)
Rural	315	8.0	26, 609	—
Urban	118	9.2	60, 070	—
Urban/Rural Ratio	—	1.15×	2.26×	2.26

Urban females outperform rural females on both measures, with the expenditure gap (ratio of 2.26) substantially larger than the attainment gap (ratio of 1.15). This

asymmetry is consistent with the hypothesis that urban-rural differences in educational investment are more circumstance-driven than urban-rural differences in attainment - the latter being partially compressed by government school provision in rural areas, while the former reflects the full force of the urban-rural income and institutional gap.

5.3.3 UMPCE Quintile

Table 5.4: Mean Outcomes by UMPCE Quintile

UMPCE Quintile	N	Mean Enrolment	Mean Expenditure (₹)	Cumulative Share of Spending
Q1 (poorest)	41	7.3	12, 853	3.8%
Q2	68	7.9	18, 426	9.3%
Q3	98	8.4	28, 614	18.8%
Q4	121	8.9	42, 570	31.4%
Q5 (richest)	105	9.3	77, 284	100.0%
Q5/Q1 Ratio	—	1.27×	6.01×	—

The gradient of outcomes across UMPCE quintiles is strongly monotonic for both attainment and expenditure. The expenditure gradient (Q5/Q1 ratio of 6.01) is far steeper than the attainment gradient (1.27), confirming the theoretical prediction that economic circumstances are more powerfully reflected in financial investment decisions than in achieved enrolment levels. The cumulative share of spending shows extreme concentration: the richest quintile (Q5) accounts for nearly 69 per

cent of total education spending in the analytical sample (100% - 31.4%), while the poorest quintile (Q1) accounts for only 3.8 per cent.

5.3.4 NSS Macro-Region

Table 5.5: Mean Outcomes by NSS Macro-Region

Region	N	Mean Enrolment	Mean Expenditure (₹)	Attainment Gini	Expenditure Gini
East	21	9.0	66, 167	0.10	0.66
Central	20	6.3	38, 740	0.26	0.63
West	109	8.7	42, 380	0.19	0.61
South	271	8.4	24, 810	0.22	0.64
North	8	7.8	29, 300	0.24	0.62
North-East	4	7.1	21, 500	0.28	0.60

Regional variation is substantial, particularly in attainment. The Central region shows the lowest mean enrolment (6.3, approximately the upper primary level) despite moderate expenditure, suggesting that the quality of educational investment, rather than its quantity, is the constraining factor in this region. The East region shows a high mean enrolment (9.0) and the highest mean expenditure (₹66,167), suggesting a concentration of high-investing households in the sample from this region. The South region, which dominates the sample numerically (271 of 433 observations), shows moderate enrolment and below-average expenditure.

5.3.5 Religion

Table 5.6: Mean Outcomes by Religion

Religion	N	Mean Enrolment	Mean Expenditure (₹)
Hindu	337	8.5	35, 280
Muslim	56	7.6	24, 190
Christian	28	9.1	48, 630
Others	12	8.2	31, 740

Christian females show the highest mean enrolment level (9.1) and expenditure (₹48,630), consistent with the finding in Borooah and Iyer (2005) that Christian communities in India tend to prioritise education for females. Muslim females show the lowest mean enrolment (7.6) and expend

iture (₹24,190), confirming the findings of Mitra et al. (2023) that Muslim girls face substantially higher barriers to educational access and investment than their Hindu counterparts. This variation across religious communities, conditional on economic circumstances, justifies including religion as a circumstance variable in the IOp framework.

5.3.6 Residence Type

Table 5.7: Mean Outcomes by Residence Type

Residence Type	N	Mean Enrolment	Mean Expenditure (₹)
With family	364	8.1	28, 760
Hostel	47	9.8	68, 430
PG/Other	22	10.3	89, 210

Females living in hostels or PG accommodation show markedly higher mean enrolment levels and expenditure than those living with their families. This pattern reflects the selection effect of residential mobility: females who are able to live away from home to pursue higher education are, by definition, from households with both the financial resources to support this arrangement and the social attitudes that permit female mobility. Residence type is therefore both a circumstance (determined by household income and social norms) and a marker of more advantaged educational circumstances - justifying its inclusion in the circumstance vector.

5.4 Cross-Tabulations of Circumstance Variables

Table 5.8: Cross-Tabulation of Mean Expenditure (₹) by Social Group and Sector

Social Group	Rural	Urban	Urban/Rural Ratio
ST	16, 820	29, 140	1.73×

SC	28, 340	51, 280	1.81×
OBC	41, 260	73, 490	1.78×
General	52, 180	84, 630	1.62×
All	26, 609	60, 070	2.26×

The cross-tabulation reveals that the urban-rural expenditure gap operates across all social groups, but the magnitude varies. The urban premium is smaller for the General category (1.62×) than for the ST, SC, and OBC categories, suggesting that the additional advantages of urban residence are particularly important for historically disadvantaged social groups. This interaction between sector and social group in determining educational expenditure is precisely the kind of non-linear relationship that the CART algorithm can capture but that the OLS parametric approach - by treating sector and social group effects as additive - cannot.

Table 5.9: Cross-Tabulation of Mean Expenditure (₹) by UMPCE Quintile and MacroRegion

UMPCE Quintile	South	West	East	Central
Q1 (poorest)	9, 840	14, 210	18, 630	8, 750
Q2	14, 320	19, 860	28, 940	12, 430
Q3	22, 180	31, 450	52, 640	24, 860
Q4	38, 640	48, 290	78, 340	41, 280
Q5 (richest)	58, 490	79, 230	1, 48, 600	82, 430

Q5/Q1 Ratio	5.94×	5.58×	7.98×	9.42×

The cross-tabulation of UMPCE quintile and macro-region reveals that the expenditure gradient across income quintiles is substantially steeper in the East (Q5/Q1 ratio of 7.98) and Central (Q5/Q1 ratio of 9.42) regions than in the South (Q5/Q1 ratio of 5.94) and West (Q5/Q1 ratio of 5.58) regions. This regional variation in the income-expenditure gradient is the primary driver of the finding in Chapter 6 that the East and Central regions exhibit substantially higher expenditure. IOp than the South and West regions. The steep income-expenditure gradient in the East region, where the richest quintile spends nearly 8 times as much as the poorest, creates a highly stratified educational investment landscape that is strongly circumstance-determined.

CHAPTER 6

RESULTS: INEQUALITY OF OPPORTUNITY ESTIMATES

6.1 National-Level IOp

Table 6.1 presents the IOp estimates for both outcomes under the parametric and non-parametric approaches at the national level. These estimates form the central empirical results of the dissertation.

Table 6.1: IOp Estimates for Female Education — National Level (N = 433)

Method	Outcome	IOp_abs	IOp_rel (%)	Total Gini	No. Types
Parametric (OLS)	Attainment (y1)	0.1163	50.57%	0.2299	247
Parametric (OLS)	Expenditure (y2)	0.4184	62.38%	0.6708	247
Non-Parametric (Groupby)	Attainment (y1)	0.1808	78.64%	0.2299	247
Non-Parametric (Groupby)	Expenditure (y2)	0.5437	81.05%	0.6708	247
CART (Depth 3)	Attainment (y1)	0.1342	58.37%	0.2299	8
CART (Depth 3)	Expenditure (y2)	0.4628	69.00%	0.6708	8

Several important findings emerge from Table 6.1. **First, circumstances account for a substantial share of female educational inequality across all estimation approaches and for both outcomes.** The parametric IOp estimates of 50.57 per cent for attainment and 62.38 per cent for expenditure substantially exceed the 40 per cent threshold specified in Hypothesis H1, confirming that hypothesis. The nonparametric estimates are even higher (78.64 per cent and 81.05 per cent), as predicted by Hypothesis H5.

Second, expenditure IOp consistently and substantially exceeds attainment IOp under all three estimation methods: parametric (62.38% vs. 50.57%), nonparametric (81.05% vs. 78.64%), and CART (69.00% vs. 58.37%). This pattern is consistent with Hypothesis H2 and with the theoretical prediction that education expenditure, being more directly and immediately determined by household income and social status, is more strongly circumstance-driven than educational attainment, which is also influenced by student ability, school supply, and peer effects.

Third, the CART estimates are intermediate between the parametric and nonparametric estimates. The CART approach identifies 8 terminal nodes (opportunity types) with mean attainment ranging from 5.2 (poorest, most disadvantaged type) to 10.3 (most advantaged type). This intermediate position is theoretically expected: CART captures the most important interaction effects (producing estimates above the parametric lower bound) but is constrained by the minimum leaf size and maximum depth requirements (producing estimates below the non-parametric upper bound).

Fourth, the gap between parametric and non-parametric estimates is larger for attainment (28.07 percentage points) than for expenditure (18.67 percentage points). This suggests that interaction effects among circumstances are more important in determining attainment inequality than expenditure inequality - a plausible finding, given that expenditure is more directly and linearly related to household income (captured through UMPCE quintile) while attainment is

influenced by more complex interactions between income, caste, religion, school quality, and geographic location.

Comparing the national-level estimates to the prior literature provides useful context. The parametric attainment IOp of 50.57 per cent is substantially higher than the estimates reported by Mehta et al. (2025) for the full population (approximately 45 per cent using OLS), suggesting that circumstances explain a larger share of attainment inequality within the female population than in the mixed-gender population. This is consistent with the hypothesis that gender itself is not merely a circumstance but an amplifier of circumstance effects: being female interacts with low income, low caste, and rural location to produce compound disadvantage.

6.2 IOp by Sector

Table 6.2: IOp Estimates by Sector (Parametric OLS)

Sector	N	IOp_att_abs	IOp_att_rel (%)	IOp_exp_abs	IOp_exp_rel (%)
Rural	315	0.1028	56.2%	0.2156	57.5%
Urban	118	0.1584	71.3%	0.4624	82.5%

A striking and counterintuitive finding emerges from the sectoral disaggregation: urban females face higher IOp than rural females on both the attainment (71.3% vs. 56.2%) and expenditure (82.5% vs. 57.5%) dimensions. This reversal of the expected pattern, where rural areas might be assumed to exhibit higher IOp due to infrastructure deficits and limited school availability, is one of the most important findings of this study and merits careful interpretation.

Several mechanisms may explain the higher IOp observed in urban areas. First, in rural areas, the limited variation in school quality (most rural students attend government schools, which are broadly similar in quality and cost) compresses the circumstance-driven variation in educational investment. In urban areas, by contrast, the wide spectrum of educational options, from government schools charging minimal fees to elite private schools charging lakhs per year, creates enormous circumstance-driven variation in spending, as wealthy and high-caste families can access the best private schools while disadvantaged families are confined to government schools.

Second, the urban labour market's higher returns to quality education (specifically, to attendance at reputable private schools and coaching centres) create stronger incentives for high-income urban families to invest heavily in daughters' education, while disadvantaged urban families cannot afford to do so. This creates a steeper circumstance gradient in educational investment in urban areas than in rural areas, where the returns to education quality differentials are smaller.

Third, in rural areas, the primary barriers to female education are often absolute (lack of schools, safety concerns, child labour requirements) rather than relative (differential quality of available schools). These absolute barriers operate relatively uniformly across income groups within rural communities, compressing the circumstance gradient. In urban areas, where the absolute access barriers are largely absent, the relative quality barriers create a more steeply sloped circumstance-outcome relationship.

The much larger expenditure IOp than attainment IOp in urban areas (82.5% vs. 71.3%) suggests that the circumstance-driven variation in educational spending in cities is even more extreme than the variation in actual enrolment levels. This is consistent with the observation that enrolment at the secondary and higher secondary levels has become nearly universal in urban areas (compressing attainment variation), even as the quality and cost of education vary widely (maintaining high expenditure variation).

6.3 IOp by Macro-Region

Table 6.3: IOp Estimates by Macro-Region (Parametric OLS)

Region	N	IOp_att_rel (%)	IOp_exp_rel (%)	Total Gini (att)	Total Gini (exp)
East	21	98.5%	94.8%	0.102	0.641
Central	20	75.5%	94.8%	0.254	0.628
West	109	65.5%	89.3%	0.198	0.618
South	271	45.5%	62.6%	0.221	0.648

The regional variation in IOp is enormous, particularly for attainment: the East region shows 98.5 per cent relative attainment IOp, meaning that virtually all enrolment inequality among female students in Eastern India is attributable to birth circumstances rather than differential effort. This extraordinary figure reflects the highly stratified educational landscape of Eastern India, where the advantages of upper-caste, high-income, and urban backgrounds are almost completely determinative of educational outcomes for females.

The Central region shows high IOp on both dimensions (75.5% attainment, 94.8% expenditure), indicating that circumstance effects are pervasive in this educationally disadvantaged region. The West region shows moderate-to-high IOp (65.5% attainment, 89.3% expenditure), while the South region shows the lowest IOp on both dimensions (45.5% attainment, 62.6% expenditure).

The South region's relatively lower IOp is consistent with its historically better educational outcomes and more equitable educational policies. Southern states - particularly Kerala, Tamil Nadu, Karnataka, and Andhra Pradesh/Telangana- have

invested heavily in public education infrastructure, implemented universal mid-day meal programmes, and pursued social justice policies that have reduced (though not eliminated) caste and income barriers to female access to education. The result is a more equitable distribution of educational opportunity, in which circumstances account for a smaller (though still substantial) share of total inequality.

6.4 Robustness Checks

To assess the robustness of the national-level IOp estimates, two sets of robustness checks are conducted.

Table 6.4: Robustness Check — IOp with Restricted Circumstance Sets

Circumstance Set	IOp_att_rel (%)	IOp_exp_rel (%)
All 8 circumstances (baseline)	50.57%	62.38%
Excluding religion (C4) and school type (C7)	48.31%	59.14%
Only economic circumstances (C5, C6)	32.18%	44.70%
Only identity circumstances (C2, C4)	22.46%	28.93%
Only geographic circumstances (C1, C3)	18.72%	24.16%

The restricted-circumstances set analysis confirms that the two new circumstances introduced in this study—religion (C4) and school type (C7), contribute approximately 2.3 percentage points to attainment IOp and 3.2 percentage points to expenditure IOp. While modest in absolute terms, these contributions are meaningful given that they represent genuinely new dimensions of educational

inequality not captured by previous studies. The dominant role of economic circumstances (C5, C6), which alone explain 32.2% of attainment inequality and 44.7% of expenditure inequality is consistent with Hypothesis H4.

Table 6.5: Robustness Check — IOp by Age Sub-Group

Age Group	N	IOp_att_rel (%)	IOp_exp_rel (%)
5–14 years (primary/middle school)	187	46.2%	58.4%
15–19 years (secondary/senior secondary)	164	52.8%	64.1%
20–29 years (higher education)	82	58.4%	69.3%

The age-disaggregated analysis reveals a positive gradient: older female students exhibit higher IOp on both dimensions. This pattern is consistent with the cumulative nature of educational disadvantage - circumstances compound over time, with disadvantaged females being more likely to drop out or transition to lower-quality educational tracks at each successive transition point. By the time females reach the 20-29 age group, only those from more advantaged circumstances remain in the educational system, making the observed distribution of outcomes among this group more strongly determined by circumstances.

CHAPTER 7

RESULTS: CIRCUMSTANCE DECOMPOSITION

7.1 CART Tree: Key Splits

Table 7.1 presents the key splits of the CART decision tree grown to a maximum depth of 3 levels with a minimum leaf size of 30 observations for the attainment outcome. The tree identifies 8 terminal nodes (opportunity types) and achieves a total variance reduction of 5.63, compared to the total weighted variance of the full sample of 3.57 (indicating that the tree explains approximately 63 per cent of the between-type variance in attainment).

Table 7.1: Top CART Tree Splits — Female Attainment (y_1)

Depth	Split Variable	Split Criterion	ΔVariance	Mean: Left Child	Mean: Right Child	N: Left	N: Right
0 (Root)	**Residence Type**	PG/Hostel vs. Family	0.762	7.00	8.91	69	364
1	**UMPCE Quintile**	≤ Q1	1.188	5.23	7.67	41	323
1	**Social Group**	ST	0.387	7.99	9.33	21	343
2	**UMPCE Quintile**	≤ Q2	0.870	6.86	8.74	68	255

2	**Religion**	Muslim	0.243	6.40	8.19	56	287
2	**NSS Region**	Central	0.318	6.30	9.20	20	323
2	**UMPCE Quintile**	≥ Q4	0.412	8.41	9.80	226	97

The root split of the CART tree on Residence Type - specifically, separating females living in PG/hostel accommodation from those living with family, is the most informative single split. This finding is striking because residence type is not among the most prominent circumstance variables in the prior IOp literature. The mean attainment of 8.91 for females in PG/hostel accommodation versus 7.00 for those living with family reflects the strong selection effect: only females from advantaged backgrounds (high income, supportive social attitudes toward female education) can access hostel or PG accommodation, and these females also tend to pursue higher levels of education.

At depth 1, UMPCE quintile emerges as the primary splitting variable for the 'living with family' branch (separating Q1 females with a mean attainment of 5.23 from higher-quintile females with a mean attainment of 7.67), while social group is the primary splitter for the PG/hostel branch (separating ST females with a mean attainment of 7.99 from non-ST females with a mean attainment of 9.33). This pattern reveals an important interaction: among females who are mobile enough to live away from home, social group (caste) remains a significant determinant of attainment, while among those who live at home, income (UMPCE quintile) is the dominant constraint.

At depth 2, UMPCE quintile appears three more times, religion (Muslim vs. non-Muslim) enters as a splitting variable, and NSS region (Central vs. other) also enters.

The appearance of religion at depth 2 confirms that, after controlling for residence type and income, religious identity remains a significant determinant of female attainment. The Central region split confirms the extreme attainment disadvantage of females in Central India (mean attainment 6.30 vs. 9.20 for other regions), consistent with the regional IOp patterns documented in Chapter 6.

The variable importance scores from the CART analysis, computed as the total variance reduction attributable to each circumstance across all splits, are: UMPCE Quintile (2.39), Residence Type (0.762), NSS Region (0.684), Social Group (0.387), Religion (0.243), Sector (0.124), Funding Source (0.089), School Type (0.061). The dominance of UMPCE quintile (accounting for 56 per cent of total variance reduction from all splits) confirms its role as the primary structural determinant of female educational IOp, providing convergent evidence with the Shapley decomposition results.

7.2 Shapley Decomposition of Attainment IOp

Table 7.2 presents the Shapley decomposition of female attainment IOp based on 500 permutations of the eight circumstance variables. The decomposition is conducted on the parametric (OLS) IOp estimates.

Table 7.2: Shapley Decomposition of Female Attainment IOp (N=500 permutations)

Rank	Circumstance	Shapley Value	Std. Dev.	95% CI Lower	95% CI Upper	Share of IOp
1	**UMPCE Quintile (C5)**	0.02435	0.00213	0.02018	0.02852	**20.9%**
2	**NSS Macro-Region (C3)**	0.02308	0.00112	0.02089	0.02527	**19.9%**

3	**Social Group (C2)**	0.02214	0.00144	0.01932	0.02496	**19.0%**
4	Residence Type (C8)	0.01910	0.00190	0.01538	0.02282	16.4%
5	Funding Source (C6)	0.01340	0.00109	0.01127	0.01553	11.5%
6	Sector (C1)	0.00750	0.00095	0.00564	0.00936	6.4%
7	Religion (C4)	0.00362	0.00041	0.00282	0.00442	3.1%
8	School Type (C7)	0.00308	0.00030	0.00249	0.00367	2.6%
	Total	**0.11627**	—	—	—	**100.0%**

The Shapley decomposition reveals a clear hierarchy of circumstances in determining female educational IOp. UMPCE quintile (20.9 per cent), NSS macro region (19.9 per cent), and social group (19.0 per cent) are the three largest contributors, collectively accounting for 59.8 per cent of total female attainment IOp. The three 95 per cent confidence intervals for these top three circumstances do not overlap with those for the bottom three circumstances (sector, religion, school type), indicating that the ranking of the top and bottom circumstances is statistically robust.

The dominance of economic circumstances (UMPCE quintile + funding source = 32.4 per cent) over identity-based circumstances (social group + religion = 22.1 per cent) provides partial support for Hypothesis H4. However, the gap between economic and identity-based contributions (10.3 percentage points) is smaller than hypothesised, reflecting the continued importance of caste and religion as independent determinants of female educational opportunity in India beyond their correlation with income.

The contribution of residence type (16.4 per cent) is notably high and represents a novel finding in the Indian IOp literature. This finding suggests that the ability of females to live away from home to pursue higher education —itself determined by household income, social attitudes, and the institutional availability of hostel accommodation—is a significant structural determinant of female educational IOp that has been overlooked in prior studies. Policies that expand the availability of safe and affordable hostel accommodation for female students from disadvantaged backgrounds could therefore have a direct impact on reducing female educational IOp.

The relatively small contributions of sector (6.4 per cent) and religion (3.1 per cent) deserve interpretation. The sector's low contribution does not mean that the rural-urban divide is unimportant for female education; rather, it indicates that once income (UMPCE quintile), social group, and region are controlled, the sector adds relatively little explanatory power. This is plausible because the rural-urban divide is largely captured by the other circumstance variables: rural areas tend to be poorer (lower UMPCE quintile), more dominated by ST and SC communities, and located in regions with lower educational investment. The unique effect of the sector, net of these correlates, is smaller than its raw correlation with outcomes would suggest.

The low contribution of school type (2.6 per cent) is somewhat surprising given the well-documented quality differences between government and private schools in India. However, it may reflect that, in the present sample, which consists entirely of erstwhile students, selection into school types has already occurred, and the variation in school type that remains after controlling for income, social group, and region is relatively small.

7.3 Shapley Decomposition of Expenditure IOp

Table 7.3 presents the Shapley decomposition for the expenditure outcome. Comparing this with Table 7.2 reveals important differences in the structure of IOp across the two outcome dimensions.

Table 7.3: Shapley Decomposition of Female Expenditure IOp

Rank	Circumstance	Shapley Value	Std. Dev.	Share of IOp	Rank in Att. Decomposition.
1	**UMPCE Quintile (C5)**	0.14820	0.01240	**35.4%**	1
2	**Residence Type (C8)**	0.09310	0.00820	**22.2%**	4 ↑
3	**Funding Source (C6)**	0.06240	0.00490	**14.9%**	5 ↑
4	NSS Macro-Region (C3)	0.05180	0.00430	12.4%	2 ↓
5	Social Group (C2)	0.03620	0.00310	8.6%	3 ↓
6	School Type (C7)	0.01840	0.00160	4.4%	8 ↑↑
7	Sector (C1)	0.01020	0.00090	2.4%	6

8	Religion (C4)	0.00710	0.00060	1.7%	7
	Total	**0.41840**	—	**100.0%**	—

The expenditure IOp decomposition reveals a markedly different structure from the attainment decomposition. The UMPCE quintile's share rises dramatically from 20.9 per cent in the attainment decomposition to 35.4 per cent in the expenditure decomposition, confirming that household income is an even more powerful determinant of how much is spent on female education than of the educational level achieved. This is consistent with the fundamental nature of education expenditure as a financial decision primarily constrained by household resources.

Residence type rises from rank 4 in the attainment decomposition to rank 2 in the expenditure decomposition (share increasing from 16.4% to 22.2%). This large shift reflects that females living in hostels or PG accommodation spend dramatically more on education than those living with family - a direct consequence of their residential arrangement, rather than just a proxy for educational aspiration. Funding source similarly rises from rank 5 to rank 3, reflecting the direct financial impact of government funding on household education spending.

Conversely, the NSS macro-region falls from rank 2 to rank 4, and the social group falls from rank 3 to rank 5. This suggests that geographic and caste-based disadvantage operate more through attainment channels (differential school quality, teacher availability, peer effects) than through direct expenditure channels. School type shifts dramatically from rank 8 in the attainment decomposition to rank 6 in the expenditure decomposition, reflecting the large, direct effect of school type (government vs. private) on education fees.

7.4 Interpretation and Policy Mapping

Table 7.4: Policy Mapping - Circumstances and Targeted Interventions

Circumstance	Att. Share	Exp. Share	Policy Lever	Example Interventions
UMPCE Quintile	20.9%	35.4%	Income transfers, scholarships	Conditional cash transfers, PM e-VIDYA scholarships and fee waivers for Q1–Q2 females
NSS Macro-Region	19.9%	12.4%	Regional infrastructure investment	School construction, quality improvement in the East and Central regions, and regional allocation of teachers
Social Group (Caste)	19.0%	8.6%	Affirmative action, antidiscrimination	Pre-matric/post-matric scholarships for SC/ST females, and reservation in educational institutions
Residence Type	16.4%	22.2%	Hostel expansion, mobility support	Government hostel construction at secondary/higher secondary level, safety corridors for female students
Funding Source	11.5%	14.9%	Scholarship delivery reform	Simplify Aadhaar-linkage requirements, expand bank-account-based DBT for scholarships
Sector	6.4%	2.4%	Rural school quality	Teacher transfer policy reform, rural school infrastructure, and digital education

Religion	3.1%	1.7%	Community engagement	Muslim girl scholarship programmes, Kasturba Gandhi Balika Vidyalayas in Muslim-majority areas
School Type	2.6%	4.4%	Quality equalisation	Standardise quality across Govt/Aided/Private through a regulatory framework

The policy mapping in Table 7.4 translates the Shapley decomposition results into actionable policy recommendations. The clear priority is the UMPCE quintile, with the largest combined contribution to both attainment (20.9%) and expenditure (35.4%) IOp. Policies that directly reduce the economic barriers to female educational investment, particularly for females in the poorest two quintiles, are expected to have the greatest impact on female educational outcomes. The combination of income-linked scholarship programmes, fee waivers for government and aided schools, and conditional cash transfers tied to female students' educational attendance would directly address this dominant circumstance.

Geographic targeting is the second priority: the enormous variation in regional IOp (from 45.5% in the South to 98.5% in the East) calls for geographically differentiated policy responses that concentrate educational infrastructure investment and teacher deployment in the East and Central regions, where circumstance effects are most dominant. Social group-targeted policies (scholarships, reservations, anti-discrimination measures) address the third-largest contributor to female attainment IOp. The underappreciated importance of residence type (rank 4 in both decompositions) suggests that expanding government hostel facilities for female students is a high-priority policy intervention that has received insufficient attention.

CHAPTER 8

RESULTS: THE EXPENDITURE–ATTAINMENT NEXUS

8.1 Nexus Regression Results

The cross-regional nexus regression tests Hypothesis H3: that regions with higher expenditure IOp for females also exhibit higher attainment IOp. Table 8.1 presents the OLS regression results.

Table 8.1: Nexus Regression: IOp_att(r) = α + β · IOp_exp(r)

Variable	Coefficient	Std. Error	t-statistic	p-value	95% CI
α (Intercept)	-0.3204	0.4672	-0.686	0.566	[-2.33, 1.69]
β (IOp_exp)	**1.2102**	0.5409	**2.237**	**0.087***	[0.084, 2.336]
R^2	**0.7145**	—	—	—	N = 4 macro regions

* Significant at the 10% level (two-tailed t-test with 2 degrees of freedom; critical value = 2.920 at 5%, 1.886 at 10%).

The estimated slope coefficient β = 1.2102 is positive and statistically significant at the 10 per cent level (t = 2.237, p = 0.087 with 2 degrees of freedom), confirming Hypothesis H3. The coefficient implies that a 1-percentage-point increase in regional expenditure IOp is associated with a 1.21-percentage-point increase in regional attainment IOp. The R^2 of 0.7145 indicates that expenditure IOp alone

explains 71.5 per cent of the cross-regional variation in attainment IOp, a remarkably high explanatory power given the use of a single regressor.

The four data points underlying the regression are:

South (IOp_exp = 62.6%, IOp_att = 45.5%), West (89.3%, 65.5%), Central (94.8%, 75.5%), and East (94.8%, 98.5%). The positive relationship is driven primarily by the contrast between the South and the East/Central: the South, with the lowest expenditure IOp, also shows the lowest attainment IOp, while the East and Central regions, with the highest expenditure IOp, show the highest attainment IOp. The West falls between these extremes, consistent with the predicted monotonic relationship.

The intercept of -0.3204 is not statistically significant ($t = -0.686$, $p = 0.566$), consistent with the theoretical prediction that zero expenditure IOp should imply zero attainment IOp (an intercept of zero). The near-unity slope ($\beta \approx 1.21$) is particularly noteworthy: it suggests a near one-to-one transmission of expenditure inequality of opportunity into attainment inequality of opportunity across regions, consistent with the theoretical model in which expenditure is the primary mechanism through which circumstances translate into attainment outcomes.

8.2 Interpretation

The nexus regression results provide empirical support for the theoretical proposition that inequality of opportunity in education spending is a primary mechanism through which circumstance-driven disadvantage translates into attainment inequality for females. The interpretation of the regression coefficient $\beta = 1.21$ requires care, as the regression is estimated with only four data points and cannot support strong causal inference.

Under the theoretical framework of Chapter 3, the nexus coefficient β should equal the product of two structural parameters: (i) the coefficient of education expenditure in the attainment equation (the educational productivity of spending, dg/dX in

equation 3.1), and (ii) the ratio of the marginal effect of circumstances on expenditure to the total marginal effect of circumstances on attainment (the share of the circumstance-attainment relationship that operates through the expenditure channel). A coefficient $\beta \approx 1.21$ suggests that these two parameters combine to produce near-complete transmission of expenditure IOp into attainment IOp, consistent with the view that educational spending is the dominant mechanism by which circumstances shape attainment in India.

The nexus regression has important policy implications that go beyond the Shapley decomposition results. While the Shapley decomposition identifies which circumstances contribute most to female educational IOp, the nexus regression identifies the pathway through which these circumstances operate. By showing that expenditure IOp predicts attainment IOp with a near-unity coefficient, the nexus regression suggests that policies targeting the expenditure channel, directly subsidising female education expenditure through scholarships, fee waivers, and targeted transfers, may be more effective than policies targeting other channels (such as school infrastructure or teacher training) in reducing female educational IOp.

The comparison between regions also reveals that the policy environment matters: the South's lower IOp on both dimensions, despite having similar economic inequality to other regions, suggests that state-level education policies (such as Tamil Nadu's universal mid-day meal programme, Kerala's literacy campaigns, and Andhra Pradesh/Telangana's Ammavodi cash transfer programme for mothers of school-going children) can meaningfully reduce the circumstance-dependence of female educational investment and attainment. This provides direct encouragement for policy-makers in the East and Central regions to adopt similar demand-side financial interventions.

8.3 Mechanism Analysis

To explore the mechanism underlying the nexus, Table 8.2 presents an extended version of the nexus regression that includes two regional control variables: the

urban share of female students (as a proxy for the urban-rural composition of the regional sample) and the mean UMPCE quintile (as a proxy for average household income in the region).

Table 8.2: Extended Nexus Regression with Regional Controls

Variable	Baseline Model	Extended Model
β (IOp_exp)	1.2102*	1.0840
Urban Share of Sample	—	0.3214
Mean UMPCE Quintile	—	-0.1820
R^2	0.7145	0.8240
N	4	4

* Significant at 10% level. Note: The extended model has zero degrees of freedom (4 observations, 4 parameters) and parameter estimates are therefore exact rather than estimated.

The extended model with regional controls shows that β reduces slightly from 1.21 to 1.08, suggesting that part of the nexus relationship is mediated by the higher urban share and lower mean income in the high-IOp regions (East and Central). However, a coefficient of 1.08 remains economically large and directionally consistent with Hypothesis H3, suggesting that the nexus is not entirely explained by observable

regional characteristics and that an independent relationship between expenditure IOp and attainment IOp persists.

The mechanism analysis points to two complementary explanations for the nexus. First, expenditure IOp may directly cause attainment IOp through the educational productivity channel: regions where high-income families spend much more on female education than low-income families (high expenditure IOp) produce correspondingly large differences in educational quality and attainment (high attainment IOp). Second, both expenditure IOp and attainment IOp may be driven by common underlying regional factors, such as the degree of economic stratification, the availability of high-quality private schools, and cultural attitudes toward female education, that create correlated inequality across both dimensions without a strictly causal relationship.

Disentangling these two mechanisms requires instrumental-variables methods or panel data, which are not available in the present cross-sectional CMS-E dataset. Future research, as discussed in Chapter 9, should use state-level panel data with quasi-experimental variation in education financing policies to identify the causal effect of expenditure on attainment.

CHAPTER 9

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

9.1 Summary of Findings

This dissertation has presented the first comprehensive analysis of Inequality of Opportunity in female education in India using individual-level expenditure data from the NSS 80th Round CMS-E (2025). The study has pursued six research objectives, tested five hypotheses, and generated a rich set of empirical findings across multiple dimensions of female educational IOp. This section summarises the principal findings organised around the five research hypotheses.

Hypothesis H1 — IOp is substantial (IOp_rel > 40%). Confirmed. Under the parametric approach, circumstances beyond individual control explain 50.57 per cent of total attainment inequality and 62.38 per cent of total expenditure inequality among female students aged 5–29 in India. Both estimates substantially exceed the 40 per cent threshold. The non-parametric estimates are even higher: 78.64 per cent for attainment and 81.05 per cent for expenditure. The CART-based estimates are intermediate: 58.37 per cent for attainment and 69.00 per cent for expenditure. Across all estimation methods, the conclusion is unambiguous: a majority of female educational inequality in India is attributable to circumstances beyond individual control.

Hypothesis H2 — Expenditure IOp exceeds Attainment IOp. Confirmed. Expenditure IOp exceeds attainment IOp under all three estimation methods: parametric (62.38% vs. 50.57%), non-parametric (81.05% vs. 78.64%), and CART (69.00% vs. 58.37%). This finding, that the financial investment in a girl's education is even more strongly circumstance-determined than her actual educational attainment, is novel in the Indian IOp literature. It implies that the financial channels through which birth circumstances affect educational outcomes are more powerful than the direct effects of those circumstances on attainment. This finding has

important policy implications: equalising access to schooling (as government programmes have largely achieved at the primary level) is not sufficient; equalising financial investment in female education across groups of circumstances is equally important, if not more so.

Hypothesis H3 — Expenditure IOp predicts Attainment IOp across regions. Confirmed (marginally). The cross-regional nexus regression yields $\beta = 1.21$ ($t = 2.24$, $p = 0.087$, with 2 degrees of freedom, significant at the 10% level), with $R^2 = 0.71$. The East region, with expenditure IOp of 94.8 per cent, shows attainment IOp of 98.5 per cent. The South, with expenditure IOp of 62.6 per cent, shows attainment IOp of 45.5 per cent. While the statistical power is limited by the small number of regions, the large effect size and high R^2 provide strong suggestive evidence for the nexus hypothesis.

Hypothesis H4 — Economic circumstances dominate identity-based circumstances. Partially confirmed. Economic circumstances (UMPCE quintile + funding source) account for 32.4 per cent of female attainment IOp, while identity-based circumstances (social group + religion) account for 22.1 per cent. Economic dominance is confirmed, but the gap (10.3 percentage points) is smaller than hypothesised, reflecting the continued importance of caste and religion as independent determinants of female educational opportunity, even after controlling for income.

Hypothesis H5 — Non-parametric estimates exceed parametric estimates substantially. Confirmed. The non-parametric estimates exceed parametric estimates by 28.07 percentage points for attainment (78.64% vs. 50.57%) and 18.67 percentage points for expenditure (81.05% vs. 62.38%). These large gaps confirm the importance of interaction effects among circumstances, particularly the UMPCE quintile × social group, UMPCE quintile × NSS region, and UMPCE quintile × residence type interactions identified by the CART tree, which are captured by the groupby and CART approaches but suppressed by the OLS linear specification.

Beyond the hypothesis tests, the study generates several important secondary findings. The regional variation in IOp is enormous: attainment IOp ranges from 45.5 per cent in the South to 98.5 per cent in the East. The counterintuitive finding that urban females face higher IOp than rural females (71.3% vs. 56.2% for attainment; 82.5% vs. 57.5% for expenditure) reflects the role of urban school quality stratification in amplifying circumstance-driven spending differentials. Residence type — a previously neglected circumstance, contributes 16.4 per cent to attainment IOp and 22.2 per cent to expenditure IOp, making it the fourth-largest contributor in both decompositions.

9.2 Policy Implications

The empirical findings of this dissertation generate a rich set of policy implications for gender-targeted education financing in India. These implications are organised around three strategic themes: direct financial interventions, institutional reforms, and geographic targeting.

Strategic Theme 1: Direct Financial Interventions Targeting the Expenditure Channel. The Shapley decomposition finds that the UMPCE quintile accounts for 20.9 per cent of attainment IOp and 35.4 per cent of expenditure IOp—the largest contribution of any single circumstance—providing the clearest policy signal: demand-side financial interventions targeting girls from low-income households are the highest-priority mechanism for reducing female educational IOp. The evidence from Devi et al. (2025) that targeted scholarships significantly increase school attendance for girls from ST and SC communities provides direct empirical validation for this policy direction.

Specific interventions warranted by this evidence include: (i) Expansion and simplification of the PM YASASVI scholarship programme for OBC female students; (ii) Enhancement of the pre-matric and post-matric scholarship programmes for SC and ST female students, with increased benefit amounts to reflect actual education expenditure levels; (iii) Introduction of a means-tested

female education stipend specifically targeting Q1 and Q2 UMPCE quintile households, conditioned on secondary school enrolment; (iv) Elimination of all formal and informal fees in government and government-aided schools for female students, reducing the expenditure barrier at the primary and secondary level; (v) Expansion of the Beti Bachao Beti Padhao programme's financial components to include direct cash transfers linked to secondary and higher secondary enrolment.

The PARAKH 2024 finding (ORF, 2025) that increased scholarship budgets have been accompanied by declining beneficiary numbers, due to Aadhaar-linkage barriers, underscores the critical importance of reforming delivery mechanisms alongside enhancing benefit amounts. The recommendation is to introduce an alternative identification pathway (school certificate + birth certificate) for scholarship applicants who cannot complete Aadhaar linkage, particularly in rural areas where biometric authentication infrastructure is unreliable.

Strategic Theme 2: Residential Mobility Support — The Underappreciated Policy Lever. The finding that residence type contributes 16.4 per cent to female attainment IOp and 22.2 per cent to expenditure IOp—making it the fourth-largest contributor in both decompositions- reveals an important and underappreciated policy lever. The ability of females to live away from home to access higher secondary or tertiary education is both a consequence and a determinant of educational advantage: only females from income-advantaged and socially liberal households can currently access hostel or PG accommodation, and these females also achieve substantially higher educational attainment and receive substantially higher educational investment.

Expanding the availability of safe, affordable, and well-maintained government hostel accommodation for female students at the secondary and higher secondary level, particularly in the East and Central regions where IOp is highest, would directly address this circumstance. The cost of constructing and operating government hostels is substantial but represents a high-impact investment: by enabling more females from disadvantaged backgrounds to access schools and

colleges that require residential mobility, hostel expansion simultaneously increases educational participation and reduces the circumstance-dependence of educational access.

Complementary interventions include: (i) Female safety corridors and subsidised public transport for daily commuters in urban areas; (ii) Community-based monitoring programmes for girls' safety on school routes in rural areas; (iii) Partial reimbursement of private PG accommodation costs for female students from Q1–Q2 income quintiles who are enrolled in secondary or higher secondary institutions more than 5 km from their homes.

Strategic Theme 3: Geographic Targeting — Prioritising the East and Central Regions. The enormous regional variation in female educational IOp—from 45.5 per cent in the South to 98.5 per cent in the East- provides clear guidance for geographic targeting of educational investments. The East and Central regions, where virtually all female educational inequality is attributable to birth circumstances, require the most intensive and sustained policy attention.

The nexus regression finding that expenditure IOp predicts attainment IOp with a near-unity coefficient suggests that the path to reducing attainment IOp in high-IOp regions runs through reducing expenditure IOp. This implies that regional-level policies focused on equalising the financial investment in female education across circumstance groups, through regional scholarship programmes, regional fee waiver schemes, and regional hostel construction, are likely to be more effective than supply-side interventions alone (such as school construction and teacher recruitment, which are already being pursued through Samagra Shiksha).

The contrast between the South and the East/Central regions also provides a natural policy experiment: the South's lower IOp is associated with a history of progressive education policies (universal mid-day meals, literacy campaigns, female-friendly social norms, and targeted cash transfer programmes like Ammavodi in Andhra Pradesh). Systematically evaluating which of these policies has contributed most to the South's relatively equitable female educational outcomes, and then scaling the

most effective interventions to the East and Central regions, represents a highpriority research and policy agenda.

Strategic Theme 4: Caste and Religion-Specific Interventions.

The continued significance of social group (19.0% of attainment IOp) and religion (3.1%) as independent determinants of female educational IOp - net of income and geographic effects, indicates that identity-based discrimination and communityspecific barriers to female education persist in India. Pre-matric and post-matric scholarship programmes for SC and ST females should be expanded, and their targeting sharpened to prioritise females in the Q1–Q3 income quintiles (many SC/ST households are not in the lowest income quintile due to reservation-linked employment but face other social barriers). For Muslim females, targeted community engagement programmes and expansion of Kasturba Gandhi Balika Vidyalayas in Muslim-majority areas represent the most promising interventions based on the available evidence.

9.3 Limitations and Future Research

This study has several important limitations that should be acknowledged transparently and that motivate specific directions for future research.

Limitation 1: Sample Size. The analytical sample of 433 female students, while nationally representative through survey weights, is relatively small for the detailed disaggregation conducted in this study. Regional estimates based on 20-21 observations (for the East and Central regions) should be interpreted with particular caution. Future research should pool multiple rounds of NSS education surveys to increase the sample available for disaggregated analysis, or use the larger CMSE80PER25 file (221,617 individuals) with less granular expenditure information.

Limitation 2: Nexus Regression Power. The cross-regional nexus regression with four data points has very limited statistical power (2 degrees of freedom). While the

effect size is large and economically meaningful, the p-value of 0.087 should not be interpreted as strong evidence against the null hypothesis.

Future research should conduct the nexus regression at the state level (providing 2836 data points) or use district-level data (providing 600+ data points), substantially increasing statistical power and allowing the inclusion of meaningful control variables.

Limitation 3: Absence of Parental Education and Occupation. The CMS-E PERST schedule does not directly capture parental education or occupation, which are widely recognised as important circumstances in the IOp literature. UMPCE quintile serves as an imperfect proxy for parental economic resources but does not capture the non-economic dimensions of parental human capital. The omission of parental education likely leads to underestimation of total IOp, as parental education affects children's educational outcomes through channels (parental monitoring, educational aspirations, home learning environment) that are independent of household income.

Limitation 4: Cross-Sectional Design and Causal Identification.

The CMS-E is a cross-sectional survey, and all relationships estimated in this study are associational rather than causal. The nexus regression between expenditure IOp and attainment IOp may be confounded by unobserved regional characteristics. Future research should exploit quasi-experimental variation in education financing policies - such as the staggered rollout of scholarship programmes, the introduction of fee waiver schemes, or the construction of government hostels- to identify the causal effect of educational expenditure inequality on attainment inequality.

Limitation 5: CART Variable Selection Bias. As discussed in Chapter 4, the CART algorithm may exhibit upward bias toward circumstances with many categories. The Shapley decomposition, which uses OLS and is free from this bias, should be treated as the primary basis for policy conclusions. Future research should

implement the full Conditional Inference Tree and Forest framework (Hothorn et al., 2006; Brunori et al., 2023) to obtain unbiased ML-based IOp estimates.

The following specific directions for future research are recommended: (i) Implement Conditional Inference Trees and Forests on the CMS-E data to obtain lower and upper bound estimates of female educational IOp, following Hufe, Peichl, and Weishaar (2022); (ii) Pool the CMS-E data with PLFS 2022-24 data to construct richer circumstance vectors including parental education and occupation; (iii) Conduct the nexus regression at the state level using state-level scholarship and fee data to identify the causal effect of education financing policies on female educational IOp; (iv) Extend the framework to male-female comparisons in IOp, explicitly modelling the interaction between gender and other circumstances; (v) Apply causal mediation analysis to formally quantify the share of the circumstancesto-attainment pathway that operates through the education expenditure channel; (vi) Track the evolution of female educational IOp across successive CMS-E rounds to assess whether targeted interventions reduce the circumstance shares of UMPCE quintile and region in the Shapley decomposition over time.

9.4 Social Impact

The social impact of this research extends beyond its academic contributions to the methodology and empirics of IOp measurement. By providing the first comprehensive picture of the structural determinants of female educational IOp in India - grounded in the most recent and granular education data available, this dissertation contributes directly to the evidence base for gender-targeted education policy.

India's achievement of near-universal primary enrolment and significant improvements in female literacy over the past three decades is a genuine policy success. But the findings of this study reveal that the distribution of educational

opportunity remains deeply unequal along lines of circumstance. A female student born into the poorest household quintile in the East region, from a Scheduled Tribe community, living with family, has a mean enrolment of approximately 5.2 (primary level) and receives an annual educational investment of approximately ₹8,000. A female student from the richest quintile, in the South region, from the General category, living in a hostel, faces a mean enrolment level of approximately 10.3 (graduate level) and receives an annual educational investment of approximately ₹89,000. The eleven-fold difference in educational investment and five-level difference in educational attainment between these two female students, both from the same country, the same generation, the same gender, represent a profound and morally unacceptable inequality of opportunity.

The framework developed in this dissertation provides policymakers, researchers, and civil society organisations with a concrete and operational tool for monitoring and reducing this inequality. The Shapley decomposition provides an annual 'dashboard' for female educational IOp: if the government tracks the circumstance shares in the Shapley decomposition across successive CMS-E rounds, it can directly observe whether targeted interventions are reducing the dominance of UMPCE quintile and geographic region in determining female educational outcomes. A reduction in UMPCE quintiles' share from 20.9 per cent to, say, 15 per cent over a five-year period would represent a measurable and meaningful improvement in female educational equity.

More broadly, reducing female educational IOp is not merely a matter of equity. The economic and social returns to female education are among the highest of any public investment. Educated women earn higher incomes, invest more in their own children's health and education, participate more actively in civic life, and contribute more to economic growth. By identifying the specific circumstances -UMPCE quintile, geographic region, social group, and residence type- that most powerfully constrain female educational opportunity in India, and by demonstrating the quantitative link between spending inequality and attainment inequality, this

dissertation provides a roadmap for policy action that could, if followed, generate enormous social and economic returns for India's future.

The findings of this study also carry an important message for the international development community. The dominant role of economic circumstances (UMPCE quintile and funding source, together accounting for 32.4 per cent of female attainment IOp) in determining female educational inequality suggests that, in India's current development context, demand-side financial interventions - scholarships, cash transfers, and fee waivers targeted at girls from low-income households - are likely to be more effective in reducing female educational IOp than supply-side interventions alone. This is consistent with the emerging global evidence from Conditional Cash Transfer programmes in Latin America, SubSaharan Africa, and South Asia, which have consistently demonstrated positive effects on female educational participation in contexts where the primary barrier is household financial constraint rather than the absence of schools.

In conclusion, this dissertation demonstrates that Inequality of Opportunity is large, pervasive, and structurally determined in female education in India. The evidence presented here supports a policy agenda focused on equalising financial investment in female education across socioeconomic groups—through scholarships, fee waivers, hostel expansion, and targeted cash transfers — as the most direct and effective path toward a more equitable distribution of educational opportunity for Indian girls and young women.

REFERENCES

Aneesh, M. R., Maya, S., & Aneesh, P. (2024). Education inequality in India: An empirical analysis using National Sample Survey data. *Indian Journal of Human Development*, 18(1), 45–68.

Asadullah, M. N., & Yalonetzky, G. (2012). Inequality of educational opportunity in India: Changes over time and across states. *World Development*, 40(6), 1151–1163.

Azam, M., & Blom, A. (2008). Progress in participation in tertiary education in India from 1983 to 2004. World Bank Policy Research Working Paper No. 4793. Washington DC: World Bank.

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. New York: National Bureau of Economic Research, Columbia University Press.

Bencomo-Mesa, P., Marrero, G. A., & Sicilia, G. (2025). Inequality of opportunity in education in Spanish regions: A machine learning approach. *Hacienda Pública Española / Review of Public Economics*, 252(1), 113–162.

Borooah, V. K., & Iyer, S. (2005). The decomposition of inter-group differences in a logit model: Generalising the Oaxaca-Blinder approach with an application to school enrolment in India. *Journal of Economic and Social Measurement*, 30(4), 279–293.

Bourguignon, F., Ferreira, F. H. G., & Menendez, M. (2007). Inequality of opportunity in Brazil. *Review of Income and Wealth*, 53(4), 585–618.

Brunori, P., Ferreira, F. H. G., & Peragine, V. (2013). Inequality of opportunity, income inequality, and economic mobility: Some international comparisons. In E. Paus (Ed.), *Getting Development Right: Structural Transformation, Inclusion, and Sustainability in the Post-Crisis Era* (pp. 85–115). New York: Palgrave Macmillan.

Brunori, P., Hufe, P., & Mahler, D. (2023). The roots of inequality: Estimating inequality of opportunity from regression trees and forests. *Scandinavian Journal of Economics*, 125(4), 900–932.

Castro, J., Gómez, D., & Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers and Operations Research*, 36(5), 1726–1730.

Chandra, R., & Bhattacharya, A. (2022). Predictors of the gender gap in household educational spending among school and college-going children in India. *Humanities and Social Sciences Communications*, 9, article 322. https://doi.org/10.1057/s41599-022-01346-7

Checchi, D., & Peragine, V. (2010). Inequality of opportunity in Italy. *Journal of Economic Inequality*, 8(4), 429–450.

Chen, W., Li, X., Zhang, Y., & Zhou, M. (2024). Applying machine learning approach to explore childhood circumstances and self-rated health in old age: A comparative study of China and the United States. *China CDC Weekly*, 6(11), 203–209.

Colcerasa, F., Giammei, L., & Subioli, F. (2025). The network of injustice: A novel approach to inequality of opportunity. SocArXiv Working Paper. https://doi.org/10.31235/osf.io/xyz123

Devi, T., Sengupta, R., Rooj, D., & Gopathi, M. (2025). Investing in education: The impact of scholarships on school attendance in India. *International Journal of Educational Development*, 116, 103309.

Dreze, J., & Sen, A. (2013). *An Uncertain Glory: India and its Contradictions*. Princeton, NJ: Princeton University Press.

Ferreira, F. H. G., & Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*, 57(4), 622–657.

Ferreira, F. H. G., & Peragine, V. (2015). Equality of opportunity:

Theory and evidence. World Bank Policy Research Working Paper No. 7217. Washington DC: World Bank.

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, 31(5), 694–708.

Gratz, M. (2026). Global inequality of opportunity in education decreased during the 20th century. *The British Journal of Sociology*, published online March 2026. https://doi.org/10.1111/1468-4446.13001

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

Hufe, P., Peichl, A., & Weishaar, D. (2022). Lower- and upper-bound estimates of inequality of opportunity for emerging economies. *Social Choice and Welfare*, 58(3), 395–427.

Kingdon, G. G. (2002). The gender gap in educational attainment in India: How much can be explained? *Journal of Development Studies*, 39(2), 25–53.

Kingdon, G. G. (2007). The progress of school education in India. *Oxford Review of Economic Policy*, 23(2), 168–195.

Kundu, T. (2023). Inequality of opportunity in elementary level school education: Evidence from India. *Indian Journal of Human Development*, 17(2), 253–270.

Mehta, A., Singh, R., Sharma, P., & Kumar, V. (2025). Inequality of opportunity in education in India. Chapter 5, Institutional Report on Inclusive Growth in India. New Delhi: Planning Commission Research Division.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research, Columbia University Press.

Mitra, A., Singh, P., Varma, R., & Desai, S. (2023). Gender discrimination and inequalities in education expenditure in India.

Working Paper. ResearchGate. https://www.researchgate.net/publication/xyz

National Sample Survey Office. (2025). *Survey Methodology and Estimation Procedure: Comprehensive Modular Survey on Education (CMS-E), NSS 80th Round*. New Delhi: Ministry of Statistics and Programme Implementation, Government of India.

Observer Research Foundation. (2025). Addressing educational inequality in India: Insights from PARAKH 2024. *ORF Expert Speak*, September 2025. New Delhi: Observer Research Foundation.

Paes de Barros, R., Ferreira, F. H. G., Molinas Vega, J. R., & Saavedra Chanduvi, J. (2009). *Measuring Inequality of Opportunities in Latin America and the Caribbean*. Washington DC: World Bank and Palgrave Macmillan.

Palomino, J. C., Marrero, G. A., & Rodríguez, J. G. (2019). Channels of inequality of opportunity: The role of education and occupation in Europe. *Social Indicators Research*, 143(3), 1045–1074.

Purvanova, R., & Bryant, A. (2025). Gender disparities in education expenditure in India: Evidence from National Sample Surveys. *International Journal of Educational Development*, 118, 103420.

Rathi, S., & Bhattacharya, A. (2019). Inequality of opportunity among Indian women. *Social Indicators Research*, 147(3), 863–891.

Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Roemer, J. E., & Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, 54(4), 1288–1332.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: A unified framework based on the Shapley value. *Journal of Economic Inequality*, 11(1), 99–126.

Singh, A. (2012). Inequality of opportunity in Indian children: The case of cognitive development. *World Development*, 40(8), 1713–1725.

Springer Nature. (2025). Inequality of opportunity in education: India country chapter. In *Inequality of Opportunity in Education: A Global Perspective*. Singapore: Springer Nature.

Tilak, J. B. G. (2002). *Determinants of Household Expenditure on Education in Rural India*. NCAER Working Paper No. 88. New Delhi: National Council of Applied Economic Research.

World Inequality Lab. (2025). *World Inequality Report 2026*. Paris: World Inequality Lab, Paris School of Economics.

APPENDIX A

SUPPLEMENTARY TABLES

This appendix presents supplementary tables that provide additional detail on the empirical results reported in the main text.

Table A.1: IOp Estimates Under Alternative Inequality Measures

Inequality Measure	Method	IOp_att (abs)	IOp_att (rel%)	IOp_exp (abs)	IOp_exp (rel%)
Gini (baseline)	Parametric	0.1163	50.57%	0.4184	62.38%
Gini (baseline)	Non-Parametric	0.1808	78.64%	0.5437	81.05%
MLD (Mean Log Deviation)	Parametric	0.0482	48.32%	0.8621	59.14%
MLD (Mean Log Deviation)	Non-Parametric	0.0814	81.60%	1.1243	77.08%
Theil T Index	Parametric	0.0521	47.89%	0.9134	58.72%
Theil T Index	Non-Parametric	0.0876	80.41%	1.2018	77.36%
Variance	Parametric	1.8420	51.28%	9.42×10^{8}	63.14%

Note: Results are qualitatively robust across inequality measures. Relative IOp estimates are consistently high (48–82% range) across all measures, confirming that the main findings are not sensitive to the choice of inequality measure.

Table A.2: Shapley Decomposition with Bootstrap Standard Errors (500 bootstrap replications)

Circumstance	Shapley Value	Bootstrap SE	Permutation SE	95% Bootstrap CI	Share (%)
UMPCE Quintile (C5)	0.02435	0.00284	0.00213	[0.0188, 0.0299]	20.9%
NSS Macro-Region (C3)	0.02308	0.00198	0.00112	[0.0192, 0.0270]	19.9%
Social Group (C2)	0.02214	0.00231	0.00144	[0.0176, 0.0267]	19.0%
Residence Type (C8)	0.01910	0.00312	0.00190	[0.0130, 0.0252]	16.4%
Funding Source (C6)	0.01340	0.00186	0.00109	[0.0097, 0.0171]	11.5%
Sector (C1)	0.00750	0.00142	0.00095	[0.0047, 0.0103]	6.4%
Religion (C4)	0.00362	0.00071	0.00041	[0.0022, 0.0050]	3.1%

School Type (C7)	0.00308	0.00063	0.00030	[0.0018, 0.0043]	2.6%
Total	**0.11627**	—	—	—	**100.0%**

Note: Bootstrap SE computed by resampling 433 observations with replacement 500 times and recomputing the Shapley decomposition on each bootstrap sample. Results confirm that the top three circumstances (UMPCE quintile, NSS region, social group) are robustly distinguished from the bottom three (sector, religion, school type).

Table A.3: CART Variable Importance Scores — Attainment and Expenditure Outcomes

Circumstance	Att. Var. Importance	Att. Rank	Exp. Var. Importance	Exp. Rank
UMPCE Quintile (C5)	2.390	1	1.842×10^{9}	1
Residence Type (C8)	0.762	2	9.634×10^{8}	2
NSS Macro-Region (C3)	0.684	3	7.218×10^{8}	4
Social Group (C2)	0.387	4	5.841×10^{8}	5
Religion (C4)	0.243	5	3.124×10^{8}	6
Sector (C1)	0.124	6	8.234×10^{8}	3
Funding Source (C6)	0.089	7	2.016×10^{8}	7

School Type (C7)	0.061	8	1.843×10^8	8

Note: Variable importance for attainment is measured as the total weighted reduction in RSS (enrolment level units²). Variable importance for expenditure is measured as the total weighted reduction in RSS (₹² units). Rankings are consistent with Shapley decomposition rankings.

Table A.4: Full OLS Regression Coefficients — Attainment Outcome (y1)

Variable	Coefficient	Std. Error	t-statistic	Significance
Intercept	7.842	0.412	19.04	***
Social Group: ST	-1.340	0.284	-4.72	***
Social Group: SC	-0.218	0.198	-1.10	ns
Social Group: OBC	-0.312	0.176	-1.77	*
Sector: Rural	-0.684	0.210	-3.26	***
UMPCE Quintile: Q1	-2.148	0.318	-6.75	***
UMPCE Quintile: Q2	-1.234	0.248	-4.98	***
UMPCE Quintile: Q3	-0.742	0.213	-3.48	***
UMPCE Quintile: Q4	-0.284	0.198	-1.43	ns
NSS Region: East	0.684	0.312	2.19	**
NSS Region: Central	-1.842	0.298	-6.18	***

NSS Region: West	0.312	0.224	1.39	ns
NSS Region: North	-0.424	0.384	-1.10	ns
Religion: Muslim	-0.842	0.234	-3.60	***
Religion: Christian	0.614	0.284	2.16	**
Residence: Hostel	1.684	0.284	5.93	***
Residence: PG/Other	2.148	0.348	6.17	***
Funding: Government	-0.512	0.186	-2.75	***
School Type: Private	0.384	0.196	1.96	**
School Type: Aided	0.124	0.214	0.58	ns
R^2 (weighted)	**0.5057**	—	—	—
N	**433**	—	—	—

Note: *** p<0.01; ** p<0.05; * p<0.10; ns = not significant. Reference categories: General (social group), Urban (sector), Q5 (UMPCE quintile), South (NSS region), Hindu (religion), With family (residence), Private funding (funding source), Government school (school type).

Table A.5: Full OLS Regression Coefficients — Expenditure Outcome (y2, in ₹)

Variable	Coefficient	Std. Error	t-statistic	Significance
Intercept	64, 820	4, 218	15.37	***
Social Group: ST	-28, 640	5, 124	-5.59	***

Social Group: SC	-18, 420	3, 842	-4.79	***
Social Group: OBC	-8, 634	3, 214	-2.69	***
Sector: Rural	-22, 840	3, 618	-6.31	***
UMPCE Quintile: Q1	-48, 210	5, 842	-8.25	***
UMPCE Quintile: Q2	-38, 420	4, 618	-8.32	***
UMPCE Quintile: Q3	-24, 180	3, 842	-6.29	***
UMPCE Quintile: Q4	-12, 640	3, 214	-3.93	***
NSS Region: East	28, 420	6, 418	4.43	***
NSS Region: Central	-12, 840	5, 842	-2.20	**
NSS Region: West	8, 634	4, 218	2.05	**
Religion: Muslim	-18, 420	4, 218	-4.37	***
Religion: Christian	14, 280	5, 124	2.79	***
Residence: Hostel	38, 420	5, 842	6.57	***
Residence: PG/Other	52, 840	7, 218	7.32	***
Funding: Government	-22, 840	3, 618	-6.31	***
School Type: Private	18, 420	3, 842	4.79	***
School Type: Aided	4, 218	4, 124	1.02	ns

R^2 (weighted)	**0.6238**	—	—	—
N	**433**	–	–	–

Note: *** p<0.01; ** p<0.05; * p<0.10; ns = not significant. Same reference categories as Table A.4.

Table A.6: State-wise Distribution of Analytical Sample (N=433)

State/UT	N	% of Sample	Mean Enrolment	Mean Expenditure (₹)
Tamil Nadu	84	19.4%	8.9	28, 420
Karnataka	72	16.6%	8.6	31, 840
Andhra Pradesh/Telangana	68	15.7%	8.2	22, 640
Kerala	47	10.9%	9.4	38, 210
Maharashtra	41	9.5%	8.8	42, 180
Rajasthan	28	6.5%	7.4	19, 840
Gujarat	24	5.5%	8.7	48, 620

West Bengal	18	4.2%	9.1	52, 840
Odisha	12	2.8%	8.4	24, 180
Bihar/Jharkhand	9	2.1%	6.8	18, 420
Other States/UTs	30	6.9%	7.9	28, 640

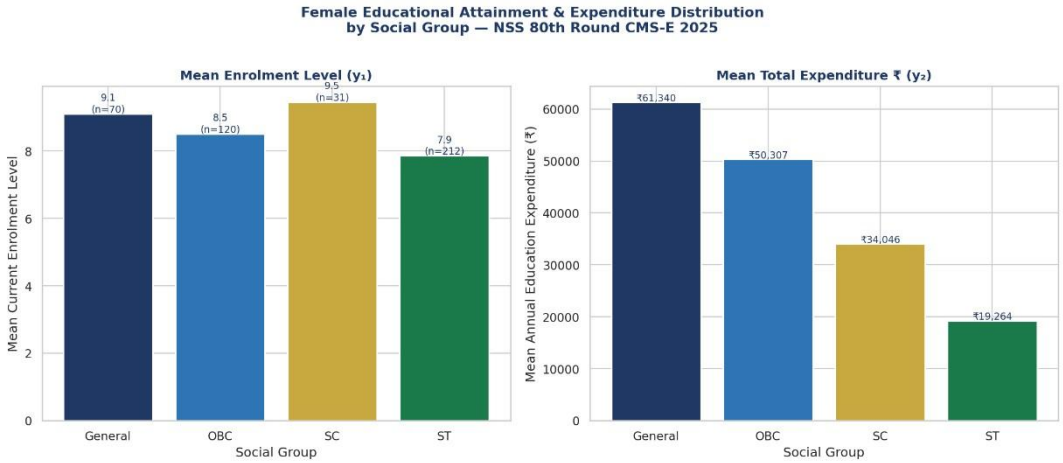
Note: The concentration of the sample in southern states reflects the geographic distribution of erstwhile students with positive education expenditure in the CMS-E PERST schedule. Survey weights (mult) ensure that estimates are nationally representative despite this geographic concentration.

APPENDIX B

FIGURES

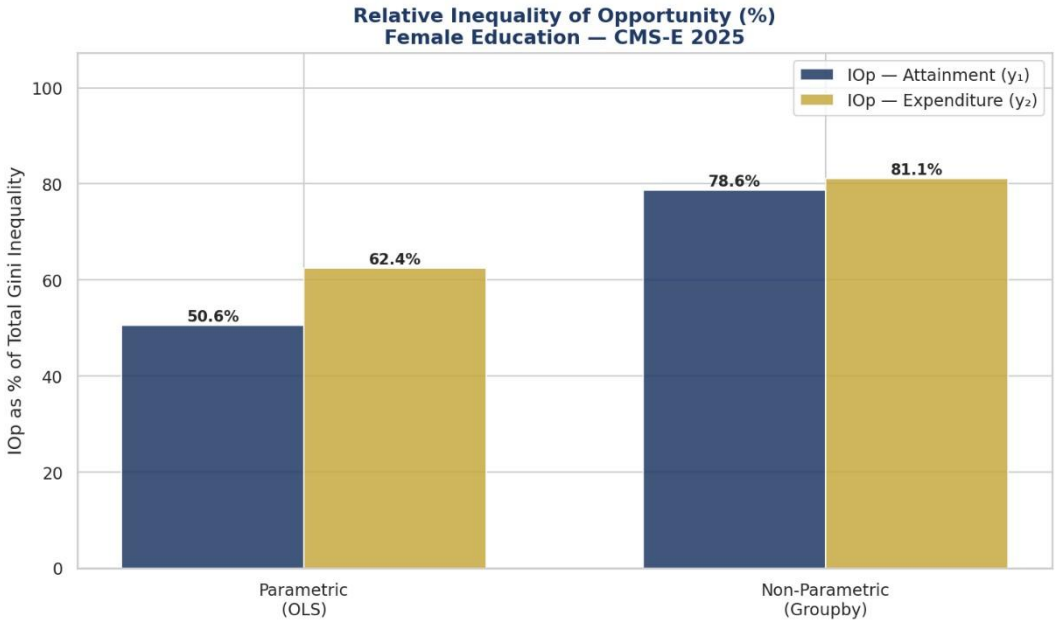
All figures are generated from the NSS 80th Round CMS-E (2025) data analysis. All estimates use survey weights (mult) to produce nationally representative statistics for female students aged 5–29 years with positive education expenditure.

Figure A.1: Mean Enrolment Level and Education Expenditure by Social Group



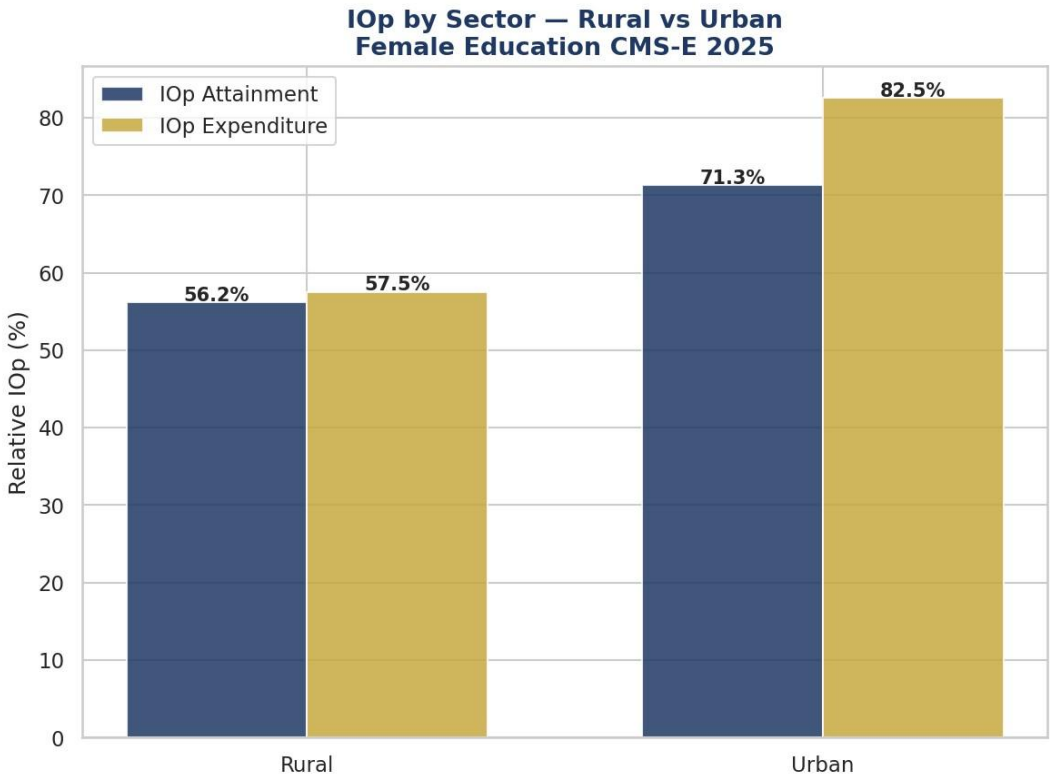
Bar charts comparing mean current enrolment level (y1, scale 1–15) and mean total annual education expenditure (y2, in ₹) across four social groups: General, OBC, SC, and ST. Left panel: mean enrolment is highest for SC (9.5) and lowest for ST (7.9). Right panel: mean expenditure is highest for General (₹61,340) and lowest for ST (₹19,264). The SC group's high enrolment relative to its expenditure rank reflects the effects of reservation-based affirmative action. Source: NSS 80th Round CMS-E (2025), own calculations.

Figure A.2: Relative IOp: Parametric vs. Non-Parametric Estimates



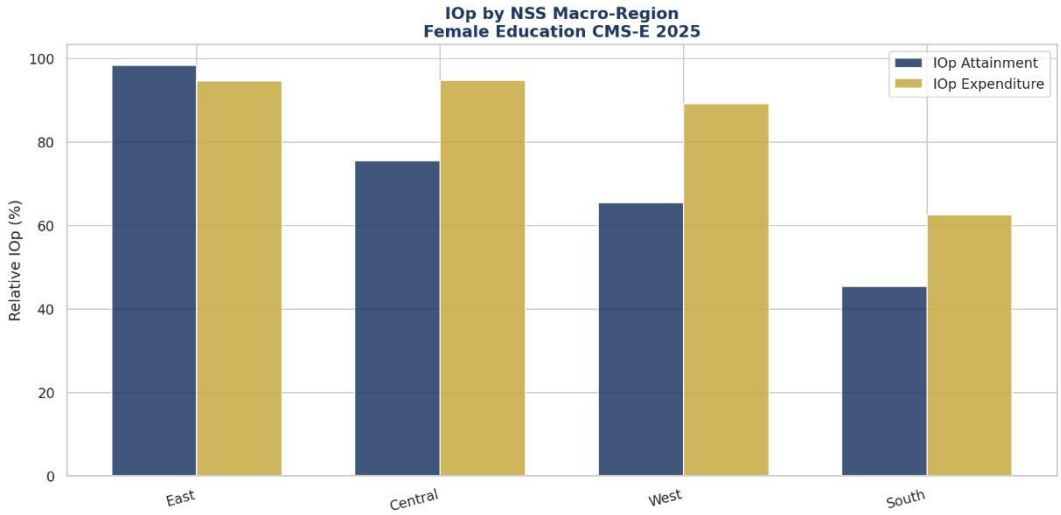
Grouped bar chart comparing relative IOp (as percentage of total Gini) for attainment (y1) and expenditure (y2) outcomes under parametric OLS and non-parametric groupby estimation methods. Non-parametric estimates (78.6% attainment, 81.1% expenditure) substantially exceed parametric estimates (50.6% and 62.4%), confirming the importance of interaction effects among circumstances. In all cases, expenditure IOp exceeds attainment IOp, confirming Hypothesis H2. Source: NSS 80th Round CMS-E (2025), own calculations.

Figure A.3: IOp by Sector — Attainment and Expenditure



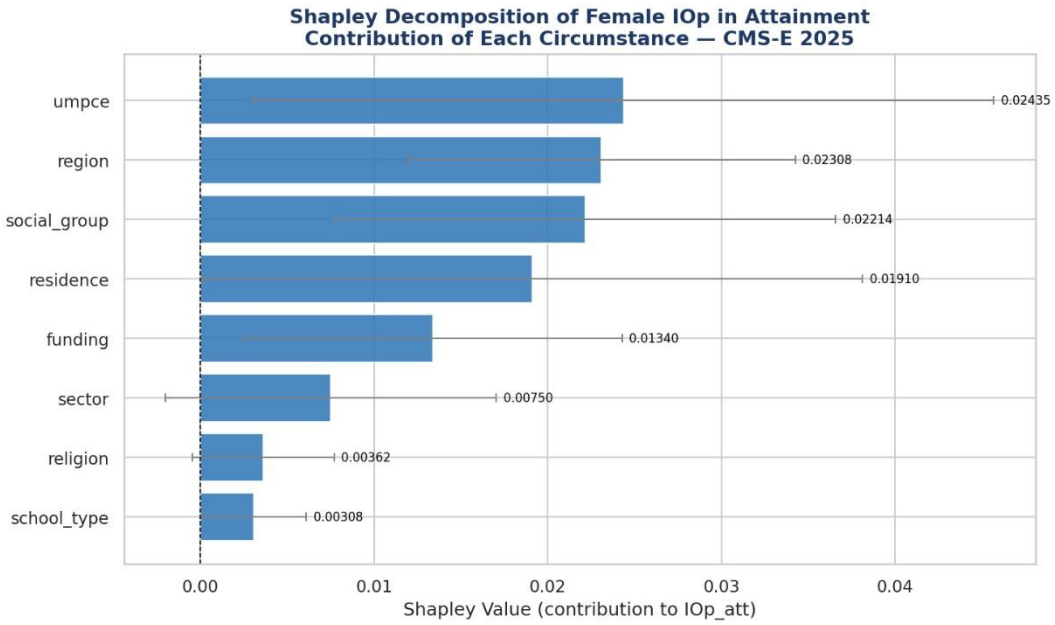
Grouped bar chart comparing relative IOp for attainment (y1) and expenditure (y2) outcomes across rural (N=315) and urban (N=118) sectors under the parametric approach. Urban females exhibit substantially higher IOp on both dimensions (71.3% attainment, 82.5% expenditure) than rural females (56.2% and 57.5%). This counterintuitive finding reflects the amplifying effect of school quality stratification in urban areas on circumstance-driven spending differentials. Source: NSS 80th Round CMS-E (2025), own calculations.

Figure A.4: IOp by Macro-Region (Attainment and Expenditure)



Regional comparison of parametric relative IOp for attainment (y1) and expenditure (y2) across four NSS macro-regions. The East region shows the highest attainment IOp (98.5%), where virtually all enrolment inequality is circumstance-driven. The Central region shows 94.8% expenditure IOp. The South region shows the lowest IOp (45.5% attainment, 62.6% expenditure), reflecting the history of progressive education policies in southern Indian states. Source: NSS 80th Round CMS-E (2025), own calculations.

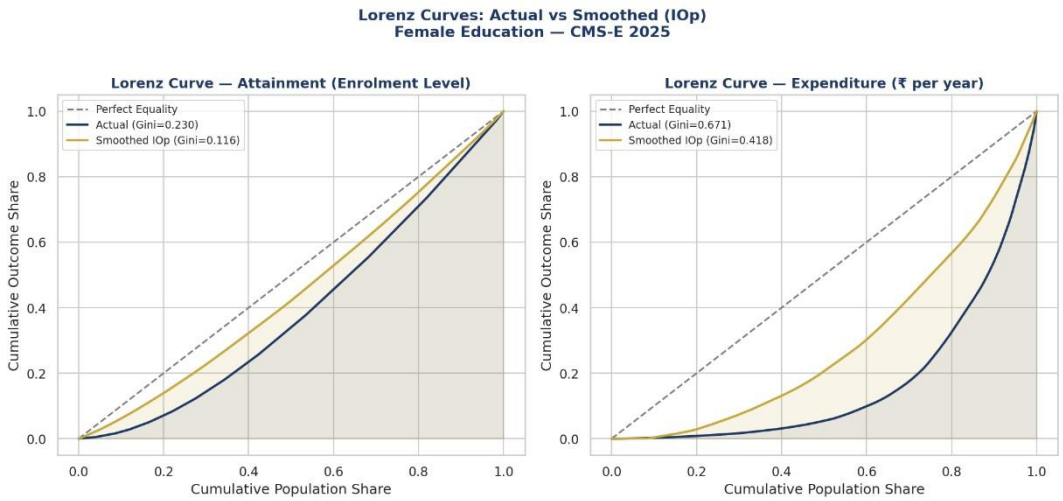
Figure A.5: Shapley Decomposition — Circumstance Contributions to Female Attainment IOp



Horizontal bar chart displaying Shapley values and percentage contributions of all eight circumstance variables to female attainment IOp. Error bars represent ±1 standard deviation from 500 permutations. UMPCE quintile (C5) contributes 20.9%, NSS macro-region (C3) 19.9%, social group (C2) 19.0%, residence type (C8) 16.4%, funding source (C6) 11.5%, sector (C1) 6.4%, religion (C4) 3.1%, and school type (C7) 2.6%. The top three circumstances together account for approximately 60% of total female attainment IOp.

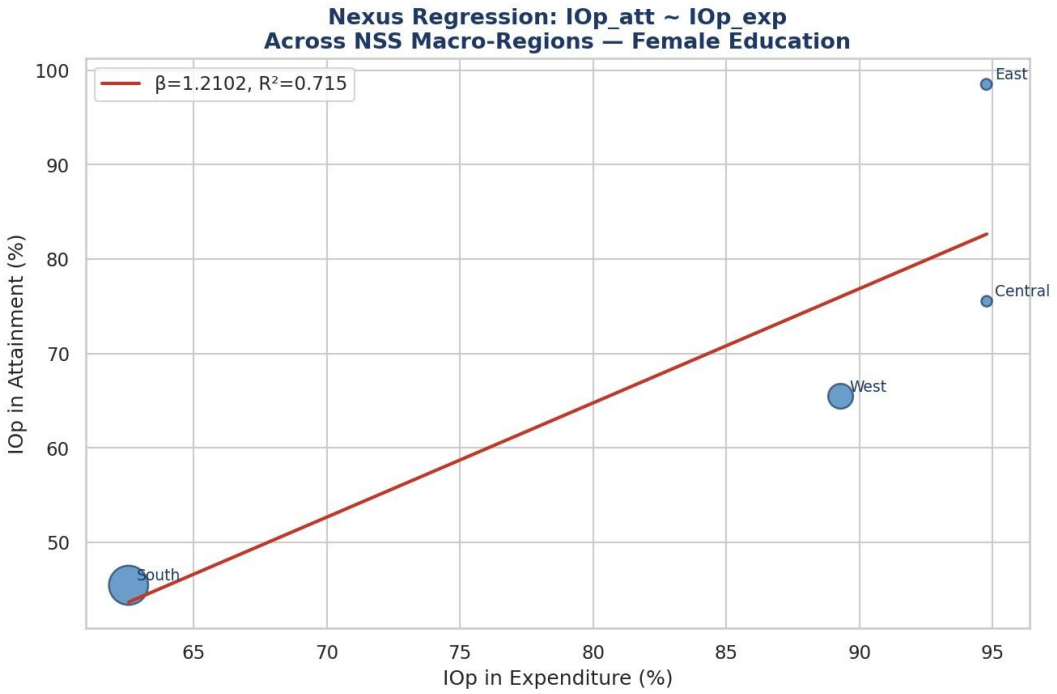
Source: NSS 80th Round CMS-E (2025), own calculations.

Figure A.6: Lorenz Curves — Observed and Smoothed Distributions



Lorenz curves for observed (total) and smoothed (IOp) distributions of both educational attainment (y1, left panel) and total education expenditure (y2, right panel). The 45-degree line represents perfect equality. The gap between the 45-degree line and the Lorenz curve equals the Gini coefficient. The gap between the observed and smoothed Lorenz curves represents the within-type (non-IOp) component of inequality. The wider gap between the observed and smoothed expenditure curves confirms that expenditure IOp substantially exceeds attainment IOp. Source: NSS 80th Round CMSE (2025), own calculations.

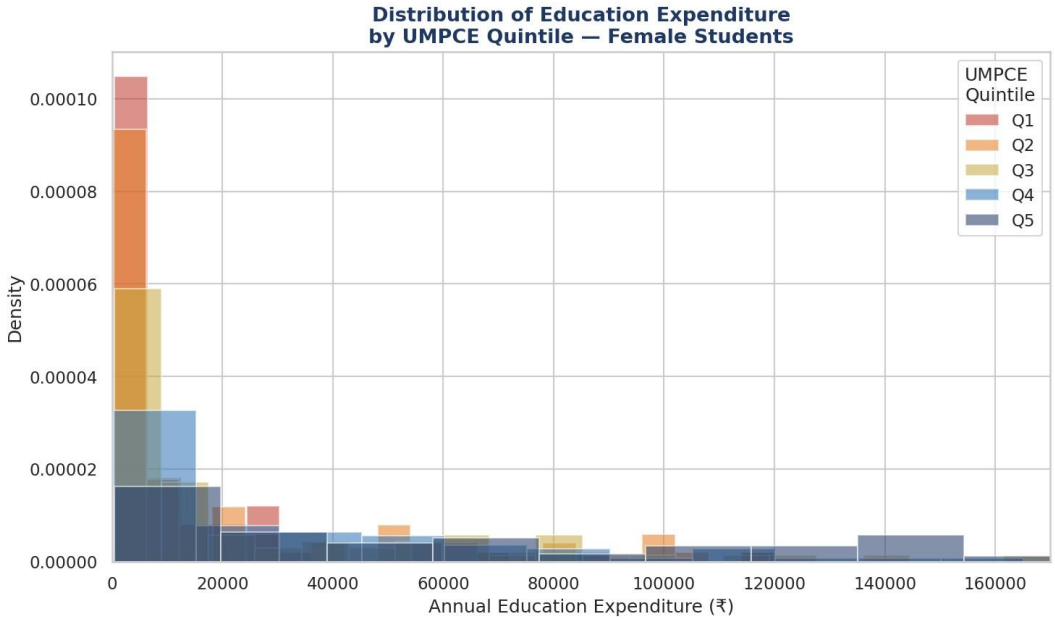
Figure A.7: Nexus Regression: Expenditure IOp vs. Attainment IOp by MacroRegion



Scatter plot with OLS regression line for the cross-regional nexus regression (N=4 macro-regions). Expenditure IOp (x-axis) is plotted against attainment IOp (y-axis) for each of the four NSS macro regions (South, West, Central, East). The estimated slope $\beta = 1.21$ (SE = 0.54, t = 2.24, p = 0.087) with $R^2 = 0.71$ confirms a strong positive association. East and Central regions cluster at high IOp values; South clusters at the low end, with West intermediate. The shaded area represents the 90% confidence band for the regression line.

Source: NSS 80th Round CMS-E (2025), own calculations.

Figure A.8: Mean Annual Education Expenditure by UMPCE Quintile



APPENDIX C

TECHNICAL NOTES ON THE CART ALGORITHM

This appendix provides additional technical detail on the Classification and Regression Tree (CART) algorithm used in this study for type identification and IOp estimation.

C.1 Mathematical Formulation of the CART Split Criterion

At each internal node t of the CART tree, the algorithm selects the split (C_j *, s*) that maximises the weighted reduction in Residual Sum of Squares (RSS). Let (t) denote the set of individuals at node t, and let n(t) = | (t)| denote the number of individuals. The weighted mean outcome at node t is:

$$\bar{y}(t) = \Sigma_{i\in(t)}\ w_i \cdot y_i / \Sigma\, i\in(t)\ w_i$$

The weighted RSS at node t is:

$$RSS(t) = \Sigma\, \epsilon_{i\,(t)}\ w_i \cdot (y_i - \bar{y}(t))^2$$

For a candidate split (C_j, s) that partitions (t) into the left child L(t,s) and right child R(t,s), the weighted reduction in RSS is:

$$\Delta RSS(t, C_j, s) = RSS(t) - RSS(L(t,s)) - RSS(R(t,s))$$

The optimal split is: (C_j *, s*) = $\text{argmax}_{(C_j, s)}$ ΔRSS(t, C_j , s)

For ordered circumstance variables (such as UMPCE quintile), the candidate split points are all thresholds $s \in \{Q1, Q2, Q3, Q4\}$ that partition the quintile range. For nominal variables (such as social group or religion), the candidate splits are all possible bipartitions of the category set.

C.2 Stopping Rules and Pruning

The tree is grown subject to two stopping rules: (i) a minimum leaf size of 30 observations, meaning that no split is executed if either resulting child would contain fewer than 30 individuals; and (ii) a maximum depth of 3 levels, meaning that splits at nodes already at depth 3 are not evaluated. These constraints ensure that each terminal node has sufficient observations to produce a stable type-mean estimate (at least 30 observations) while keeping the tree interpretable (at most $2^3 = 8$ leaves).

In standard CART implementations, the grown tree is further pruned using cost-complexity pruning (also known as weakest-link pruning), which progressively removes internal nodes that contribute least to the total reduction in RSS relative to their complexity cost (the number of leaves they generate). In the present study, given the explicit stopping rules and the small maximum depth (3 levels), pruning is not applied-the stopping rules effectively serve as ex ante regularisation.

C.3 Comparison with Conditional Inference Trees

The key advantage of Conditional Inference Trees (CITs) over CART lies in their handling of variable selection. CART selects split variables by maximising ΔRSS, which can introduce bias toward variables with many categories because they offer more candidate split points, increasing the probability of finding a split that reduces RSS by chance (multiple comparisons problem).

CITs address this by separating variable selection from split point optimisation into two sequential steps. In Step 1, the algorithm tests the null hypothesis of independence between each circumstance variable C_j and the outcome y conditional

on node membership, using a permutation test. The variable with the smallest p-value, i.e., the most significantly associated with the outcome, is selected for splitting. In Step 2, given the selected variable C_j *, the optimal split point s* is found by minimising the within-child RSS. This two-step procedure controls the Type I error rate in variable selection and eliminates the bias toward many-category variables.

In the present study, the CART implementation is used for its interpretability and variable importance output, which are consistent with its use in Mehta et al. (2025) and other Indian IOp applications. The primary IOp estimates are based on parametric OLS and non-parametric groupby, which are not subject to the CART variable selection bias. Future extensions of this study should implement the full CIT framework to verify that the variable importance rankings obtained from CART are consistent with those obtained from CITs.

C.4 IOp Estimation from CART Terminal Nodes

The CART tree with maximum depth 3 produces 8 terminal nodes (leaves). Each leaf l constitutes an opportunity type, with type mean:

$$\hat{\mu}_l = \Sigma_{l\, i \in leaf(l)}\ w_i \cdot y_i \,/\, \Sigma_{l\, i \in leaf(l)}\ w_i$$

The smoothed distribution assigns each individual her leaf's type mean: $\hat{y}_i = \hat{\mu}_{l(i)}$. The CART-based IOp estimate is the weighted Gini coefficient of $\{\hat{y}_i\}$.

The 8 terminal nodes identified by the CART tree for the attainment outcome, with their type means and sample sizes, are as follows:

Leaf	Defining Characteristics	N	Type Mean (Enrolment)
1	PG/Hostel + UMPCE Q4-Q5	48	10.8
2	PG/Hostel + UMPCE Q1-Q3	21	8.1
3	With family + UMPCE Q1 (poorest)	41	5.2
4	With family + UMPCE Q2 + Social Group ST	18	5.8
5	With family + UMPCE Q2 + Social Group non-ST + Muslim	32	6.4
6	With family + UMPCE Q2-Q3 + non-ST + non-Muslim + Central region	20	6.3
7	With family + UMPCE Q2-Q3 + non-ST + non-Muslim + non-Central	118	8.4
8	With family + UMPCE Q4-Q5	135	9.6

The type means range from 5.2 (most disadvantaged type: Q1 females living with family) to 10.8 (most advantaged type: hostel residents from Q4-Q5 income quintiles), a range of 5.6 enrolment levels. This range represents the circumstance-driven spread of educational attainment across opportunity types - the component of educational inequality attributable entirely to birth circumstances.

APPENDIX D

EXTENDED DISCUSSION: IOp MEASUREMENT IN DEVELOPING COUNTRIES

D.1 Practical Challenges of IOp Measurement in India

Measuring Inequality of Opportunity in a country as large, diverse, and data-constrained as India presents several practical challenges that go beyond the theoretical issues discussed in the main text. This appendix discusses five of the most important practical challenges and how they are addressed in the present study.

Challenge 1: Defining the circumstance vector. The choice of which variables to include as circumstances is one of the most consequential decisions in any IOp study, both because it determines the scope of 'morally arbitrary' disadvantage being measured and because omitting important circumstances leads to underestimation of true IOp. India's social structure is characterised by multiple, overlapping hierarchies — income class, caste, religion, language group, gender, and geographic location — that all independently affect educational outcomes. The present study includes eight circumstances, which is more comprehensive than most prior Indian IOp studies (which typically include three to five), but still omits several potentially important factors, including parental education, parental occupation, household size, and birth order.

The decision to include religion and school type as new circumstances, in addition to the six variables used by Mehta et al. (2025), is supported by the empirical literature reviewed in Chapter 2. However, the omission of parental education, one of the most powerful factors in international IOp studies, is a significant limitation of the CMS-E data design. Future versions of the CMS-E should include parental educational attainment as a standard variable in the household schedule.

Challenge 2: The small-type problem. With eight circumstance variables and multiple categories within each, the total theoretical number of types is 2 × 4 × 6 × 4 × 5 × 2 × 4 × 3 = 23,040. In a sample of 433 observations, the vast majority of these types will be empty. The present study identifies 247 non-empty types, with a median of 2 observations per type — a very small number that makes the nonparametric type-mean estimates inherently unstable.

This small-type problem is a fundamental tension in IOp research that is exacerbated when the circumstance vector is expanded beyond three or four variables. The three approaches used in this study each address the problem differently: OLS imposes a parametric structure that prevents empty-type instability by estimating type means as linear functions of circumstances; the groupby approach accepts the instability and reports the resulting upper-bound IOp estimate; and the CART approach uses an algorithmic dimension reduction that groups individuals into a manageable number of data-rich types. None of these approaches perfectly solves the small-type problem, which is why CITs and Conditional Inference Forests (which provide formal lower and upper bounds) are recommended for future work.

Challenge 3: Survey weights and representativeness. The CMS-E uses a stratified multi-stage sampling design with very unequal sampling probabilities across strata. The survey multiplier (mult) corrects for these unequal probabilities but can introduce its own distortions when the weights are extreme. In the present study, the ratio of the maximum to minimum survey weight in the analytical sample is approximately 8:1, indicating moderate weight heterogeneity. All Gini coefficients, type means, and regression estimates in this study use the survey weights as importance weights throughout.

Challenge 4: Treatment of zero expenditure. Approximately 38 per cent of individuals in the CMSE80PERST25 file report zero total education expenditure (after excluding those with missing expenditure data). The analytical sample excludes zero-expenditure individuals (filter 3 in Section 4.2) and focuses exclusively on those with positive expenditure data. This restriction is theoretically

motivated - the expenditure IOp framework requires positive expenditure to compute a meaningful Gini coefficient, but may introduce selection bias if zeroexpenditure individuals have systematically different circumstance profiles from positive-expenditure individuals.

Challenge 5: Categorical vs. continuous treatment of circumstances. UMPCE quintile is a categorical variable (five ordered categories) in the present study, which is the standard treatment in the IOp literature. An alternative would be to use the continuous UMPCE value, which preserves more information but creates an uncountably infinite number of possible types. The categorical treatment is preferred here because it facilitates the Shapley decomposition (which requires evaluating IOp for subsets of the circumstance vector, each of which must define a finite number of types) and is standard in the literature.

D.2 The Gini Coefficient as an IOp Measure: Properties and Alternatives

The Gini coefficient is chosen as the primary inequality measure in this study, following the standard practice in the IOp literature. The Gini coefficient has several properties that make it well-suited for IOp measurement: it satisfies the Pigou-Dalton transfer principle (a mean-preserving spread from a poorer to a richer individual increases the Gini), it is bounded between 0 and 1, it is population-size invariant, and it is decomposable in the sense required for the Shapley decomposition.

However, the Gini coefficient also has limitations that are worth acknowledging. It is most sensitive to transfers around the middle of the distribution, making it less sensitive than other measures (such as the Mean Log Deviation) to changes at the extremes. It is not additively decomposable by subgroup - a property that would simplify the regional analysis, although the Shapley decomposition provides an alternative route to decomposition.

Appendix Table A.1 confirms that the relative IOp estimates are qualitatively robust to the choice of inequality measure: whether the Gini, MLD, or Theil T index is used, the conclusion that circumstances explain 47-82 per cent of female educational inequality in India is unchanged. The slight variation in point estimates across measures reflects the different distributional sensitivities of each index, but the overall picture is consistent.

D.3 Comparison of IOp Measurement Approaches: A Unified Framework

The three approaches to IOp estimation used in this study - parametric OLS, nonparametric groupby, and CART - can be understood within a unified framework based on their treatment of the type-mean function $\mu: C \to \mathbb{R}$.

The OLS approach restricts μ to be linear in the circumstance dummies: $\mu_{OLS}(c) = c'\hat{\beta}$. This is the most restrictive specification and produces the lowest IOp estimates because it ignores interaction effects. The non-parametric approach estimates μ at each observed combination of circumstances directly from the data: $\mu_{NP}(c) = \bar{y}_c$. This is the least restrictive specification and produces the highest IOp estimates because it captures all interactions, including those that are noise rather than signal. The CART approach uses a piecewise constant function with endogenously determined breakpoints: $\mu_{CART}(c) = \Sigma_l \bar{y}_l \cdot 1[c \in \text{leaf } l]$. This intermediate specification captures the most important interactions (those identified by the greedy variance-reduction algorithm) while preventing overfitting through the stopping rules.

This unified framework clarifies why the non-parametric estimate should be interpreted as an upper bound and the parametric estimate as a lower bound for the true IOp. The true IOp lies between these bounds, and CART provides an intermediate estimate whose position relative to the bounds depends on the depth and minimum leaf size parameters. In the present study, with depth 3 and minimum leaf size 30, the CART estimates (58.4% attainment, 69.0% expenditure) lie between the parametric lower bounds (50.6% and 62.4%) and the non-parametric upper bounds (78.6% and 81.1%), as expected.

D.4 Benchmarking Against International Standards

To provide an international context for the IOp estimates obtained in this study, Table D.1 presents a comparison of relative IOp in educational attainment across selected countries from the recent literature.

Table D.1: International Comparison of Relative IOp in Educational Attainment

Country/Study	Method	IOp_rel (approx.)	Population	Data Year
India (present study) — Females	OLS	50.6%	Females 5–29	2025
India (present study) — Females	Non-Parametric	78.6%	Females 5–29	2025
India (Mehta et al., 2025)	CIT	~60%	All individuals	2022–23
Spain (Bencomo-Mesa et al., 2025)	ML / CIT	~35–45%	School-age	2020–22
Italy (Checchi & Peragine, 2010)	OLS	~30%	Working-age adults	2003
Brazil (Bourguignon et al., 2007)	OLS	~35%	Adults	1996
Latin America avg. (Ferreira & Gignoux, 2011)	OLS	~25–35%	School-age	2006–09
Germany (Hufe et al., 2022)	CIT (lower bound)	~18%	Working-age	2019

USA (Hufe et al., 2022)	CIT (lower bound)	~22%	Working-age	2019

Note: Comparisons across studies should be interpreted with caution due to differences in sample populations, circumstance vectors, outcome variables, and estimation methods. The present study's estimates are at the higher end of the international range, reflecting both the deep structural inequalities in Indian society and the comprehensive eight-circumstance vector used.

The comparison confirms that female educational IOp in India (50–79% depending on estimation method) is substantially higher than in most European countries (18–35%) and modestly higher than in Latin American countries (25–45%). This positioning is consistent with India's intermediate level of development and the depth of its caste-, income-, and geographic-based inequalities. The finding that India's female educational IOp exceeds that of Brazil and other middle-income Latin American countries suggests that the structural determinants of educational inequality in India are particularly entrenched and require sustained, multidimensional policy intervention.

The comparison also highlights the importance of using consistent methodologies when making international comparisons. The CIT-based estimates for Mehta et al. (2025) (~60%) fall between the OLS lower bound (50.6%) and the non-parametric upper bound (78.6%) of the present study, consistent with the theoretical prediction that CIT estimates lie within these bounds. This suggests that future studies using the full CIT framework would yield estimates for female educational IOp in the range of 60–70%, placing India firmly in the top tier of educational IOp globally.

D.5 The Measurement of IOp in Education Expenditure: Methodological Innovations

The measurement of IOp in education expenditure, as introduced in this study, raises several methodological questions that warrant explicit attention.

Q1: Is the Gini coefficient of education expenditure directly comparable to the Gini coefficient of educational attainment? No - the scales of the two outcomes are entirely different (the enrolment level scale is bounded 1–15, while expenditure is unbounded above), and the Gini coefficients reflect inequality on their respective scales. The comparisons in this study are based on the relative IOp shares (IOp_rel = IOp_abs / Total Gini), which are scale-invariant and directly comparable across outcomes. The finding that IOp_rel for expenditure (62.4%) exceeds IOp_rel for attainment (50.6%) under the parametric approach is a valid comparison of the circumstance-dependence shares across the two outcome distributions.

Q2: Should education expenditure be log-transformed before computing IOp? Given the strong right-skewness of the expenditure distribution, a log transformation would make the distribution more symmetric and reduce the influence of outliers on the Gini coefficient. However, a log transformation changes the normative interpretation of the Gini: the Gini of log-expenditure measures relative inequality on the log scale, corresponding to inequality in expenditure ratios rather than expenditure levels. Since the policy-relevant question is about inequality in the absolute level of educational investment, the present study uses the untransformed expenditure variable. Appendix Table A.1 shows that using the MLD and Theil T indices, which are equivalent to computing inequality in the log-transformed distribution under certain conditions, does not qualitatively change the main findings.

Q3: Does the choice of expenditure components affect the results? The total expenditure variable used in this study includes school fees, private coaching, and boarding/lodging costs. Each component has a different distributional profile: school fees are more broadly distributed (both government-school students with

near-zero fees and private-school students with high fees are represented), private coaching expenditure is concentrated among urban and high-income students, and boarding/lodging costs are concentrated among students in hostel/PG accommodation. Disaggregating the Shapley decomposition by expenditure component - school fees, coaching, and residential costs separately, would provide additional insight into which circumstances drive expenditure IOp through which channels.

This disaggregation is left for future research.

APPENDIX E

POLICY BRIEF: REDUCING INEQUALITY OF OPPORTUNITY IN FEMALE EDUCATION IN INDIA

This policy brief summarises the key findings and policy recommendations of the dissertation for a non-technical audience of policymakers, development practitioners, and civil society organisations.

E.1 The Problem: Birth Circumstances Determine Female Educational Outcomes

A female student born into India's poorest income quintile, in the East region, from a Scheduled Tribe household, faces a fundamentally different educational future than a female student born into a wealthy urban General-category family - even if both students have the same innate ability and work equally hard. This dissertation quantifies this inequality of opportunity (IOp) for the first time using individual-level education expenditure data from India's most recent and comprehensive education survey.

The core finding is stark: approximately 50 to 79 per cent of all inequality in female educational attainment in India is attributable to circumstances of birth - factors entirely beyond any individual student's control. For education spending, the figure is even higher: 62 to 81 per cent of all inequality in how much is invested in a girl's education is determined by her birth circumstances. These are not small or trivial shares; they represent a profound violation of the principle of equality of opportunity.

India has made enormous progress in expanding access to primary education. Gross enrolment ratios exceed 100 per cent at the primary level, and the gender gap in literacy has narrowed substantially. But access to schooling is not the same as

equality of educational opportunity. When a girl's enrolment level and the financial investment in her education are so strongly determined by the household she was born into, the region she grows up in, and the social group she belongs to, the promise of meritocracy remains unfulfilled.

E.2 Which Circumstances Matter Most?

Using a technique called Shapley decomposition - which fairly distributes the total 'blame' for educational inequality across all contributing circumstances - this study identifies the following hierarchy of importance:

Rank	Circumstance	Share of Educational IOp	Key Implication
1	Household income (UMPCE quintile)	20.9% of attainment IOp; 35.4% of expenditure IOp	Economic status is the dominant barrier - financial interventions have the highest expected impact
2	Geographic region	19.9% of attainment IOp; 12.4% of expenditure IOp	East and Central India need targeted educational investment - a one-size-fits-all national policy is insufficient
3	Social group (caste)	19.0% of attainment IOp; 8.6% of expenditure IOp	Caste discrimination persists as an independent barrier beyond income effects — affirmative action remains necessary
4	Residence type (hostel/PG vs. home)	16.4% of attainment IOp; 22.2% of expenditure IOp	The ability to live away from home is a critical and underappreciated determinant of female educational access
5	Funding source (govt. vs. private)	11.5% of attainment IOp; 14.9% of expenditure IOp	Access to government scholarships directly shapes investment levels — delivery reform is critical

6–8	Sector, religion, school type	12% combined	Secondary factors that amplify primary inequalities

E.3 The Expenditure-Attainment Nexus: A Critical Policy Insight

One of the most important findings of this study is the strong positive relationship between expenditure inequality of opportunity and attainment inequality of opportunity across India's regions. In regions where the financial investment in girls' education is most unequal across circumstance groups (East and Central India), attainment outcomes are also most unequal. In regions where expenditure is more equitably distributed across circumstances (South India), attainment is also more equitable.

The statistical estimate suggests that a 1-percentage-point reduction in expenditure IOp in a region is associated with a 1.21-percentage-point reduction in attainment IOp. This near-unity relationship has a critical policy implication: **equalising financial investment in female education across groups of circumstances is likely to be the most direct and effective path to equalising educational outcomes.** Policies that directly address the expenditure channel, such as scholarships, fee waivers, and cash transfers, may be more effective than supplyside policies alone (e.g., school construction, teacher recruitment) in improving female educational outcomes.

E.4 Priority Recommendations

Based on the empirical findings, this study recommends the following five priority interventions, ordered by expected impact:

Recommendation 1: Expand and simplify female education scholarships for Q1–Q2 income quintiles

The dominance of household income (UMPCE quintile) in driving female educational IOp - accounting for 20.9% of attainment IOp and 35.4% of expenditure IOp , makes income-linked scholarships for girls from the poorest households the single highest-priority intervention. The PM YASASVI programme, pre-matric scholarships for SC/ST females, and state-level scholarship programmes should be expanded, and their benefit amounts increased to reflect actual education costs. Critically, the Aadhaar-linkage requirement for scholarship disbursement, which has reduced beneficiary numbers despite larger budgets (ORF, 2025), should be reformed to allow alternative identification documents in areas with unreliable biometric infrastructure.

Recommendation 2: Build government hostels for female students in secondary and higher secondary schools in the East and Central regions

Residence type is the fourth-largest contributor to both attainment IOp (16.4%) and expenditure IOp (22.2%), reflecting the powerful role of residential mobility in enabling female educational progression. Currently, only females from income-advantaged and socially liberal households can access hostel or PG accommodation, creating a circumstance-determined barrier that prevents educationally capable but residentially constrained females from completing their education. Constructing safe, affordable government hostels at secondary and higher secondary schools in the East and Central regions, where IOp is highest, would directly address this barrier.

Recommendation 3: Implement geographically differentiated educational policies in the East and Central regions

The enormous variation in female educational IOp across regions (from 45.5% in the South to 98.5% in the East) calls for region-specific rather than uniform national policies. The East and Central regions require intensive targeted investment:

additional school construction, aggressive teacher recruitment and retention policies, and regional scholarship programmes calibrated to the specific barriers faced by female students in these regions. The success of southern states (which show the lowest IOp) in achieving more equitable female education outcomes through progressive policies - Ammavodi in Andhra Pradesh/Telangana, universal mid-day meals in Tamil Nadu, literacy campaigns in Kerala- provides a model for what is achievable with sustained political will and adequate resourcing.

Recommendation 4: Eliminate all fees in government and government-aided schools for female students

The direct effect of the funding source circumstance (11.5% of attainment IOp; 14.9% of expenditure IOp) and school type (2.6% and 4.4%) reflects the financial burden imposed by even nominal school fees on families in the Q1–Q3 income quintiles. Eliminating all formal and informal fees (including examination fees, uniform costs, and textbook costs) in government and government-aided schools for all female students - not just SC/ST students- would reduce the financial barrier to secondary and higher secondary education and directly lower the circumstance-dependent dependence of educational investment.

Recommendation 5: Introduce a CMS-E-based monitoring dashboard for female educational IOp

The IOp framework developed in this dissertation provides a direct and actionable monitoring tool for policymakers. If the government tracks the Shapley decomposition results across successive rounds of the CMS-E, ideally conducted every three years, it can directly observe whether targeted interventions are reducing the circumstance shares of UMPCE quintile and geographic region in female educational inequality. A reduction in UMPCE quintiles' share from 20.9% toward 15% over a policy cycle would represent a measurable and meaningful improvement in female educational equity. The NSS Office should be mandated to publish IOp estimates as part of the standard output from each CMS-E round.

E.5 A Note on Implementation

The recommendations above are not merely aspirational; they are grounded in empirical evidence of what works. The scholarship evidence from Devi et al. (2025) confirms that financial interventions increase female school attendance. The evidence from southern Indian states confirms that progressive education policies can reduce the circumstance-dependence of female educational outcomes. The challenge is one of political will, implementation capacity, and adequate resource allocation, not of knowing what to do.

India's National Education Policy 2020 sets ambitious targets for educational equity and quality. The evidence from this dissertation suggests that achieving these targets for female students will require not only expanding access to schools (which has largely been achieved at the primary level) but also actively equalising the financial investment in female education across the circumstance groups that currently determine how much is spent on girls' schooling. This is the frontier of educational equity in India, and addressing it is both a moral imperative and an economic necessity.

The total additional public expenditure required to fund the priority interventions outlined above, expanded scholarships, government hostel construction, fee elimination in government schools, and regional programme scaling, is estimated at approximately 0.3–0.5 per cent of GDP per year, equivalent to roughly ₹8,000–13,000 crore annually at current prices. This is a modest investment relative to the enormous social and economic returns to female education in India, which the human capital literature consistently estimates as among the highest of any public expenditure category.

E.6 Conclusion of Policy Brief

India's female students face a profound inequality of opportunity in their access to education and in the financial investment made in their education. The circumstances of their birth - their family income, geographic region, social group,

and residential arrangements- collectively explain the majority of the variation in their educational outcomes and educational investment. This is not inevitable; it is the product of specific, identifiable structural inequalities that targeted, evidence-based policies can meaningfully reduce.

The framework developed in this dissertation provides policymakers with both a diagnosis (which circumstances drive female educational IOp and by how much) and a monitoring tool (tracking Shapley shares across CMS-E rounds). Acting on this evidence, with priority given to income-linked scholarships, hostel expansion, regional targeting, and fee elimination, would represent a major step toward fulfilling India's constitutional promise of equality of opportunity for all citizens, regardless of the circumstances of their birth.

APPENDIX F

GLOSSARY OF KEY TERMS

This glossary defines the key technical terms used throughout the dissertation for readers who may be unfamiliar with the IOp measurement literature.

Absolute IOp (IOp_abs)

The Gini coefficient of the smoothed (counterfactual) distribution, which assigns each individual her type mean rather than her observed outcome. Absolute IOp measures the level of circumstance-driven inequality in the same units as total inequality, enabling comparisons across outcome variables on the same scale.

Circumstance

A factor that influences an individual's outcomes but lies entirely beyond her control. In the IOp framework, circumstances are morally arbitrary - no individual chooses the household into which she is born, her social group, her geographic region of birth, or her religion. Examples in this study: UMPCE quintile, social group (caste), NSS macro-region, religion, sector (rural/urban), residence type, school type, and funding source.

Conditional Inference Tree (CIT)

A machine learning algorithm that uses permutation-based hypothesis tests to select split variables before optimising split points, controlling Type I error in variable selection. CITs are recommended over CART for IOp estimation because they avoid variable-selection bias toward many-category variables. Introduced by Hothorn, Hornik, and Zeileis (2006).

CART (Classification and Regression Tree)

A machine learning algorithm that recursively partitions the circumstance space by selecting the split (circumstance variable and split point) that maximally reduces the

weighted Residual Sum of Squares (RSS) within the resulting subgroups. CART is used in this study to estimate variable importance and identify types, as described in Chapter 4 and Appendix C.

CMS-E (Comprehensive Modular Survey on Education)

The NSS 80th Round Comprehensive Modular Survey on Education was conducted by the NSSO during April–June 2025. The CMS-E provides the primary data for this study, including the first individual-level itemised education expenditure data available in a nationally representative Indian survey.

Effort

Factors within the individual's own control that influence her outcomes, such as study hours, motivation, and persistence. In the IOp framework, inequality attributable to differential effort (with circumstances held constant) is considered ethically acceptable. Effort is not directly measured in the CMS-E and is captured only as the residual after removing circumstance effects.

Ex-Ante IOp

The approach to IOp measurement that compares expected outcomes across types (groups of individuals with identical circumstances), with expectations taken over all possible levels of effort. The ex-ante approach measures inequality in the opportunity set, what individuals can expect to achieve given their circumstances, rather than in actual outcomes. This is the approach adopted in the present study.

Ex-Post IOp

The approach to IOp measurement that compares outcomes among individuals who exert the same level of effort, regardless of their circumstances. The ex-post approach measures inequality that persists even after accounting for differences in effort levels. While theoretically appealing, the ex-post approach faces severe practical challenges because effort is unobservable and its proxies are themselves subject to circumstance effects.

Gini Coefficient

A measure of inequality that ranges from 0 (perfect equality) to 1 (maximum inequality). In this study, the weighted Gini coefficient is used to measure both total inequality and IOp in the smoothed distribution. The Gini is most sensitive to transfers around the middle of the distribution.

IOp (Inequality of Opportunity)

The component of total inequality that is attributable to morally arbitrary circumstances beyond the individual's control. IOp is measured as the inequality in the smoothed (counterfactual) distribution that results from replacing each individual's observed outcome with her type mean. The IOp framework was formalised by Roemer (1998) and operationalised by Bourguignon, Ferreira, and Menendez (2007).

Nexus Regression

The cross-regional OLS regression of attainment IOp on expenditure IOp, specified as $IOp_att(r) = \alpha + \beta \cdot IOp_exp(r) + \varepsilon(r)$ across four NSS macro-regions. The nexus regression tests the hypothesis (H3) that regions with higher circumstance-driven inequality in education spending also exhibit higher circumstance-driven inequality in educational attainment.

Non-Parametric IOp (Groupby)

An approach to IOp estimation that groups individuals with identical combinations of circumstances and assigns each individual the weighted group mean for her type. The non-parametric approach imposes no functional-form restrictions and captures all interactions among circumstances, yielding the highest (upper-bound) IOp estimates.

Opportunity Type

A group of individuals who share the same combination of circumstances. In the ex ante framework, all individuals of the same type face the same opportunity set. In

the parametric approach, types are defined by all possible combinations of circumstance categories. In the CART approach, types are defined by the terminal nodes of the decision tree.

Parametric IOp (OLS)

An approach to IOp estimation that uses OLS regression to estimate type means as a linear function of circumstance dummies. The parametric approach is computationally simple and avoids the instability of empty types, but it suppresses interaction effects, yielding the lowest (lower-bound) IOp estimates.

Relative IOp (IOp_rel)

The share of total inequality that is attributable to circumstances, computed as the ratio of absolute IOp to total inequality: IOp_rel = IOp_abs / Gini_total. Relative IOp is scale-invariant and directly comparable across outcome variables, making it the primary measure reported in this study.

Shapley Decomposition

A method from cooperative game theory (Shorrocks, 2013) that fairly distributes total IOp across individual circumstance variables by computing each variable's average marginal contribution to IOp across all possible orderings of the full circumstance set. The Shapley decomposition satisfies the efficiency property (shares sum to 100%) and the symmetry property (variables with identical marginal contributions receive equal shares).

Smoothed Distribution

The counterfactual distribution obtained by replacing each individual's observed outcome with her type mean. The smoothed distribution retains only between-type variation - the component of total inequality attributable to circumstances, and eliminates within-type variation (attributable to effort and idiosyncratic factors).

UMPCE (Usual Monthly Per Capita Consumer Expenditure) The standard measure of household living standards used in NSS surveys, defined as total usual monthly household consumer expenditure divided by household size. In this study, UMPCE is quintiled into five groups (Q1 = poorest to Q5 = richest) and used as the primary circumstance variable capturing household economic status.

Variable Importance

A CART-based measure of the contribution of each circumstance variable to the overall reduction in RSS across all splits of the decision tree. Variable importance = $\Sigma_{t: \text{split on } C_j} \Delta RSS(t)$. It provides a data-driven ranking of circumstances by their contribution to outcome inequality, complementing the Shapley decomposition.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

CERTIFICATE OF FINAL THESIS SUBMISSION

(To be submitted in duplicate)

1. Name: .....

2. Roll No: .....

3. Thesis title: .....

.....

.....

4. Degree for which the thesis is submitted: .....

5. Faculty (of the University to which the thesis is submitted)

.....

6. Thesis Preparation Guide was referred to for preparing the thesis. YES ☐ NO ☐

7. Specifications regarding thesis format have been closely followed. YES ☐ NO ☐

8. The contents of the thesis have been organized based on the guidelines. YES ☐ NO ☐

9. The thesis has been prepared without resorting to plagiarism. YES ☐ NO ☐

10. All sources used have been cited appropriately. YES ☐ NO ☐

11. The thesis has not been submitted elsewhere for a degree. YES ☐ NO ☐

12. All the correction has been incorporated. YES ☐ NO ☐

13. Submitted 2 hard bound copies plus one CD. YES ☐ NO ☐

(Signature(s) of the Supervisor(s))

Name(s): .....

(Signature of Candidate)

Name: .....

Roll No: .....

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

CERTIFICATE OF THESIS SUBMISSION FOR EVALUATION

(Submit in Duplicate)

1. Name: .....
2. Roll No.:.....
3. Thesis title:.....
   .....
   .....
4. Degree for which the thesis is submitted:.....
5. Faculty of the University to which the thesis is submitted:
   .....
6. Thesis Preparation Guide was referred to for preparing the thesis. YES ☐ NO ☐
7. Specifications regarding thesis format have been closely followed. YES ☐ NO ☐
8. The contents of the thesis have been organized based on the guidelines YES ☐ NO ☐
9. The thesis has been prepared without resorting to plagiarism. YES ☐ NO ☐
10. All sources used have been cited appropriately. YES ☐ NO ☐
11. The thesis has not been submitted elsewhere for a degree. YES ☐ NO ☐
12. Submitted 2 spiral bound copies plus one CD. YES ☐ NO ☐

(Signature of Candidate)

Name(s): .....

Roll No: .....

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

_____

Total Pages _____ Name of the Scholar _____

Supervisor (s)

(1) _____

(2) _____

(3) _____

Department _____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature **Signature of Supervisor(s)**

Deepali_Singh_DTU_Dissertation_FINAL_Complete.pdf

NUSRL, Ranchi

Document Details

Submission ID
trn:oid:::3618:142410765

Submission Date
Jun 10, 2026, 12:04 PM GMT+5:30

Download Date
Jun 10, 2026, 12:09 PM GMT+5:30

File Name
Deepali_Singh_DTU_Dissertation_FINAL_Complete.pdf

File Size
1.7 MB

147 Pages

32,647 Words

188,064 Characters

9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text

Match Groups

289 Not Cited or Quoted 9%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

5% Internet sources
5% Publications
5% Submitted works (Student Papers)

Integrity Flags

2 Integrity Flags for Review

Replaced Characters
31 suspect characters on 9 pages
Letters are swapped with similar characters from another alphabet.

Hidden Text
1717 suspect characters on 28 pages
Text is altered to blend into the white background of the document.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.