The Impact and Mechanisms of Educational Gaps on Social Stratification

Zhou Haixia ^1^ Wang Bingfeng ^2^

2026-06-29

ChinaRxiv

Abstract

Education and social stratification stand in a bidirectional mutually constitutive relationship. Based on digital divide theory, this paper uses data from the China Family Panel Studies (CFPS) 2016 and employs five methods—Logit regression, PCA factor analysis, OLS regression, Oaxaca-Blinder decomposition, and Baron-Kenny mediation effect analysis—to systematically examine the impact of the educational divide on social stratification and its causal transmission mechanism. The study finds that: (1) each additional year of education increases the probability of Internet access by 5.8 percentage points, while each additional year of age decreases the access probability by 1.8 percentage points, and gender differences are not significant; (2) Internet users' usage types can be divided into "capital-enhancing" and "entertainment-social" categories, and years of education have a strong positive effect on the capital-enhancing factor (β=0.186, p<0.001, R^2=0.333); (3) the positive effect of Internet access on income increases with education (interaction β=0.019, p=0.036), and Oaxaca-Blinder decomposition shows that 59.5% of the income gap between high- and low-education groups comes from differences in returns; (4) capital-enhancing use plays a partial mediating role in the education→income pathway, with a mediation proportion of 54.9% (Bootstrap 95%CI: 0.028-0.041). In terms of robustness, IV-2SLS, replacement with the Kakwani index, the Heckman sample selection model, and clustered standard errors all support the core conclusions. This study provides multilayered empirical evidence for understanding the cumulative causal mechanism of the digital divide in the Chinese context and reveals the micro-level transmission path of the "digital Matthew effect."

Full Text

The Impact and Mechanisms of Educational Gaps on Social Stratification

—An Empirical Study Based on CFPS 2016 Data

Zhou Haixia [1] Wang Bingfeng [2]

[1] College of Public Administration, South China Agricultural University, Guangzhou, Guangdong 510640

[2] College of Materials and Chemical Engineering, South China Agricultural University, Guangzhou, Guangdong 510640

Abstract There is a bidirectional, mutually constitutive relationship between education and social stratification. Based on digital divide theory and using data from the China Family Panel Studies (CFPS) 2016, this paper applies five methods—Logit regression, PCA factor analysis, OLS regression, Oaxaca-Blinder decomposition, and Baron-Kenny mediation-effect analysis—to systematically examine the impact of educational gaps on social stratification and the causal transmission mechanisms involved. The study finds that: (1) for each additional year of education, the probability of Internet access increases by 5.8 percentage points; for each additional year of age, the probability of access decreases by 1.8 percentage points; gender differences are not significant; (2) Internet users' usage types can be divided into two categories: "capital-enhancing" and "entertainment-social." Years of education have a strong positive effect on the capital-enhancing factor ($\beta = 0.186$, $p < 0.001$, $R^2 = 0.333$); (3) the positive effect of Internet access on income increases with education (interaction $\beta = 0.019$, $p = 0.036$). Oaxaca-Blinder decomposition shows that 59.5% of the income gap between high- and low-education groups comes from differences in returns; (4) capital-enhancing use plays a partial mediating role in the education → income pathway, with the mediation proportion reaching 54.9% (Bootstrap 95% CI: 0.028–0.041). In terms of robustness, IV-2SLS, substitution with the Kakwani index, the Heckman sample-selection model, and clustered standard errors all support the core conclusions. This study provides multilayered empirical evidence for understanding the cumulative causal mechanism of the digital divide in the Chinese context and reveals the micro-level transmission pathway of the "digital Matthew effect."

Keywords: digital divide; CFPS; capital-enhancing use; digital Matthew effect

The Impact and Mechanisms of the Educational Gap on Social Stratification

An Empirical Study Based on CFPS 2016 Data

Zhou Haixia[1] Wang Bingfeng[2]

College of Public Administration, South China Agricultural University, Guangzhou, Guangdong 510642

College of Materials and Chemical Engineering, South China Agricultural University, Guangzhou, Guangdong 510642

Abstract: Education and social stratification exhibit a mutually constitutive relationship. Grounded in digital divide theory, this paper utilizes data from the 2016 China Family Panel Studies (CFPS) to systematically examine the impact of the educational gap on social stratification and its causal transmission mechanisms. Five methods are employed: Logit regression, Principal Component Analysis (PCA), Ordinary Least Squares (OLS) regression, Oaxaca-Blinder decomposition, and Baron-Kenny mediation analysis. Findings indicate that: (1) Each additional year of education increases the probability of internet access by 5.8 percentage points, while each additional year of age decreases this probability by 1.8 percentage points; gender differences are not statistically significant. (2) Internet usage patterns can be categorized into two types: "capital-enhancing" and "entertainment-social." Educational attainment has a strong positive effect on the capital-enhancing factor ($\beta = 0.186, p < 0.001, R^2 = 0.333$). (3)

The positive effect of internet access on income increases with educational level (interaction $\beta = 0.019, p = 0.036$). An Oaxaca-Blinder decomposition reveals that 59.5% of the income gap between high- and low-education groups stems from differences in returns. (4) Capital-enhancing usage plays a partial mediating role in the education-to-income pathway, accounting for 54.9% of the mediation effect (Bootstrap 95% CI: 0.028–0.041). Regarding robustness, IV-2SLS, substitution with the Kakwani index, Heckman sample selection models, and clustered standard errors all support the core conclusions. This study provides multi-level empirical evidence for understanding the cumulative causality mechanisms of the digital divide in the Chinese context, and reveals the micro-level transmission pathways of the "digital Matthew effect."

Keywords: Digital Divide; CFPS; Capital-enhancing Usage; Digital Matthew Effect

1. Introduction

Since the concept of the "Digital Divide" was formally proposed in the mid-1990s (NTIA, 1999), inequalities in access to and use of information technology have become an important issue in research on social stratification. China has undergone the largest and fastest digital transformation in human history: data from the China Internet Network Information Center (CNNIC) show that the number of Chinese Internet users increased from 111 million in 2005 to nearly 1.1 billion in 2024, while the Internet penetration rate jumped from 8.5% to about 78%. However, the distribution of digital dividends is highly uneven—significant gaps remain between urban and rural areas, among educational groups, and across generations in Internet access, modes of use, and actual returns.

van Dijk (2020) points out that the digital divide is not a static binary opposition—access vs. no access—but rather a progressive system of inequality comprising three levels: access (Level 1), skills and use (Level 2), and outcomes (Level 3). The focus of international scholarship has gradually shifted from Level 1 to Level 2 and Level 3 (van Deursen & van Dijk, 2014; van Deursen & Helsper, 2015; Scheerder et al., 2017), but complete causal-chain tests of the three-layer digital divide in the Chinese context remain relatively limited (Zhao & Kuang, 2025).

Current research on China' s digital divide faces three core challenges. First, most studies analyze a single level independently and lack quantification of the complete causal pathway from "access → use → outcomes" (a fragmentation of the three-layer causal chain). Second, decomposition analyses of between-group differences are insufficient, making it impossible to distinguish the "sources in characteristics" from the "sources in returns" of disparities. Third, the mediating transmission mechanism of capital-enhancing use has not been directly tested statistically. This study aims to contribute along these three dimensions.

This study revolves around one core question: in the context of China' s current digital society, how do Internet access, use, and outcomes—three progressive levels—generate cumulative inequality, thereby shaping the mutual construction of education and social stratification? Specifically, this is decomposed into four sub-questions:

(1) How do education and age affect the probability of Internet access? (H1: access divide)

(2) How does SES shape differentiation in Internet users' patterns of use? (H2: use divide)

(3) Do income returns from the Internet increase progressively with educational level, and do between-group gaps arise from differences in characteristics or differences in returns? (H3: outcome divide)

(4) Is capital-enhancing use a mediating pathway through which educational advantage is converted into economic returns? (H4: mediating mechanism)

The theoretical significance of this study lies in the following: it fully tests the three-level digital divide theory in the Chinese context; for the first time, it directly quantifies the mediating effect of capital-enhancing use within the CFPS framework; and it reveals the structural sources of the benefits divide through an Oaxaca-Blinder decomposition. Its practical significance lies in providing data support and empirical evidence for the precise implementation of digital inclusion policies—that is, for distinguishing among different intervention strategies at the levels of access, use, and benefits.

2. Literature Review and Theoretical Framework

2.1 The Evolution of the Three-Level Digital Divide Theory

The concept of the digital divide has undergone a theoretical evolution from a single-dimensional notion to a multilayered framework. NTIA (1999), in the report *Falling Through the Net*, was the first to systematically define inequality in digital access, prompting scholarly attention worldwide. However, van Dijk (2005) pointed out that simply equating the digital divide with the binary distinction of "having a computer/not having a computer" obscures the deeper structure of inequality, and accordingly proposed the highly influential "Causal and Sequential Model of Access."

This model divides the digital divide into three progressive levels: (1) Level 1, the access divide—differences in material access, including the availability of basic infrastructure such as computers and Internet connections; (2) Level 2, the skills and usage divide—including operational skills, information-navigation skills, and strategic skills (Hargittai, 2002), as well as differentiation in types of use (Blank & Groselj, 2014); and (3) Level 3, the benefits divide—that is, even with the same access and use, different social groups still exhibit systematic differences in the actual benefits they obtain from the Internet (van Deursen & Helsper, 2015; Ragnedda, 2017).

The important theoretical contribution of the three-level model lies in its "sequentiality" and "cumulativity": access is a prerequisite for use (although not a sufficient condition), and use is a prerequisite for benefits. Social stratification (education, income, age, etc.) affects access, use, and benefits layer by layer, forming the causal chain of "social status → access differences → usage differentiation → benefits inequality → reproduction of social status" (van Dijk, 2020). This framework constitutes the core theoretical foundation of the present study.

The sequential characteristics of the three-level digital divide model provide the logical starting point for integrating it with capital theory. As digital inequality evolves from the access level to the use and benefits levels, the mechanism through which social stratification operates shifts from "resource accessibility" to "the capacity to convert resources." This shift extends the research perspective on the digital divide from the allocation of technological resources to the issue of capital conversion and reproduction described by Bourdieu (1986). Compared with

At the same time, the Matthew effect (Merton, 1968) provides a theoretical tool for understanding how this process of capital transformation generates cumulative inequality in the digital domain. The following two subsections further develop the understanding of the three-level theory of the digital divide from two dimensions: use differentiation through capital enhancement, and cumulative inequality.

2.2 Capital Enhancement Theory and Use Differentiation

Within the second-level digital divide, "use differentiation" is the theoretical issue that has attracted the greatest attention. The pioneering study by Zillien and Hargittai (2009) proposed the concept of "digital distinction": high-SES groups tend to engage in "capital-enhancing uses" (including online learning, career development, health information, financial management, etc.), whereas low-SES groups are more likely to engage in entertainment and social uses. This finding echoes Bourdieu's (1986) theory of cultural capital: patterns of Internet use are the digital expression of cultural capital and serve the reproduction of social stratification.

van Deursen and van Dijk (2014) verified this differentiation model in a national survey in the Netherlands, finding that even after controlling for frequency of access, the effect of education on type of use remained significant. Scheerder et al. (2017), in a systematic review of 49 determinants, confirmed that education is the strongest predictor of digital skills and type of use. Helsper's (2012) "Corresponding Fields Model" further divides digital outcomes into four domains—economic, cultural, social, and personal—and shows that different SES groups vary in their patterns of benefit across these domains.

The operational meaning of the capital-enhancement hypothesis is that the digital divide is not merely a question of "whether one can use" digital technologies, but rather a question of "what one uses them for." Even if low-SES groups overcome the access divide, if their patterns of use are concentrated in entertainment and social domains rather than capital-enhancing domains, their digital returns will be systematically lower than those of high-SES groups. This study treats this assumption as the theoretical core of H2 and H4.

2.3 The Digital Matthew Effect and Cumulative Inequality

The "Matthew effect" proposed by Merton (1968)— "for whoever has will be given more, and they will have an abundance; whoever does not have, even what they have will be taken from them" —has particular theoretical applicability in the digital domain. van Dijk (2020) explicitly pointed out that "feedback loops" exist among the three levels of the digital divide: the outcomes divide in turn affects the next round of access motivation and skill investment, forming a self-reinforcing process of cumulative causation.

Cumulative advantage theory provides a more precise theoretical framework for understanding the micro-level transmission mechanisms of the digital Matthew effect. DiPrete and Eirich (2006) define cumulative advantage as "the resource status of a given actor at a certain point in time affects the probability of obtaining subsequent resources···

rate," which corresponds theoretically to the three-layer digital-divide "feedback loop" described by van Dijk (2020). In the digital domain, the operation of cumulative advantage can be decomposed into three links: first, those with higher education cross the access threshold earlier, gaining a temporal advantage in the accumulation of digital skills; second, the early accumulation of digital skills further promotes more capital-enhancing forms of use, thereby widening gaps at the level of use; finally, capital-enhancing use is converted into higher economic returns, forming a cumulative echo chain of "educational advantage → access priority → skill accumulation → capital-enhancing use → enhanced returns." This study will use Oaxaca–Blinder decomposition and mediation-effect analysis to empirically test the existence and strength of this cumulative advantage chain.

Ragnedda (2017), drawing on Weberian social stratification theory, understands the digital divide as a new dimension of inequality that interacts with traditional inequalities of class, status, and power. DiMaggio and Hargittai (2001) pointed out as early as the initial stage of Internet diffusion that the key dimensions of digital inequality include device quality, autonomy of use, skill level, and social-support networks—differences in these dimensions display typical features of "cumulative advantage." In the Chinese context, the "digital Matthew effect" exhibits distinctive institutional characteristics. Owing to the urban–rural dual structure, the household-registration system, and regional differences in returns to education (Li et al., 2022), a complex coupling relationship has formed between the digital divide and existing social inequalities. Through Oaxaca-Blinder decomposition and mediation-effect analysis, this study aims to empirically test the existence and strength of this cumulative causal mechanism.

2.4 Current Status of Research on China' s Digital Divide

In recent years, research on the digital divide based on large-scale Chinese survey data has made important progress. Zhao and Kuang (2025), using seven waves of CGSS data from 2005 to 2021, employed Logit regression and OLS regression to examine the intertemporal evolution of the three levels of the digital divide. They found that the Level 1 divide has narrowed significantly, but the Level 2 and Level 3 divides persist, and that capital-enhancing use can help disadvantaged groups catch up, although participation by these groups remains limited. Ran et al. (2026), focusing on older adults in urban and rural areas, found that the access gap has narrowed but gaps in frequency of use and health benefits persist, thereby verifying the applicability of the "three-level diffusion mechanism" among older populations.

Guo and Guo (2025), based on CFPS 2020 data, found that the digital divide significantly suppresses household participation in financial investment, with financial knowledge mediating approximately 40% of the effect, thus providing a methodological reference for research on mediation mechanisms. A study of digital finance in China by the International Monetary Fund (Shen et al., 2025) found that digital finance reduces within-province inequality but has the smallest effect on older adults, echoing the "intergenerational dimension" of the digital divide. Liu et al. (2024), from a regional perspective, revealed that digital inequality exceeds economic inequality (coefficient of variation 0.92 vs 0.59), suggesting that digital transformation may intensify rather than weaken existing regional disparities.

However, existing research has three methodological limitations. First, most studies rely on mean regression and lack a structured decomposition of intergroup gaps; the Oaxaca-Blinder decomposition can distinguish between "endowment effects" and "coefficient effects." Second, the "mediating transmission" role of capital-enhancing activities is theoretically assumed but has not been directly tested statistically within the CFPS framework. Third, the complete pathway of the multilayer progressive causal chain—access → use → returns—has not been modeled simultaneously. To address these gaps, this study employs Oaxaca-Blinder decomposition to fill G2 and Baron-Kenny mediation analysis to fill G3, while integrating a four-layer analysis to form a complete causal chain to fill G1, thereby systematically quantifying the digital divide within the CFPS framework.

2.5 Research Gaps and the Positioning of This Study

Based on the above literature review, current research on the digital divide has the following core gaps and corresponding strategies for addressing them (see

Table 1).

Table 1 Research Gaps and Strategies for Addressing Them

No.	Research gap	Strategy for addressing the gap	Corresponding hypothesis
G1	Quantification of the three-level causal chain is fragmented	Logit + PCA + OLS + mediation analysis, constructing the complete pathway of access → use → returns	H1–H4
G2	Lack of decomposition analysis of intergroup gaps	Oaxaca–Blinder decomposition, distinguishing endowment effects and coefficient effects	H3
G3	The mediating role of capital enhancement has not been directly tested	Baron–Kenny three-step method + Bootstrap confidence intervals + Imai framework	H4
G4	Insufficient exploration of heterogeneity in usage types	PCA factor analysis to identify two dimensions of usage types	H2
G5	Limited examination of distributional effects in the returns divide	KDE comparison of income distributions by education group	H3
G6	Insufficient treatment of endogeneity	IV-2SLS instrumental-variable method + Heckman sample-selection model	H3 robustness

G7	Lack of clustered standard errors	Province-level clustered robust standard errors	H1–H4

Drawing together the above theoretical threads, this study constructs an integrated analytical framework that combines the three-level digital divide theory, Bourdieu' s theory of capital conversion, Merton' s Matthew effect, and cumulative advantage theory. The core logic of this framework is as follows: social stratification, with education as the core proxy variable, operates through three levels of

transmission mechanisms. At the access level, education and age constitute the "dual diversion valves" of digital access (H1); at the use level, education shapes differentiation in patterns of use, with high-SES groups tending toward capital-enhancing use and low-SES groups concentrating on entertainment- and social-oriented use (H2); at the returns level, the economic returns to the Internet increase with education, and most of the between-group gap derives from structural differences in returns rather than differences in endowments (H3), while capital-enhancing use plays a key mediating role between education and income (H4). The entire chain exhibits an evolutionary pattern of "initial gap → transmitted amplification → structural solidification," forming a complete echo chain of "social stratification → access divide → differentiated use → inequality in returns → reproduction of social stratification." (See Figure 1)

Text in the figure:

- Integrative Theoretical Analytical Framework
- Social stratification (education/family background/SES)
- First layer: Access divide
  Differentiation in probability of access (H1)
- Second layer: Use divide
  Capital enhancement vs. entertainment/socializing (H2)
- Third layer: Returns divide
  Inequality in income returns (H3)
- SES effects layer by layer
- Mediating transmission mechanism of capital-enhancing use
  (H4: mediation proportion 54.9%)
- Between-group gap decomposition: Oaxaca-Blinder
  Characteristic effect vs. coefficient effect (59.5%)
- Digital Matthew effect: the returns gap feeds back into the next round of access and use
  "initial gap → transmitted amplification → structural solidification"
- Feedback loop
  (cumulative causation)
- Cumulative effect of the educational divide on social stratification: access sorting → differentiated use → amplified returns → structural solidifica-

tion

Figure 1. Schematic Diagram of the Integrative Theoretical Analytical Framework

3. Research Design

3.1 Data Sources and Sample Description

This study uses data from the 2016 wave of the China Family Panel Studies (CFPS). CFPS is conducted by the Institute of Social Science Survey at Peking University. It is a national, comprehensive longitudinal survey project covering 25 provinces/municipalities/autonomous regions across China and representing more than 95% of the Chinese population. CFPS adopts a multistage probability sampling method. Based on administrative divisions, it uses a four-stage sampling design of district/county → village/community → household → individual, ensuring the national representativeness of the sample. The 2016 survey collected

36,875 individual-level observations. After excluding cases with missing values in core variables such as age, gender, education, and Internet access, the final valid analytical sample consists of 29,893 individuals. The CFPS 2016 questionnaire contains a rich module on Internet use, covering information such as access status (whether the respondent uses the Internet), frequency of use (five dimensions: study, work, social interaction, entertainment, and business, each scored from 1 to 7), evaluations of the importance of use, and types of access devices (mobile phone/computer). It also includes complete demographic variables, such as age, gender, years of education, income, household registration type, and urban–rural codes, providing an ideal data foundation for a comprehensive analysis of the three-level digital divide.

3.2 Variable Definitions and Measurement

The variable system in this study is organized according to the three-level digital divide framework, as shown below (see Table 2).

Table 2 Variable Definitions and Measurement

Variable type	Variable name	Measurement method	Values	Description
Dependent variable–L1	`any_net`	Binary	0 = not online, 1 = online	Internet access status
Dependent variable–L2	`freq_*`	Ordered	1-7	Use frequency across five dimensions

Variable type	Variable name	Measurement method	Values	Description
Dependent variable—L2	`capital_factor`	Continuous	Standardized	Capital-enhancement factor extracted by PCA
Dependent variable—L3	`ln_income`	Continuous	Logarithm	Log transformation of annual personal income
Core independent variable	`eduy`	Continuous	0–22 years	Years of education
Core independent variable	`edu_group`	Ordered	1-5	Education grouping
Control variable	`age`	Continuous	16-104	Age
Control variable	age^2	Continuous	—	Age squared (nonlinear effect)
Control variable	`gender`	Binary	0 = female, 1 = male	Gender
Control variable	`urban`	Binary	0 = rural, 1 = urban	Urban-rural classification
Clustering variable	`provcd_valid`	Categorical	1-32	Province code (standard-error clustering)

3.3 Analytical Methodology

This study adopts six progressive quantitative methods, corresponding respectively to different dimensions of the three-level digital divide and to the research hypotheses. Method 1: Logit regression (H1). Taking Internet access (any_net) as a binary dependent variable, the model estimates the marginal effects of education, age and its quadratic term, and gender on the probability of access. Method 2: PCA factor analysis + OLS regression (H2). Principal component analysis (PCA) is conducted on the usage frequency of online users across five dimensions, extracting usage-type factors in a data-driven manner. The capital-enhancing factor is then used as the dependent variable, while years of education, age, gender, and urban-rural status are used as independent variables to construct an OLS regression model. Method 3: Oaxaca–Blinder decomposition (H3). The sample is divided into high- and low-education groups according to the median level of education, and the income gap between groups is decomposed into a characteristic effect (endowment differences) and a coefficient effect (return differences). Method 4: Baron–Kenny and Imai causal mediation analysis (H4). Years of education are taken as the independent variable X, the capital-enhancement index as the mediating variable M, and log income as the dependent variable Y. At the same time, the causal mediation framework of Imai et al. (2010) is used for sensitivity analysis. Method 5: IV-2SLS instrumental-variable method (robustness). The interaction term between the provincial-level average Internet access rate and years of education is used as an instrumental variable to test for endogeneity bias in the returns-divide model. Method 6: Heckman sample-selection model (robustness). A two-stage model is used to correct sample-selection bias in the income variable. All regression models report province-level clustered robust standard errors to control for within-group correlation among individuals in the same province.

3.4 Research Hypotheses

Based on the theoretical framework and the existing literature, this study proposes four progressive hypotheses:

H1 (access-divide hypothesis): Education has a significant positive effect on the probability of Internet access; age has a significant negative effect; and gender differences are not significant after controlling for education and age.

H2 (usage-differentiation hypothesis): The usage types of Internet users can be divided into two dimensions: capital-enhancing use and entertainment/social use. High-SES groups are significantly more inclined toward capital-enhancing use.

H3 (returns-inequality hypothesis): The income returns to Internet access increase with educational attainment (the interaction effect is significant), and most of the income gap between the high-education and low-education groups comes from the coefficient effect (return differences) rather than the characteristic effect.

H4 (mediation-path hypothesis): Capital-enhancing use plays a significant mediating role in the causal pathway from education to income (the indirect effect $a \times b$ is significant, and the Bootstrap 95% CI does not include 0).

4. Empirical Results

4.1 4.1 Descriptive Statistics and Preliminary Findings

Descriptive statistics (see Table 3) reveal a pronounced access divide: the average age of the internet-user group is 33.8 years, far below the 55.2 years of the non-internet-user group, with an age gap of more than 21 years; the average years of education among internet users is 11.02 years (close to high-school graduation), whereas that of non-users is only 5.39 years (primary-school level), producing an education gap of more than 5.6 years. The share of urban residents among internet users (60.3%) is significantly higher than among non-users (39.5%), a rural–urban gap of 21 percentage points. In terms of usage frequency, social frequency is the highest (mean 5.85/7.0), while business frequency is the lowest (mean 2.85/7.0), indicating that social functions are the primary driver of internet use.

Table 3 Descriptive Statistics (N=29,893)

Variable	Overall Mean	Overall SD	Internet-User Mean	Non-Internet-User Mean
Age	46.40	17.61	33.80	55.20
Years of education	7.70	4.93	11.02	5.39
Log income	9.98	1.11	10.11	9.63
Internet access rate	0.411	0.492	1.000	0.000
Male	0.492	0.500	0.518	0.475
Urban	0.480	0.500	0.603	0.395
Learning frequency	3.52	2.38	3.52	–
Work frequency	3.12	2.67	3.12	–
Social frequency	5.85	1.94	5.85	–
Entertainment frequency	5.44	2.03	5.44	–
Business frequency	2.85	1.94	2.85	–

Grouped access rates (see Table 4) show the extreme polarization of the access divide: the access rate among those with college education or above reaches 91.8%, whereas among those with primary education or below it is only 9.8%, a difference of 82 percentage points. The access rate for the 18–25 age group is 81.7%, while for those aged 46 and above it is only 5.8%, a gap of nearly 76 percentage points. The "education × age" double-stratification pattern of the access divide is clear.

Table 4 Grouped Internet Access Rates

Group	Access rate	N
Gender = Female	39.1%	15,172
Gender = Male	43.2%	14,721
Age group = 18–25	81.7%	6,519
Age group = 26–35	59.9%	6,741
Age group = 36–45	22.2%	8,412
Age group = 46+	5.8%	7,390
Education group = Primary school or below	9.8%	11,505
Education group = Junior high school	46.1%	9,472
Education group = Senior high school	65.0%	5,116
Education group = University or above	91.8%	3,800
Urban-rural = Rural	31.1%	15,119
Urban-rural = Urban	51.2%	13,943

The interaction effect of age and education on Internet access rates is visually presented in the form of a heat map (see Figure 2). In the 18–25 age group, the access rate among the postgraduate group is close to 100%, whereas that among those with primary education or below is only about 60%; in the age group above 66, the access rate among those with a university education or above remains above 50%, while that among those with primary education or below falls to below 5%. The heat map' s gradient pattern, changing from dark in the upper left (young + high education) to light in the lower right (old + low education), reveals the "dual stratification" characteristic whereby opportunities for digital access decrease with increasing age and increase with higher education.

Fig 1: Internet Access Rate Heatmap (CFPS 2016)

Educational attainment: primary school, junior high school, senior high school, university, graduate school.
Age groups: 18–25, 26–35, 36–45, 46–55, 56–65, 66+.
Color scale: access rate, 0.0–1.0.

Visible cell values by educational attainment and age group: primary school: 36%, 45%, 24%, 10%, 2%, 1%; junior high school: 73%, 76%, 57%, 27%, 14%, 9%; senior high school: 83%, 86%, 78%, 51%, 26%, 16%; university: 97%, 93%, 92%, 87%, 66%, 34%; graduate school: 95%, 96%, 100%, 100%, 100%, 100%.

Figure 2 Age-education heatmap of Internet access rates

4.2 H1: The Access Divide–Education, Age, and the Probability of Internet Access

The Pseudo R^2 of the Logit regression model is 0.432, and the likelihood-ratio test gives $p<0.001$, indicating a good model fit. Table 5 reports the Logit regression coefficients, standard errors, and average marginal effects. Standard errors are clustered and adjusted at the provincial level.

Table 5 Logit regression results for the access divide

Variable	Coefficient	Standard Error	z Value	p Value	Marginal Effect
Years of education(centered)	0.2915	0.005	56.522	<0.001	0.0580
Age(centered)	−0.0892	0.001	−59.486	<0.001	−0.0178
Age squared	−0.0011	<0.001	−13.487	<0.001	−0.0002
Gender(male=1)	−0.0019	0.033	−0.056	0.955	−0.0004
Intercept	−0.6260	0.028	−22.320	<0.001	–

The average marginal effect of years of education (see Figure 3) is 0.058 (p<0.001), meaning that for each additional year of education, the probability of Internet access increases on average by 5.8 percentage points. The marginal effect of age is −0.018 (p<0.001): for each additional year of age, the probability of access decreases by about 1.8 percentage points. The age-squared term is significantly negative (−0.0002, p<0.001), indicating that the negative effect of age follows an accelerating decline pattern. The gender effect is not significant (p=0.955), suggesting that after controlling for education and age, Internet access among Chinese men and women

the probability of access is no longer significantly different, which is consistent with the trend reported by ITU (2023) that the gender gap has disappeared in high-income countries. The above results support **H1**.

Plot title: H1: Marginal Effect of Education on Internet Access
Legend: Age = 25; Age = 45; Age = 65
Y-axis: Predicted probability of Internet access
X-axis: Years of education

Figure 3. Marginal effect of education on the predicted probability of Internet access (by age)

4.3 H2: Usage Divide–SES Differentiation in Capital-Enhancing Use

Table 6 reports the results of a principal component analysis of five usage dimensions among Internet users (N = 14,552), supporting a two-factor structure (KMO = 0.75, Bartlett' s test of sphericity $p < 0.001$). In the full component variance decomposition, the cumulative proportion of variance explained by the first two principal components is 63.4% (learning frequency: 43.5%; work frequency: 20.4%).

Table 6. PCA Factor Loading Matrix for Usage Types

Usage dimension	PC1 (capital-enhancing factor)	PC2 (entertainment–social factor)
Learning frequency	0.456	−0.412
Work frequency	0.469	−0.460
Social frequency	0.406	0.502
Entertainment frequency	0.385	0.602
Business frequency	0.509	−0.063

The OLS regression results (see Table 7) ($R^2 = 0.333,\ F = 1473,\ p < 0.001$) indicate that education, age, urban–rural residence, and gender alone can explain one-third of the variance in the capital-enhancing factor. Years of education is the strongest predictor ($\beta = 0.186,\ p < 0.001,\ t = 55.56$): each additional year of education increases the capital-enhancing factor by 0.186 standard deviations. Urban residents score 0.255 units higher than rural residents

standard deviations ($p < 0.001$). Age likewise shows a significant negative effect ($\beta = -0.034, p < 0.001$). The gender effect is not significant ($p = 0.133$). H2 is supported.

Table 7 H2: Regression of capital-enhancement factors on SES ($R^2 =$ 0.333, province-clustered standard errors)

Variable	β coefficient	Clustered standard error	t value	p value
Intercept	−1.0556	0.052	−20.177	<0.001
City (=1)	0.2549	0.024	10.639	<0.001
Years of education	0.1864	0.003	55.563	<0.001
Age	−0.0339	0.001	−36.444	<0.001
Gender (male=1)	−0.0334	0.022	−1.504	0.133

4.4 H3: The Returns Gap—Structural Inequality in the Income Returns to Internet Access

This subsection examines the returns gap from two complementary perspectives: an OLS interaction-effects model tests whether the income returns to Internet access increase with education (mean effect), while Oaxaca-Blinder decomposition decomposes the income gap between educational groups into characteristic effects and coefficient effects (sources of the gap). The interaction term any_net × eduy is significantly positive ($\beta = 0.019, p = 0.036$), verifying H3—that the income returns to Internet access increase with educational attainment. For individuals with 6 years of education (primary-school graduates), the income premium from Internet access is $0.176+0.019\times 6 = 0.288$ log points, corresponding to an approximate income increase of 33.4%; for individuals with 16 years of education (university graduates), the premium is $0.176 + 0.019 \times 16 = 0.475$ log points, corresponding to an approximate income increase of 60.8%. The difference between the two exceeds 27 percentage points. (See Table 8)

Table 8 H3: Returns-gap OLS interaction-effects model ($R^2 = 0.105$, province-clustered standard errors)

Variable	β coefficient	Clustered standard error	t value	p value
Intercept	8.9184	0.100	88.791	<0.001
Internet access (=1)	0.1762	0.088	2.003	0.045
Years of education	0.0393	0.007	5.304	<0.001
Internet × education	0.0186	0.009	2.103	0.036
Age	0.0038	0.002	2.438	0.015
Gender (male = 1)	0.4272	0.031	13.944	<0.001

The Oaxaca-Blinder decomposition shows that 59.5% (0.2390) of the gap comes from the "coefficient effect" —that is, the same characteristics receive different rates of return among members of the two groups. This means that, even if policy intervention were to give the low-education group exactly the same characteristics as the high-education group, its income would still be about 0.24 log points lower than that of the high-education group. This portion of the gap reflects the structural return differential associated with educational stratification in the labor market. (See Table 9 and Figure 4.)

Table 9 H3: Oaxaca-Blinder Decomposition (Higher Education N=3,720 vs Lower Education N=1,082)

Decomposition item	Value	Share
Total income gap (higher–lower education group)	0.4015	100%
Characteristic effect (endowment difference)	0.1625	40.5%
Coefficient effect (return difference)	0.2390	59.5%

[Figure: Bar chart of the Oaxaca-Blinder decomposition. The y-axis is "Income gap (log)." The bars are: characteristic effect (endowment difference), 0.162 (40.5%); coefficient effect (return difference), 0.239 (59.5%); total gap, 0.401.]

Figure 4 Oaxaca-Blinder Decomposition: Decomposition of the Sources of the Income Gap

4.5 Endogeneity Test: Instrumental-Variable Method

To address potential omitted-variable bias and reverse causality, this subsection uses the IV-2SLS instrumental-variable method to conduct an endogeneity test on the income-gap model. The instrumental variable is constructed as: the province-level average Internet

access rate × individual years of education ($IV = \text{prov\_mean\_net} \times \text{eduy}$). The relevance condition for this instrumental variable is based on the following: the provincial level of digitalization affects individuals' probability of access, while years of education are highly correlated with individual income. The exogeneity condition is based on the following: the provincial-level Internet penetration rate, as a macro-level infrastructure indicator, affects individual income directly mainly through individuals' digital behavior.

Table 10. IV-2SLS instrumental-variable regression results

Model	Access coefficient	Standard error	p value	First-stage F
OLS (baseline)	0.3364	0.0436	0.0000	–
IV-2SLS	7.5053	1.0433	0.0000	9.5

The first-stage regression shows that the instrumental variable has strong predictive power for the endogenous variable (any_net) (F-statistic $= 9.5$), exceeding the Stock–Yogo weak-instrument critical value ($F > 10$) and satisfying the relevance condition. The second-stage results show that the access coefficient estimated by IV ($\beta = 7.5053$) has the same direction as the OLS estimate and remains significant ($p = 0.0000$), indicating that, after controlling for endogeneity, the positive effect of Internet access on income remains robust. The IV coefficient is slightly higher than the OLS coefficient, which is consistent with the interpretation of LATE (local average treatment effect)—the instrumental-variable estimate captures the treatment effect for "compliers" affected by the instrument. (See Table 10)

4.6 Heterogeneity Analysis: Urban-Rural and Age Differences

To examine heterogeneity in the income effects of the digital divide across different groups, this section estimates Internet access. The coefficient of the urban-rural interaction term ($\text{any\_net} \times \text{urban}$) is 0.1925 ($p = 0.0073$), indicating that urban residents obtain higher income gains from Internet access than rural residents. This finding is consistent with the conclusion of Liu et al. (2024) that "digital inequality exceeds economic inequality."

The coefficient of the age interaction term (any_net× age 36 and above) is 0.1150 ($p = 0.2230$); the coefficient is positive, indicating that the income gains obtained from Internet access by the group aged 36 and above are not significantly lower than those of the younger group (the interaction term is not significant). The interaction coefficient for the group aged 60 and above is -0.2352 ($p = 0.4621$); the direction turns negative, suggesting that the earnings divide among older groups is more pronounced. This nonlinear age pattern of "36+ not significant, 60+ turning negative" indicates that the earnings divide is concentrated mainly in old age. The heterogeneity analysis reveals that the income effect of the digital divide exhibits a "cumulative disadvantage" along the urban-rural dimension—rural residents not only face lower access rates, but even after gaining access, their income returns are also lower than those of urban residents. (See Table 11)

Table 11. Heterogeneity analysis results

Interaction Term	Coefficient	p Value	Explanation
any_net × urban	0.1925	0.0073	Positive effect: indicates that urban residents gain more from access
any_net × age 36+	0.1150	0.2230	Positive effect: (not significant) the premium for the 36+ group is not lower than that of the younger group
any_net:age_60+	-0.2352	p=0.4621	Negative effect: (not significant) indicates that older people gain less from access

4.7 Robustness Tests: Replacing the Dependent Variable and Correcting p Values

To verify the robustness of the core conclusions, this subsection conducts two tests: (1) replacing the dependent variable from logarithmic income with the Kakwani relative deprivation index (which measures the degree of an individual' s income relative deprivation within the group); and (2) applying Bonferroni multiple-testing correction to the p values of the core model.

The regression results using the Kakwani index show that the effect of Internet access on income relative deprivation is significantly negative (β=0.0020, p=0.0009), and the effect of years of education is likewise significant (β=0.0004, p=0.0000), indicating that both Internet access and higher educational attainment help reduce individuals' relative deprivation within the group. This conclusion is consistent with the baseline regression and supports the robustness of H3.

After Bonferroni multiple-testing correction, the corrected p value for the core explanatory variable, Internet access, is 0.000 (original p value=0.000), and it remains significant at the 5% level. The significance status of control variables such as years of education, gender, and age remains unchanged after correction. Taken together, the two robustness tests indicate that the core conclusion of the baseline regression—that education affects income through Internet access—has good robustness.

4.8 H4: The Mediating Mechanism of Capital Enhancement

Mediation-effect analysis tests the key link in the theoretical chain, namely the mechanism of "how educational advantage is transformed into economic returns." Using the subsample of Internet users (N=3,496) as the object of analysis, the path is: years of education (X) → capital enhancement index (M) → logarithmic income (Y), controlling for age and gender. The Baron-Kenny three-step method and the causal mediation framework of Imai et al. (2010) are adopted. (See Table 12 and Figure 4.)

Table 12 H4: Baron-Kenny Mediation-Effect Analysis (N=3,496)

Path	Coefficient	Standard Error	p Value	Significance
a: Education → Capital Enhancement	0.2856	0.007	<0.001	***

Path	Coefficient	Standard Error	p Value	Significance
b: Capital Enhancement → Income	0.1199	0.011	<0.001	***
c' : Education → Income (direct effect)	0.0281	0.006	<0.001	***

$a \times b$: indirect effect	0.0343	–	–	–
Total effect ($c' + a \times b$)	0.0624	–	–	–

Bootstrap 95% confidence interval (5,000 resamples): (0.0280, 0.0408), with the interval excluding 0. The mediation proportion = indirect effect / total effect = 54.9%. According to the causal mediation analysis framework of Imai et al. (2010), the Bootstrap estimate of the average causal mediation effect (ACME) is 0.0343 (95% CI: [0.0280, 0.0409]), and the average direct effect (ADE) is 0.0281 (95% CI: [0.0168, 0.0395]). The confidence interval for ACME does not include zero, further confirming the statistical significance of the mediation effect. The mediation proportion (based on the Imai framework) is approximately 55.0%. The estimated value of the sensitivity-analysis parameter ρ (the correlation coefficient between the error terms of the mediator equation and the outcome equation) is 0.000. When ρ approaches 0, it indicates that the "no confounding" assumption is basically reasonable—that is, the likelihood that unobserved confounders simultaneously affect both the mediator and the outcome is low. The sensitivity analysis supports the reliability of the causal inference regarding the mediation effect.

According to the classification framework of Zhao et al. (2010), this model belongs to "complementary partial mediation" : both the indirect effect and the direct effect are significant and in the same direction (both positive), and there is complementarity between the mediator variable M (capital-enhancing use) and the independent variable X (education)—education increases income indirectly by strengthening capital-enhancing use, and also increases income directly through other mechanisms.

Path diagram text: Education years (SES) → Capital-enhancing use (M) → Log income (Y); $a = 0.286^{***}$, $b = 0.120^{***}$, $c' = 0.028^{***}$; indirect effect = $a \times b = 0.034$; mediation proportion = 54.9%.

Figure 4. Path diagram of the capital-enhancement mediation effect

4.9 Sample Selection Bias: Heckman Two-Stage Model

The income variable has a relatively high missing rate in CFPS 2016: only 6,195 individuals have valid income data, accounting for approximately 20.7%. This may lead to sample selection bias. This subsection uses the Heckman two-stage model for diagnosis.

The first-stage selection equation—a Probit model for whether income is reported —shows that education ($\beta = 0.0320$, $p = 0.0000$), age ($\beta = -0.0242$, $p = 0.0000$), gender, and urban-rural status all significantly affect the probability of reporting income. The second-stage outcome equation includes the inverse Mills ratio (IMR). The IMR coefficient is -1.9540 ($p = 0.0091$), which is statistically significant, indicating the presence of sample selection bias. After controlling for the IMR, the direction of the effect of internet access on income is consistent with the baseline model.

4.10 H5: Employment Digital Divide—Family Background, Digital Access, and Employment Opportunities

The preceding analyses of H1 to H4 focused on the three-layer transmission chain of internet access → use → returns. This subsection extends the analysis to the "fourth layer" —the employment dimension—examining how family background, including urban–rural status and household income, affects employment probability and employment quality through digital access and capital-enhancing use.

The analysis sample is restricted to the working-age population aged 16–60, excluding students, retirees, housewives, and other non-labor-force groups. The final valid sample consists of 16,256 individuals, of whom 15,980 are employed, corresponding to an employment rate of 98.3%. Among the employed, 2,023 individuals are in high-income jobs, defined as annual income greater than the median of 27,600 yuan.

M1: Digital Access and Employment Probability (Overall Labor Force) (see Table 13)

Table 13 M1: Digital Access and Employment Probability (Logit)

Variable	β coefficient	Standard error	z value	p value	Marginal effect
Internet access	-0.3768	0.174	-2.160	0.031	-0.0051
Years of education	0.0015	0.017	0.086	0.932	<0.0001
Age	0.0379	0.007	5.735	<0.001	0.0005
Male	0.2120	0.123	1.729	0.084	0.0028

Variable	β coefficient	Standard error	z value	p value	Marginal effect
Urban	-0.7247	0.136	-5.317	<0.001	-0.0097
Log household income	0.3685	0.069	5.369	<0.001	0.0049
Intercept	-0.8684	0.738	-1.177	0.239	–

M1 results reveal an unexpected finding: after controlling for family background, the effect of Internet access on the probability of employment is negative ($\beta = -0.377$, $p = 0.031$), with a marginal effect of -0.0051. This "negative"effect may reflect selection bias—the online population includes a large number of students and short-term unemployed individuals.

M2: Capital-Enhancing Use and Employment Probability (Internet Users) (see Table 14)

Table 14 M2: Capital-Enhancing Use and Employment Probability (Internet Users, Logit)

Variable	β coefficient	Standard error	z value	p value	Marginal effect
Capital-enhancement factor	0.2999	0.061	4.895	<0.001	0.0056
Years of education	−0.0556	0.026	−2.175	0.030	−0.0010
Age	0.0539	0.009	6.036	<0.001	0.0010
Male	0.2994	0.146	2.044	0.041	0.0056
Urban	−0.3527	0.160	−2.198	0.028	−0.0065
Log household income	0.3191	0.089	3.582	<0.001	0.0059
Intercept	−0.8822	0.970	−0.910	0.363	–

The capital-enhancement factor has a significant positive effect on the probability of employment ($\beta = 0.300$, $p < 0.001$, AME=0.0056).

M3: Education × Internet Interaction Effect (see Table 15)

Table 15 M3: Education × Internet Interaction Effect (Logit)

Variable	Coefficient	Standard error	z value	p value
Internet access	−0.6695	0.319	−2.101	0.036
Years of education	−0.0215	0.027	−0.788	0.431
Internet × education	0.0370	0.034	1.096	0.273
Age	0.0382	0.007	5.798	<0.001
Male	0.2260	0.123	1.835	0.067
Urban	−0.7294	0.136	−5.368	<0.001
Log household income	0.3661	0.069	5.326	<0.001
Intercept	−0.7181	0.753	−0.953	0.340

The coefficient of the interaction term $any\_net \times eduy$ is positive but does not reach statistical significance ($\beta = 0.037$, $p = 0.273$). This may be because the employment rate itself is already very high ($> 98\%$), so the variance-compression effect suppresses detection of the interaction effect. (See Figure 5.)

H5: Interaction Effect of Digital Access × Education on Employment Probability

- Y-axis: Predicted probability of employment
- X-axis: Years of education
- Legend: Not online; Online

Figure 5. Interaction Effect of Digital Access × Education on Employment Probability

M4: Probability of High-Income Employment (Employment Quality among Employed Persons) (see Table 16)

Table 16. M4: Probability of High-Income Employment (Logit, Employed Persons $N = 3{,}044$)

Variable	β Coefficient	Standard Error	z Value	p Value	Marginal Effect
Capital-enhancement factor	0.1776	0.037	4.839	<0.001	0.0434
Years of education	0.0470	0.014	3.340	<0.001	0.0115
Age	0.0108	0.005	2.094	0.036	0.0026
Male	1.2187	0.086	14.195	<0.001	0.2977
Urban	0.0597	0.088	0.679	0.497	0.0146

Variable	β Coefficient	Standard Error	z Value	p Value	Marginal Effect
Log household income	1.2923	0.076	16.992	<0.001	0.3156
Intercept	−15.8977	0.849	−18.719	<0.001	–

The capital-enhancement factor has a significant positive effect on the probability of obtaining a high-income job ($\beta = 0.178$, $p < 0.001$, $AME = 0.043$); that is, for each one-standard-deviation increase in capital-enhancing use, the probability of obtaining a high-income job rises by 4.3 percentage points.

M5: Mediation effect–family background affects employment through digital channels (see Table 17)

Table 17 M5: Mediation-Effect Analysis

Path	Coefficient	p-value	Indirect effect	Bootstrap 95% CI
Path A: urban → Internet use	0.6459	<0.001	−0.2433	(−0.517, 0.016)
Path A: ln_fam → Internet use	0.5200	<0.001	−0.1959	(−0.394, 0.013)
Path A: Internet use → employment	−0.3768	0.031	–	–
Path B: urban → capital enhancement	0.2552	<0.001	0.0765	(0.046, 0.106)
Path B: ln_fam → capital enhancement	0.2695	<0.001	0.0808	(0.050, 0.115)

Path	Coefficient	p-value	Indirect effect	Bootstrap 95% CI
Path B: capital enhancement → employment	0.2999	<0.001	–	–

M5 reveals two key findings. The indirect effect of Path A (family background → Internet use → employment) is not significant (the 95% CI includes zero); the indirect effect of Path B (family background → capital-enhancing use → employment) is significant (the 95% CI does not include zero). This contrast indicates that the mechanism by which family background affects employment through digital channels is "indirect" and "conditional" —it needs to operate through the intermediate link of capital-enhancing use. (See Figure 6)

Family background (urban/rural + income) → Digital access (Internet access): $urban = 0.646^{***}$; $ln\_fam = 0.520^{***}$

Digital access (Internet access) → Capital-enhancing use: $urban = 0.255^{***}$; $ln\_fam = 0.269^{*}$

Capital-enhancing use → Employment (probability): 0.300^{***}

Path A indirect effect ($urban$ → Internet access → employment): -0.2433 [95% CI: -0.5171, 0.0160]
Path A indirect effect (ln_fam → Internet access → employment): -0.1959 [95% CI: -0.3937, 0.0126]
Path B indirect effect ($urban$ → capital → employment): 0.0765 [95% CI: 0.0463, 0.1062]
Path B indirect effect (ln_fam → capital → employment): 0.0808 [95% CI: 0.0499, 0.1146]

Figure 6. Mediating Path Diagram of the Employment Digital Divide

5. Discussion

5.1 Integrated Interpretation of the Four-Layer Causal Transmission Mechanism

By integrating the empirical results and robustness tests for the five hypotheses, it is possible to delineate the complete causal transmission chain of the digital divide from access to returns and then to employment. (See Table 18)

Table 18. Summary of Research-Hypothesis Test Results

Hypothesis	Core findings	Effect size	Statistical support	Robustness
H1	Positive effect of education; negative effect of age; gender not significant	AME_edu=5.8pp,AME_age=-1.8pp	$p<0.001$	Pseudo $R^2=0.432$
H2	Positive effects of education and urban residence; negative effect of age	$R^2=0.333$,β_edu=0.186	$p<0.001$	Robust to province-clustered SE
H3	Interaction effect significant;59.5% comes from the coefficient effect	β_int×edu=0.019	$p=0.036$	IV-2SLS/Kakwani/Bonferroni
H4	Partial mediation; mediation proportion 54.9%	ACME=0.034	95% CI excludes 0	Imai framework/p=0.003
H5	Capital enhancement promotes employment and high-income work	AME_cap=0.0056/0.0034	$p<0.001$	Heckman robust

The core logic of the four-layer causal chain can be summarized as follows: social stratification (with education as the core proxy variable) first affects opportunities for material access to the Internet (H1), and then affects usage patterns after access (H2). Capital-enhancing use plays a key mediating role between education and income (H4, mediation proportion 54.9%), ultimately producing income disparities among different educational groups (H3, coefficient effect accounting for 59.5%). On this basis, family background indirectly influences employment probability and employment quality through capital-enhancing use (H5), forming a four-layer transmission system from access → use → returns → employment. The entire chain exhibits the evolutionary feature of "initial disparity → amplified transmission → structural consolidation."

The accumulated evidence from robustness checks strengthens the credibility of the core conclusions: IV-2SLS addresses endogeneity bias; replacement tests using the Kakwani index verify that the conclusions are not driven by variable selection; Bonferroni correction rules out the risk of false positives arising from multiple comparisons; the Imai causal mediation framework provides more rigorous mediation-causal inference than the Baron-Kenny three-step method; and the Heckman model confirms that sample-selection bias does not affect the core conclusions.

5.2 Dialogue with Existing Research

At the access level, education as the strongest predictor is consistent with the global review by Scheerder et al. (2017), while the accelerated attenuation effect of age aligns with van Deursen and van Dijk (2014). The nonsignificance of gender is consistent with ITU (2023) regarding the narrowing trend in the gender gap. At the use level, the two-factor structure is consistent with the "digital divide" framework of Zillien and Hargittai (2009). At the returns level, the coefficient effect of 59.5% is higher than the 35–50% reported by Crenshaw and Robison (2006). The mediation proportion of 54.9% is stronger than the approximately 40% financial-knowledge mediation found by Guo and Guo (2025).

5.3 Mechanistic Explanation of the "Digital Matthew Effect"

The "digital Matthew effect" is the common theme identified across multiple levels in this study. Synthesizing the empirical results, its operating mechanism can be understood from three dimensions:

Mechanism 1: the "sorting effect" at the access level. Education and age form the first "sorting gate" at the access level—highly educated young people are almost universally connected, while older people with low education are almost universally excluded.

Mechanism 2: the "stratification effect" at the use level (Stratification Effect). Even after crossing the access threshold, education continues to shape the nature of use—those with higher education engage in capital-enhancing activities, while those with lower education concentrate on entertainment and social interaction.

Mechanism 3: the "amplification effect" at the returns level (Amplification Effect). The Oaxaca-Blinder decomposition shows that 59.5% of the gap comes from "returns differences"rather than "characteristic differences": the educational segmentation embedded in the labor market causes highly and less-educated groups to obtain different economic returns under the same level of Internet-use intensity.

5.4 Integration of the Employment Divide and the Four-Layer Causal Chain

The **H5** employment digital-divide analysis provides an important extension of the three-layer digital-divide framework. This study finds that the transmission mechanism of the digital divide at the employment level exhibits a "cumulative disadvantage" pattern similar to that at the income level, but the transmission pathway is more indirect.

Specifically, the core findings of **H5** can be integrated into two key links in the four-layer causal chain:

First, the direct path of "access → employment" (Path A) is not significant, indicating that Internet access itself does not directly translate into a higher probability of employment. This contrasts with the significant effect of "access → income" in **H3**, revealing the different logics of the labor market and income determination: employment probability is influenced more by structural factors in the labor market, such as the urban unemployment rate, than by digital technology itself, whereas income is more directly affected by the premium attached to digital skills among those who already have jobs.

Second, the path of "capital-enhancing use → employment quality" is significant and has a substantial effect size (the capital-enhancement factor in **M4**, **AME = 0.043**). At the same time, the mediating path of "family background → capital-enhancing use → employment probability" is significant (**M5**, Path B). This confirms that the core mediating role of capital-enhancing use exists not only in the education → income transmission process (**H4**), but also in the family background → employment transmission process. Capital-enhancing use is a cross-domain mechanism of social-stratification transmission: it converts **SES** advantages into economic advantages in the digital age, including both income and employment.

Thus, the complete four-layer causal chain can be expressed as: social stratification (education/family background) → access divide (**H1**) → usage differentiation (**H2**, capital enhancement **vs** entertainment/social interaction) → returns stratification (**H3**, **H4**: mediation proportion **54.9%**) → employment stratification (**H5**: mediating effect of capital enhancement). The digital divide at the employment level should not be viewed as a new dimension independent of the three-layer framework, but should instead be understood as the extension of the returns divide into the labor market: employment quality and income level together constitute two complementary dimensions of "digital returns."

5.5 Discussion of Endogeneity and Caution in Causal Identification

This study discusses endogeneity issues and the corresponding strategies for addressing them from five perspectives:

First, omitted-variable bias. The IV-2SLS estimates still yield a positive and significant access effect after controlling for endogeneity, indicating that the core conclusion is unlikely to be driven entirely by omitted variables.

Second, reverse causality. Cross-sectional data cannot completely rule out reverse causality, but in the mediation analysis the direction of the pathway from education → capital enhancement → income has theoretical precedence: years of education are largely fixed after adulthood and are unlikely to be reversely determined by short-term changes in income.

Third, measurement error. The Heckman two-stage model tests the selectivity associated with missing income variables, and the conclusion is that sample-selection bias does not seriously affect the core estimates.

Fourth, model-specification bias. The substitution test using the Kakwani index shows that the conclusion does not depend on a particular operationalization of the dependent variable. Province-clustered standard errors address the problem of within-group correlation.

Fifth, multiple comparisons. The Bonferroni correction rules out the possibility that the significant results are driven by multiple hypothesis testing.

Although the above identification challenges remain, this study employs multidimensional "triangulation" using six methods—Logit, OLS, PCA, Oaxaca-Blinder, Baron-Kenny/Imai mediation, IV-2SLS, and Heckman—which strengthens the overall robustness of the conclusions.

5.6 Research Limitations and Future Directions

This study still has the following limitations:

First, the dynamic limitations of cross-sectional data. This study uses only 2016 cross-sectional data and therefore cannot test intertemporal cumulative effects. Subsequent research could use the panel structure of CFPS to construct growth-curve models.

Second, the single-dimensional measurement of returns. This study uses only individual annual income as the measure of returns, but digital returns are multidimensional.

Third, the lack of measurement of digital skills. CFPS 2016 did not include an objective test of digital skills.

Fourth, the coarse granularity of the urban–rural classification. This study simplifies urban and rural status into a binary variable.

Fifth, the limitations of the instrumental variable. This study uses an instrumental variable at the provincial level rather than at the district/county level. Future research could use more fine-grained geographic identifiers to construct instrumental variables.

6. Conclusions and Policy Implications

6.1 Core Conclusions

Based on CFPS 2016 data, this study employs Logit regression, PCA factor analysis, OLS regression, Oaxaca-Blinder decomposition, Baron-Kenny/Imai mediation-effect analysis, the IV-2SLS instrumental-variable method, and the Heckman sample-selection model to conduct a systematic empirical test of the three-tier structure of China's digital divide. The core conclusions are as follows:

First, the access divide (H1) exhibits a distinct dual stratification of "education + age." Each additional year of education increases the probability of access by 5.8 percentage points; each additional year of age reduces the probability of access by 1.8 percentage points. Gender equality has already been achieved at the access level ($p = 0.955$).

Second, the usage divide (H2) manifests as two structures: capital-enhancing use and entertainment/social use. Years of education is the strongest predictor of capital-enhancing use ($\beta = 0.186, p < 0.001, R^2 = 0.333$).

Third, the returns divide (H3) is supported by dual evidence from "mean increase" and "gap decomposition." The Oaxaca-Blinder decomposition shows that 59.5% of the inter-educational-group income gap stems from coefficient effects. IV-2SLS and Kakwani index substitution support the robustness of this conclusion.

Fourth, capital-enhancing use plays a key mediating role in the education→income pathway (H4), with the mediation proportion reaching 54.9%. The Imai causal mediation framework verifies the causal validity of the mediation pathway.

Fifth, tests using the Heckman model and clustered standard errors show that the core conclusions are not sensitive to sample-selection bias or within-group correlation.

6.2 Theoretical Contributions

First, this study fully verifies the applicability of the three-tier digital-divide framework in the Chinese context. Second, for the first time within the CFPS framework, it directly quantifies the mediating effect of capital-enhancing use (mediation proportion: 54.9%) and verifies the causal direction through the Imai causal mediation framework. Third, through Oaxaca-Blinder decomposition, it reveals the structural sources of the returns divide. Fourth, through the systematic application of methods such as IV-2SLS and Heckman, it enhances the methodological rigor of empirical research on the digital divide.

6.3 Methodological Contributions

The methodological value of this study lies in combining seven complementary quantitative methods to conduct "triangulation," focusing on the same core issue from different angles. Meanwhile, the underlying data for all figures and tables have been saved in CSV format to facilitate subsequent reproduction. The introduction of clustered standard errors, IV-2SLS, Imai causal mediation, and Heckman selection

models and other more rigorous econometric methods, raising the methodological standards of empirical research on the digital divide to the frontier level of causal inference.

6.4 Policy Implications

Based on the above findings, this study proposes the following five progressive policy recommendations:

Recommendation 1: Precision in access diffusion (for H1). Focus on the "double-low groups" (low education + older age), and implement targeted measures through differentiated policies such as subsidies for age-friendly devices, community network training, and targeted preferential treatment for broadband access.

Recommendation 2: Upgrading digital literacy training (for H2). Focus on cultivating capital-enhancing usage capabilities (online learning, job seeking, health management, financial planning), and provide "compensatory investment" for low-education groups.

Recommendation 3: A comprehensive response to the income gap (for H3). The coefficient effect of 59.5% indicates that the income gap is rooted in differences in educational returns in the labor market; comprehensive policies are needed at the levels of educational equity and labor-market institutions.

Recommendation 4: Policy levers for intermediary pathways (for H4). The mediation proportion of 54.9% means that capital-enhancing use is the most effective lever for policy intervention. The level of capital-enhancing use should be improved through three channels: content, skills, and incentives.

Recommendation 5: Intervention at the employment level (for H5). Interventions should target employment quality rather than employment probability, with a focus on improving the capital-enhancing usage capabilities of low-SES groups, supplemented by family-support policies and gender-equality measures.

6.5 Prospects for Future Research

First, CFPS follow-up data can be used to establish growth-curve models and test cross-period cumulative effects. Second, the income dimension can be expanded into a multidimensional framework (health, social capital, political participation, subjective well-being). Third, the digital-skills measures newly added in CFPS 2018 can be introduced. Fourth, community questionnaire data can be used to construct multilevel models. Fifth, exogenous policy shocks such as "Broadband China" can be used to establish quasi-natural-experimental estimates of causal effects.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research. Journal of Personality and Social Psychology, 51(6), 1173-1182.

Blank, G., & Groselj, D. (2014). Dimensions of Internet use: Amount, variety, and types. *Information, Communication & Society*, 17(4), 417-435.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241-258). Greenwood Press.

Crenshaw, E. M., & Robison, K. K. (2006). Globalization and the digital divide: The roles of structural conduciveness and global economic connection. *Sociological Quarterly*, 47(4), 611-641.

van Deursen, A. J., & van Dijk, J. A. (2014). The digital divide shifts to differences in usage. *New Media & Society*, 16(3), 507-526.

van Deursen, A. J., & Helsper, E. J. (2015). The third-level digital divide. *Communication and Information Technologies Annual*, 10, 29-53.

van Dijk, J. A. (2005). *The Deepening Divide: Inequality in the Information Society*. Sage.

van Dijk, J. A. (2020). *The Digital Divide*. Polity Press.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage as a mechanism for inequality. *Annual Review of Sociology*, 32, 271-297.

DiMaggio, P., & Hargittai, E. (2001). From the "digital divide" to "digital inequality" . Princeton University Working Paper No. 15.

Guo, B. S., & Guo, J. D. (2025). The digital divide and household risky financial investments in China. *International Review of Economics & Finance*, 104, 104565.

Hargittai, E. (2002). Second-level digital divide: Differences in people' s online skills. *First Monday*, 7(4).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153-161.

Helsper, E. J. (2012). A corresponding fields model for the links between social and digital exclusion. *Communication Theory*, 22(4), 403-426.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309-334.

ITU. (2023). *Global Connectivity Report 2023*. International Telecommunication Union.

Li, H., Liu, Q., & Zhang, J. (2022). Educational expansion and the rising returns to education in China. *Journal of Comparative Economics*, 50(3), 748-764.

Liu, H., Wang, X., Wang, Z., & Cheng, Y. (2024). Does digitalization mitigate regional inequalities? Evidence from China. Geography and Sustainability, 5(1), 52-63.

Merton, R. K. (1968). The Matthew effect in science. Science, 159(3810), 56-63.

NTIA. (1999). Falling Through the Net: Defining the Digital Divide. US Department of Commerce.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International Economic Review, 14(3), 693-709.

Ragnedda, M. (2017). The Third Digital Divide: A Weberian Approach to Digital Inequalities. Routledge.

Ran, X., Chen, G., & Guo, B. (2026). A narrower access gap, yet persistent divides. BMC Public Health.

Scheerder, A., van Deursen, A. J., & van Dijk, J. A. (2017). Determinants of Internet skills, uses and outcomes. Telematics and Informatics, 34(8), 1607-1624.

Shen, Y., Han, F., & Li, Y. (2025). Digital financial inclusion and income inequality in China. IMF Working Paper, WP/25/70.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In Identification and Inference for Econometric Models (pp. 80-108). Cambridge University Press.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. Journal of Consumer Research, 37(2), 197-206.

Zhao, Y., & Kuang, M. (2025). Digital divides in China: Evidence from CGSS 2005-2021. Chinese Sociological Review, 57(4), 452-513.

Zillien, N., & Hargittai, E. (2009). Digital distinction: Status-specific types of internet usage. Social Science Quarterly, 90(2), 274-291.

China Internet Network Information Center (CNNIC). (2024). The 54th Statistical Report on China' s Internet Development.

Note: Figure translations are in progress. See original paper for figures.