Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond digital access: how internet use patterns shape labour income returns—evidence from China

Qi Xiao

qi.xia@st.econ.tu.ac.th

Thammasat University

Dingyi Qin

Thammasat University

Article

Keywords: internet use, labour income, second-level digital divide, productive vs. consumptive use, panel fixed effects, China

Posted Date: September 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10710847/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Does internet access raise labour income, or do higher-earning, more capable people simply select into going online? Using the China Family Panel Studies (CFPS), this study argues that what matters is not whether one goes online but how, and decomposes internet use into productive (online learning), mixed (online shopping), and consumptive (online gaming) activities. A cross-sectional internet-adoption premium of roughly 11% is stable across OLS, propensity score matching, a Lewbel instrument, and a Heckman correction—yet it collapses to essentially zero in a multi-wave (2014–2020) individual fixed-effects panel, revealing that the aggregate access premium is largely selection. What robustly predicts income is instead the composition of use: conditional on being online, productive and mixed uses are each associated with about 8–9% higher income (robust to Oster bounds), while gaming is not. These productive-use returns are 2–3 times larger for women and older workers—a redistributive reversal confirmed by interaction tests and false-discovery-rate correction. Suggestive mediation evidence, reported in an online appendix, points to information-acquisition and employment-formalisation channels. The policy implication is sharp: expanding access alone does little for earnings; converting access into productive use, especially among disadvantaged groups, is likely to matter more for an inclusive digital dividend.

1. Introduction

Does going online raise a worker's income? A large empirical literature answers yes: across developing economies internet users systematically out-earn non-users (Galperin & Fernández-Ardèvol, 2019; Ma et al., 2020), and estimates for China are consistently positive (Si & Li, 2023; Yuan & Zhang, 2024). The question matters at scale—by the end of 2024 China counted more than 1.1 billion internet users, a penetration rate of 78.6% (CNNIC, 2025)—and it underwrites large public investments in connectivity. This paper documents a fact that complicates the consensus: in a national household panel, a cross-sectional internet premium of roughly 11%, stable across every standard cross-sectional identification strategy, disappears the moment the comparison is made within individuals.

Yet this literature shares a critical blind spot: it treats 'going online' as a homogeneous treatment, collapsing fundamentally different behaviours—learning, shopping, gaming, social networking—into a single binary indicator. This conflation is problematic both empirically and theoretically. Empirically, if distinct use types have heterogeneous income effects, the average treatment effect from a binary internet variable provides an incomplete picture of the digital dividend. Theoretically, human capital theory and information economics predict that using the internet to acquire productive knowledge or reduce job-search frictions should yield income returns, whereas using it primarily for entertainment may not. A second generation of digital divide research—'Digital Divide 2.0'—has accordingly shifted attention from mere access to the quality and purposiveness of use (Hargittai, 2002; van Dijk, 2006; Warschauer, 2003). Despite this conceptual reorientation, evidence on how specific use patterns translate into differential income returns remains sparse.

This paper fills that gap. Using CFPS data, we decompose internet use into online learning (productive use), online shopping (mixed use), and online gaming (consumptive use), and bring two complementary forms of identification to bear: a battery of cross-sectional estimators on the 2020 wave (OLS, propensity score matching, a community-level instrument, a Lewbel instrument, and a Heckman selection correction), and a four-wave 2014–2020 individual fixed-effects panel that differences out time-invariant unobservables.

Our analysis yields four main findings. First, the aggregate internet-adoption premium is largely selection: although every cross-sectional estimator places it near 11%, it falls to zero—or below—once we add individual fixed effects in the 2014–2020 panel, indicating that more capable individuals both go online and earn more rather than that connectivity per se raises income. Second, what robustly predicts income, conditional on being online, is the composition of use: online learning and online shopping are each associated with roughly 8.4% and 8.5% higher income, while online gaming shows a precisely estimated zero effect, and these use-type associations survive Oster (2019) coefficient-stability bounds. Third, the returns to productive use are sharply unequal but in a redistributive direction: women and older workers (46–60) obtain roughly 2–3 times higher returns to online learning than men and younger workers—a reversal confirmed by interaction tests and false-discovery-rate correction—while the rural premium points the same way; the access margin, by contrast, weakly favours the advantaged, so the second-level divide operates opposite to the first. Fourth, suggestive mediation evidence (Appendix A) points to two plausible channels—information acquisition and, more robustly, employment formalisation—both interpreted cautiously given their sensitivity to unobserved confounding.

The paper makes three contributions. Methodologically, the paper is organised around an identification contrast: using a four-wave individual fixed-effects panel alongside cross-sectional OLS, PSM, and Lewbel and Heckman estimators, we show that the binary access premium is largely a selection artifact—the one causal claim the design supports—whereas, conditional on being online, it is the composition of use that robustly predicts income, bounded by Oster (2019) coefficient-stability statistics. Empirically, we document a novel reversal: the productive-use return is concentrated among disadvantaged groups—women and older workers—a pattern prior work has missed. Relative to the closest prior studies—cross-sectional analyses of use-type returns such as Yuan and Zhang (2024)—the panel dimension is the decisive addition: it is what exposes the access premium as selection and disciplines the causal interpretation of every use-type estimate. Theoretically, a simple human-capital/matching framework rationalizes the findings—no return to consumptive use, a redistributive reversal in productive-use returns, and a more tentative access-margin gradient—and motivates the two transmission channels we examine. The remainder proceeds as follows: Section 2 reviews related literature and develops a conceptual framework; Section 3 describes data and methods; Section 4 presents results; Section 5 concludes.

2. Literature Review

2.1 Internet Use and Labour Market Outcomes

The labour-economic case for internet-induced income gains rests on three channels. First, the internet lowers job-search costs, improving match quality between workers and employers (Autor, 2001; Kuhn & Mansour, 2014). Second, online platforms enable skill acquisition at lower cost than traditional training (Belo et al., 2014; Fairlie, 2004). Third, broader market information reduces monopsonistic exploitation and strengthens wage-bargaining power (Aker, 2010). Early empirical studies typically found a substantial 'internet premium' in wages, though this partly reflected selection of skilled workers into internet use (DiNardo & Pischke, 1997).

For China, recent studies using national surveys and explicit identification strategies consistently find positive income effects of internet use. Ma et al. (2020), applying an endogenous treatment regression to rural household data, find that internet adoption significantly raises household income and expenditure. Si and Li (2023) use CFPS 2018 and an endogenous switching regression to show that internet use raises the wages of informal workers, while Weng et al. (2024) document that internet use promotes the upward income mobility of rural low-income households through entrepreneurship and non-farm employment. Our OLS-PSM baseline of approximately 10.9% is consistent with this body of evidence. A parallel strand reaches mixed conclusions on distribution: Deng et al. (2023) find that digital-economy development raises urban incomes more than rural incomes, widening the urban-rural gap, whereas Zhang et al. (2023) find that internet use can narrow rural households' income gap by expanding non-farm employment. Our paper reconciles these by showing that the distributional pattern depends critically on use type, not just access.

2.2 The Second-Level Digital Divide: From Access to Use Quality

The 'second-level digital divide' highlights that even among internet users, substantial heterogeneity exists in how digital resources are utilised (Hargittai, 2002; Selwyn, 2004). Van Dijk (2006) distinguished motivational, material, skills, and usage access, arguing that upstream inequalities in skills and motivation generate downstream inequalities in returns. Warschauer (2003) similarly emphasised that technology's social impact depends on surrounding social practices, not connectivity per se.

Empirically, Yuan & Zhang (2024) used CGSS 2017 data to document that online business activities yield higher income returns than online entertainment, with information access and social networking occupying intermediate positions. However, that study relies on cross-sectional OLS without addressing endogeneity, uses a different national survey (CGSS vs. CFPS), and does not examine heterogeneity reversals across demographic groups. Our paper extends this line in three ways: (i) we exploit CFPS 2020's richer internet-use module; (ii) we bring panel fixed effects and Oster (2019) bounds to bear, showing that the aggregate access premium is largely selection while the use-type effects are robust; and (iii) we document the novel within-group reversal that has escaped prior detection.

The productive/consumptive distinction connects to the economics of time allocation (Becker, 1965): gaming is a leisure activity with high immediate utility but low returns to labour productivity, predicting zero income effects consistent with our finding. Online learning directly substitutes for formal education

and vocational training, implying returns via the human capital channel. Online shopping occupies a middle ground: beyond consumption, it exposes users to information-processing environments (price comparison, seller evaluation, product research) and may enable platform-based self-employment and informal entrepreneurship, providing plausible channels for the income premium we observe.

2.3 Mechanism Channels and Heterogeneity

Two principal mechanisms link internet use to income. The information channel posits that the internet reduces search costs, broadening workers' information sets regarding available jobs, wage norms, and skill requirements (Kuhn & Mansour, 2014; Si & Li, 2023). The institutional channel proposes that internet use—by connecting workers to formal labour-market platforms and increasing awareness of labour rights—raises the probability of formal employment with associated wage protections (Zhan & Yang, 2024). We examine suggestive mediation evidence on both channels in Appendix A.

On demographic heterogeneity, Zhang et al. (2023) found that internet use narrows rural households' income gap, consistent with a 'digital catch-up' hypothesis. Our results refine this picture: any cross-sectional access advantage for urban workers—itself statistically fragile—is reversed for productive use (online learning), suggesting that catch-up operates through use quality rather than mere connectivity. The age dimension has received less attention; our finding that workers aged 46–60 obtain the largest returns from online learning challenges the view that older workers are inherently ill-positioned to benefit from digital technology.

2.4 A Simple Conceptual Framework

We formalise the intuitions above in a stylized framework that yields the paper's testable predictions. Following the Mincerian tradition, individual i's log labour income depends on effective human capital H_i and on the quality of labour-market matching M_i:

$$ln(w_i) = \alpha + \rho \cdot H_i + \lambda \cdot M_i + \gamma' X_i + \varepsilon_i \quad (M1)$$

Internet use enters through these two intermediate inputs, and its effect is purpose-specific. Decomposing use into productive (P_i, learning), mixed (S_i, shopping), and consumptive (G_i, gaming) activities, productive and mixed use augment human capital and information-based matching, whereas consumptive use does neither: $H_i = H_{0i} + \theta_P \cdot P_i + \theta_S \cdot S_i + \theta_G \cdot G_i$ and $M_i = M_{0i} + \varphi_P \cdot P_i + \varphi_S \cdot S_i + \varphi_G \cdot G_i$, with $\theta_G \approx \varphi_G \approx 0$ and the remaining coefficients positive. Substituting into (M1) gives the reduced-form return to each use type, $\beta_k = \rho \cdot \theta_k + \lambda \cdot \varphi_k$. This delivers Prediction 1: $\beta_P, \beta_S > 0 = \beta_G$—productive and mixed uses raise income while consumptive use does not.

The reversal follows from letting the marginal human-capital yield of productive use decline in a worker's baseline information access A_i (the ease with which they already obtain job, wage, and skill information through conventional channels): $\theta_P = \theta - \psi \cdot A_i$, $\psi > 0$. Groups historically starved of such access—rural residents, women, and older workers—have low A_i and therefore a higher marginal return to online learning, yielding Prediction 2: $\partial\beta_P / \partial A < 0$, the productive-use premium is larger for low-access

(disadvantaged) groups. Education is the exception: low schooling also lowers the rate δ at which online content is converted into human capital (θ_P scaled by δ(Educ), δ' > 0), so for the education margin the access effect and the conversion effect offset, predicting no reversal.

Finally, because advantaged workers retain greater absorptive capacity and better baseline matching (higher M_0 and ρ), the return to mere connectivity—the binary access margin—can still favour them. This produces Prediction 3: a Matthew pattern in aggregate access returns can coexist with the disadvantaged-favouring reversal in productive-use returns. Sections 4.3–4.4 take these predictions to the data. Prediction 3 is the most tentative of the three: because the aggregate access effect is itself largely selection (Section 4.2), the access-margin gradient admits only an exploratory reading rather than a sharp test.

3. Data and Methodology

3.1 Data Source and Sample Construction

This study draws principally on the 2020 wave of the China Family Panel Studies (CFPS), administered by the Institute of Social Science Survey at Peking University, and supplements it with the 2014, 2016, and 2018 waves to form a four-wave individual panel. CFPS is a nationally representative, stratified, multistage probability-sampled panel survey covering all 31 mainland provincial-level administrative units (including provinces, municipalities, and autonomous regions). The 2020 wave includes a dedicated internet-use module that records not only whether respondents go online but also their specific use purposes and the devices employed—making it uniquely suited to our research question.

We restrict the analysis to working-age individuals (16–60 years) who report positive annual labour income (emp_income > 0). We further exclude observations with missing or anomalously coded values (≤ 0) for education, health, marital status, or urban–rural residence, and drop individuals with a missing community identifier (fid_cid20), which is required to construct the instrument. After these restrictions, the analytic sample comprises 6,948 individuals spanning 31 provinces and approximately 1,000 communities. The full derivation is: 28,530 adult respondents; 19,707 after restricting to ages 16–60; 9,551 with positive labour income; 7,845 after dropping missing or anomalously coded education, health, urban–rural, and marital records; and 6,948 after excluding 897 respondents in single-respondent communities, for whom the leave-one-out instrument is undefined. The 16–60 restriction targets the population at working age under Chinese retirement norms; full-time students and retirees are largely excluded by the positive-labour-income requirement, and the self-employed and flexibly employed may follow distinct behavioural patterns—a possibility that Section 4.3 examines directly for the self-employed. Among these, 5,783 (83.2%) report internet use. Within the user group, 28.0% use the internet for learning, 59.9% for shopping, and 26.1% for gaming (unconditional rates across the full sample are 23.3%, 49.9%, and 21.7%, respectively, as reported in Table 1). Table 1 provides full descriptive statistics.

For the fixed-effects analysis we additionally draw the 2014, 2016, and 2018 waves, applying identical sample restrictions and harmonising the derived variables across waves (the internet-adoption indicator and the controls carry consistent definitions; we use the main-job labour-income variable incomeb, which is available in every wave). Stacking the four waves yields 17,232 individuals and 28,781 person-wave observations; 8,088 individuals are observed in two or more waves, of whom 2,168 change online status across waves and thus identify the within-individual estimator. Because CFPS recorded internet use in two non-comparable formats—a frequency-and-importance module in 2014–2018 versus binary purpose indicators in 2020—the cross-sectional use-type (Section 4.3) and heterogeneity (Section 4.4) analyses use the cleaner 2020 binary indicators, while the binary adoption effect is identified both cross-sectionally and within individuals (Section 4.2).[1]

3.2 Variable Definitions

Dependent variable. The outcome is the natural log of annual labour income (ln_income = ln[emp_income]). Log transformation mitigates right-skewness and allows coefficients to be interpreted as semi-elasticities. Zero and negative labour incomes are excluded by the positive-income restriction, and no truncation is applied at baseline; sensitivity to extreme incomes is assessed in Section 4.5 by trimming and by winsorising the income distribution at the 1st and 99th percentiles.

Core explanatory variables. Throughout the paper, 'internet adoption' (equivalently 'access', the first-level margin) denotes whether a respondent goes online at all, and 'internet use' (the second-level margin) denotes what an online respondent does; we keep this distinction consistent. We construct two treatment layers. The first is a binary internet adoption indicator (internet = 1 if the respondent goes online via mobile device [qu201] or computer [qu202]). The second layer decomposes use into three binary purpose indicators, conditional on being online: use_learn (online learning, qu94), use_shop (online shopping, qu92), and use_game (online gaming, qu91). We classify online learning as ‘productive use,’ online gaming as ‘consumptive use,’ and online shopping as ‘mixed use’ on the grounds that shopping can serve both productive purposes (price discovery, market access, platform-based entrepreneurship) and pure consumption. The three indicators are not mutually exclusive—19.6% of users both learn and shop online, and 5.9% report all three activities—so the joint specification in Eq. (2) estimates each activity's partial association holding the others fixed. The pairwise correlations are modest (learning–shopping r = 0.13, learning–gaming r = 0.05, shopping–gaming r = 0.07), indicating that the three indicators are not simply proxies for a single underlying digital-skill factor.

Instrumental variable. To address endogeneity in internet adoption, we construct a community-level leave-one-out (LOO) internet penetration rate (iv_comm): for individual i in community c containing n_c members, iv_comm equals the mean adoption rate of the remaining $n_c - 1$ members. This instrument captures the local peer-use environment that shapes individual adoption decisions through social learning and peer effects, independently of the individual’s own adoption choice.

The exclusion restriction requires that iv_comm affects individual income only through the individual’s own internet adoption. This is plausible for three reasons. First, we control for individual characteristics

that capture own productivity and selection; province fixed effects absorb regional economic conditions (e.g., local broadband infrastructure, labour demand). Second, the LOO construction ensures the instrument is not mechanically inflated by the individual's own adoption, mitigating reverse causality at the instrument level. Third, the remaining threat—peer adoption raising community-level productivity spillovers that directly affect individual wages—is unlikely to be the dominant channel relative to the direct peer-effect path, given that we control for urban/rural status and province-level conditions. This assumption is fundamentally untestable—a limitation shared by all IV applications of this type (Ma et al., 2020); Section 4.6 reports falsification and local-conditions checks that probe it directly. As a supplementary check, a Wu–Hausman endogeneity test confirms that the OLS estimator is inconsistent ($F = 9.03$, $p = 0.003$), consistent with the selection bias we later document; we nonetheless rely primarily on the panel fixed-effects design (Section 4.2) rather than on this instrument. The first-stage F-statistic (116.07) far exceeds the Stock–Yogo (2005) threshold of 10, confirming instrument strength.

Control variables. Following the Mincerian earnings framework (Mincer, 1974), we include years of education (cfps2020eduy), age and age-squared (to allow non-linear experience effects), gender (male = 1), marital status (married = 1), self-reported health (qp201, 1 = best to 5 = worst), and urban–rural residence (urban = 1). All regressions include province fixed effects.

Table 1
Descriptive Statistics

Variable	N	Mean	SD	Min	Max
Annual labour income (CNY)	6,948	46,124.5	42,067.8	1.0	722,050.0
Log labour income	6,948	10.376	0.972	0.000	13.490
Internet adoption	6,948	0.832	0.374	0	1
Online learning (use_learn)	6,948	0.233	0.423	0	1
Online shopping (use_shop)	6,948	0.499	0.500	0	1
Online gaming (use_game)	6,948	0.217	0.412	0	1
Community internet rate (IV)	6,948	0.832	0.184	0.000	1.000
Age (years)	6,948	39.334	10.754	18	60
Male (= 1)	6,948	0.582	0.493	0	1
Years of education	6,948	10.777	4.038	0	22
Health status (1–5, 5 = worst)	6,948	2.722	1.061	1	5
Urban resident (= 1)	6,948	0.574	0.495	0	1
Married (= 1)	6,948	0.775	0.418	0	1

Notes: Annual labour income in Chinese yuan (CNY). Health status (qp201): 1 = very healthy, 5 = very unhealthy; the coefficient in regressions is therefore expected to be negative. Means for use_learn, use_shop, and use_game are unconditional (full sample denominator). Conditional on internet use, the proportions are 28.0%, 59.9%, and 26.1%, respectively.

3.3 Empirical Strategy

Baseline model. Our benchmark is a semi-log OLS regression:

$$ln(Income_i) = \alpha + \beta \cdot Internet_i + \gamma' X_i + \delta_{} + \varepsilon_i \quad (1)$$

where X_i is the vector of controls, $\delta_{}$ denotes province fixed effects, and ε_i is the error term. HC3 heteroskedasticity-robust standard errors are reported in the tables; all inferences are robust to clustering standard errors at the community level (≈ 1,000 clusters)—for the preferred specification the clustered standard error on the internet coefficient (0.032) is virtually identical to its HC3 counterpart (0.030). The coefficient β captures the average income premium associated with internet adoption.

Endogeneity and identification. OLS estimates of β may be biased through two channels: reverse causality (higher earners can more readily afford internet access and devices) and omitted-variable bias (unobserved ability or motivation may simultaneously drive internet adoption and income). We address these concerns via two complementary strategies. IV-2SLS instruments internet adoption with iv_comm (see Section 3.2 for the exclusion restriction defence). PSM (Rosenbaum & Rubin, 1983) matches internet users and non-users on a rich set of observable characteristics using a Logit propensity score and estimates the average treatment effect on the treated (ATT) within the region of common support. Because the community-level instrument is mechanically related to peer behaviour and its exclusion restriction is untestable, we additionally report two further checks that do not rely on it (Section 4.6): a Lewbel (2012) heteroskedasticity-based instrument constructed internally from the model's own controls, and a Heckman (1979) two-step correction for selection into employment, with the number of co-resident children under 16 serving as the selection-equation exclusion restriction.

Panel fixed effects. All of the foregoing estimators are cross-sectional and can remove only selection on observed characteristics. Our strongest defence against selection on time-invariant unobservables—ability, motivation, family background—is a multi-wave panel. Pooling the 2014–2020 waves, we estimate $\ln(\text{Income})_{it} = \alpha + \beta \cdot \text{Internet}_{it} + \gamma' Z_{it} + \mu_i + \tau_t + \varepsilon_{it}$, where μ_i are individual fixed effects, τ_t a wave effect, and Z_{it} the time-varying controls; β is then identified purely from within-individual changes in online status, with age and time-invariant traits absorbed by μ_i and τ_t. We report pooled, fixed-effects, and first-difference estimates side by side (Table 3) and cluster standard errors by individual.

Use-type heterogeneity. To open the 'black box' of internet use, we estimate a classification regression restricted to internet users:

$$\ln(\text{Income}_i) = \alpha + \beta_1 \cdot \text{Learn}_i + \beta_2 \cdot \text{Shop}_i + \beta_3 \cdot \text{Game}_i + \gamma' X_i + \delta_t + \varepsilon_i \quad (2)$$

Comparing β_1, β_2, and β_3 identifies the differential income returns to each use type conditional on other observables. Note that Eq. (2) is identified by OLS; applying IV to use-type indicators would require separate instruments for each purpose variable, which are unavailable in the data. The use-type estimates are therefore best interpreted as conditional correlations rather than strictly causal effects, though the consistency of results across multiple robustness checks (Section 4.5) supports their robustness without establishing causation. To gauge how sensitive these estimates are to selection on unobservables, we additionally report Oster (2019) coefficient-stability bounds (Section 4.6), which quantify how large selection on unobservables would have to be, relative to selection on the included observables, to drive each use-type coefficient to zero.

Heterogeneity analysis. We estimate Eq. (1) separately for subgroups defined by urban/rural status, education (above/below 12 years), gender, and age cohort (16–30, 31–45, 46–60). We similarly estimate Eq. (2) within each subgroup, focusing on the online-learning coefficient. We supplement the subgroup regressions with pooled interaction-term specifications (Internet × Group dummy) to formally test whether differences across subgroups are statistically significant. Because the heterogeneity analysis

involves twelve subgroup tests, we guard against false positives by reporting sharpened false-discovery-rate q-values (Anderson, 2008) alongside conventional p-values.

Mediation analysis. As a complement to the main results, Appendix A reports a mediation exercise on two candidate channels—perceived information access and formal-contract holding—using the Baron–Kenny procedure with bootstrap confidence intervals and the sensitivity parameters of Imai, Keele and Yamamoto (2010). Because the exercise rests on strong no-confounding assumptions, we treat it as suggestive and keep it outside the main analysis.

4. Results

4.1 Baseline OLS Estimates

We begin with cross-sectional benchmarks; readers primarily interested in the paper's central identification contrast—individual fixed effects against these benchmarks—will find it in Section 4.2 and Fig. 1. Table 2 reports baseline OLS results across three nested specifications. Column (1) includes only the internet adoption indicator, yielding a large raw coefficient of 0.468 ($p < 0.01$), which translates into an income difference of approximately 60% [$\exp(0.468) - 1 \approx 59.7\%$]. This estimate primarily reflects compositional differences between users and non-users in education, age, health, and location.

Column (2) adds individual controls; the internet coefficient falls sharply to 0.095—a reduction of approximately 80%—indicating that most of the raw gap is attributable to observable sorting. Column (3) further adds province fixed effects, yielding a cross-sectional OLS coefficient of 0.103 ($p < 0.01$). The implied income premium is $\exp(0.103) - 1 \approx 10.9\%$.

Control variable estimates align with established findings. Education returns are 7.4% per additional year [$\exp(0.071) - 1 \approx 7.4\%$], consistent with Mincerian estimates for China (Li et al., 2012). Age effects exhibit the expected inverted-U pattern. The male wage gap is approximately 66% and urban workers earn about 12.7% more than rural workers, conditional on other characteristics. The negative coefficient on health status (− 0.041) indicates that poorer health (higher score on the 1–5 scale) is associated with lower income, consistent with health-human-capital theory.

Table 2
Baseline OLS Regression Results (Dependent Variable: Log Labour Income)

Variable	(1)	(2)	(3)
Internet adoption	0.468***	0.095***	0.103***
	(0.029)	(0.030)	(0.030)
Age		0.110***	0.107***
		(0.009)	(0.009)
Age²		−0.001***	−0.001***
		(0.000)	(0.000)
Male (= 1)		0.496***	0.505***
		(0.022)	(0.021)
Years of education		0.075***	0.071***
		(0.003)	(0.003)
Health status		−0.032***	−0.041***
		(0.010)	(0.010)
Urban (= 1)		0.198***	0.120***
		(0.023)	(0.023)
Married (= 1)		0.035	0.050*
		(0.031)	(0.030)
Province FE	No	No	Yes
Observations	6,948	6,948	6,948
R^2	0.032	0.212	0.261
Adj. R^2	0.032	0.211	0.257

*Notes: Dependent variable is log annual labour income (ln_income). HC3-robust standard errors in parentheses. Column (3) is the preferred specification. * p < 0.10, ** p < 0.05, *** p < 0.01.*

4.2 Endogeneity, Panel Fixed Effects, and the Selection Interpretation

Table 3 confronts the central identification problem: do people earn more because they go online, or do more capable, higher-earning individuals simply select into going online? Panel A reports cross-sectional estimators on the 2020 wave. A Wu-Hausman test rejects exogeneity of internet adoption (F = 9.03, p = 0.003), and the community-instrument first stage is strong (coefficient on iv_comm 0.242; first-stage F =

116.07, far above the Stock-Yogo (2005) threshold of 10). Read in isolation, these cross-sectional estimators appear mutually corroborating: OLS (0.103), PSM-ATT (0.109), the Lewbel instrument (0.119), and the Heckman-corrected estimate (0.100; Section 4.6) all cluster near a 10–12% premium.

The community IV-2SLS coefficient is far larger (1.037); we treat this as a fragile, complier-specific Local Average Treatment Effect rather than a population parameter (Imbens & Angrist, 1994) and do not rely on it. The PSM-ATT of 0.109, estimated by 1:1 nearest-neighbour matching without replacement (caliper 0.01) within the common-support region [0.047, 0.994], yielding 977 matched pairs with post-matching covariate imbalance below 0.05, closely tracks OLS. Yet every one of these is a cross-sectional estimator: each conditions only on observed characteristics, and none can purge selection on time-invariant unobservables such as ability, motivation, or family background.

Panel B exploits the panel structure of CFPS, and we read its two designs together. The cleaner design is a two-wave panel restricted to 2018–2020, where the after-tax wage measure (emp_income) is defined consistently across waves: pooling reproduces the cross-sectional premium, yet the individual fixed-effects estimate is essentially zero (− 0.002, p = 0.97), as is the first difference. To sharpen the within variation we then stack four waves (2014, 2016, 2018, 2020) using the longest consistently available labour-income measure (incomeb): the panel holds 17,232 individuals and 28,781 person-wave observations, of whom 8,088 are observed in two or more waves and 2,168 (26.8% of those) change online status across waves and identify the within estimator. These switchers are a meaningful middle group rather than an extreme tail: at first observation they average 37.5 years of age and 9.4 years of schooling, between the always-online (31.5 and 12.2) and never-online (46.0 and 6.9) groups, with urbanization and baseline income ordered the same way. The within estimate is therefore informative about the population actually at the adoption margin during this period, though it need not generalize to infra-marginal early adopters. The pooled estimate again matches the cross-section (0.125; +13.3%), but the fixed-effects coefficient is, if anything, slightly negative (− 0.075, p = 0.03) and, on the cleanest switcher-only subsample, statistically indistinguishable from zero (− 0.010, p = 0.75). The negative point estimate itself should not be taken literally—binary-treatment measurement error and the mid-decade redefinition of the CFPS income variable both bias differenced estimates downward, and the consistent-income two-wave panel lands at exactly zero—but the conclusion it supports is clear: the within-individual return is far below the cross-sectional premium and not robustly positive. One threat the design cannot fully dispel is reverse causality at the switching margin—a worker who loses a job may go online to look for one, so that an income change drives adoption rather than the reverse—but this too would push the within estimate downward and so reinforces, rather than overturns, our reading. Across every specification the within-individual estimate is far below the + 11% cross-sectional figure and never positive (Fig. 1). The message—consistent with the Oster bound for the access effect ($\delta^* = 0.83 < 1$; Section 4.6)—is that the aggregate access premium is substantially a selection artifact: simply being online does not robustly raise income once we hold the individual fixed.

This finding sharpens, rather than closes, the central question: if mere connectivity carries little robust causal content, then whatever the internet does for income must operate through how it is used. The

remainder of the paper shows that the robust, policy-relevant variation lies in the composition of use—productive versus consumptive—not in access per se.

Table 3
Identification: Cross-Sectional Estimates versus Panel Fixed Effects

Panel A: Cross-sectional estimators (CFPS 2020; dependent variable = log labour income)			
	OLS	**IV-2SLS**	**PSM-ATT**
Internet adoption	0.103***	1.037***	0.109***
	(0.030)	(0.253)	(0.037)
Diagnostic	—	1st-stage F = 116	977 matched pairs
Observations	6,948	6,948	1,954
Panel B: Four-wave panel fixed effects (CFPS 2014–2020; income = incomeb)			
	Pooled OLS	Individual FE	FE, switchers
Internet adoption	0.125***	−0.075**	−0.010
	(0.015)	(0.035)	(0.033)
Individual FE	No	Yes	Yes
Year FE	Yes	Yes	Yes
Switchers / ≥2-wave persons	—	2,168 / 8,088	2,168 / 2,168
Person-wave obs	28,781	28,781	5,530

Notes: Panel A reports cross-sectional estimators on the 2020 wave; the Lewbel and Heckman estimates referenced in the text appear in Table 7. *Panel B stacks four waves (2014, 2016, 2018, 2020) into an individual panel and re-estimates within individuals; the internet coefficient is identified off the 2,168 respondents who change online status across waves, with age and time-invariant traits absorbed by the individual and year effects. Column 'FE, switchers' restricts the sample to those ever-switchers. The harmonised income measure is incomeb (main-job income); a cleaner two-wave panel (2018–2020) using the consistently defined after-tax wage variable emp_income gives an FE estimate of −0.002 (p = 0.97), confirming the result. In no specification does the within-individual estimate approach the + 11% cross-sectional premium—it is statistically indistinguishable from zero (switcher subsample) or slightly negative—indicating that the cross-sectional premium is substantially selection on time-invariant unobservables. Standard errors clustered by individual (Panel B) or HC3-robust (Panel A) in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.*

4.3 Internet Use Type and Income Returns

Table 4 presents the classification regression results for internet users (n = 5,783). These regressions are identified by OLS and are best read as conditional associations; a causal interpretation would additionally require that, within the user group, unobservables do not differentially select workers into particular use types after conditioning on individual characteristics and province fixed effects.

Online learning carries a coefficient of 0.081 ($p < 0.01$), implying an income premium of approximately 8.4% [exp(0.081) − 1]. Online shopping yields an essentially identical coefficient of 0.081 ($p < 0.01$). Online gaming has a coefficient of − 0.002 and is statistically indistinguishable from zero ($p = 0.937$). In the joint regression (Column 4), coefficients are unchanged: learning 0.076, shopping 0.077, gaming − 0.008. A Wald test for equality of the learning and shopping coefficients cannot reject the null ($F = 0.02$, $p = 0.89$), confirming that these two effects are statistically indistinguishable.

The comparable magnitudes of learning and shopping premiums warrant discussion. We propose two complementary explanations. First, online shopping platforms in China (Taobao, JD, Pinduoduo) are major channels for small-scale entrepreneurship and platform-based self-employment; users who shop online are more likely to also sell online, generating direct income effects. Second, online shopping requires a suite of information-processing tasks—price comparison, seller evaluation, logistics coordination—that may develop transferable cognitive and organisational skills. The zero gaming effect supports both explanations by contrast: gaming involves neither productive information processing in market contexts nor potential entrepreneurial income. The precise zero on gaming doubles as a soft falsification: a generic confounder under which abler individuals do more of everything online and also earn more would inflate the gaming coefficient too, yet gaming is a tight zero, arguing against such a mechanical proxy. CFPS 2020 does not, however, separately record online selling or e-commerce entrepreneurship, so the shopping indicator cannot be partitioned into productive (selling) and purely consumptive components; the ‘mixed-use’ label is therefore inferential, and the comparable shopping and learning premiums are best read as showing that income returns require information-processing engagement rather than entertainment—not that shopping is as productive as formal learning. A direct check reinforces this reading: splitting users by whether they operate a family business (the CFPS self-employment module; 6.8% of users), the shopping coefficient is significant among wage employees (0.063, $p < 0.01$) and larger but imprecise among the self-employed (0.190, $p = 0.10$), with an insignificant interaction ($p = 0.49$). The shopping premium is thus not confined to those positioned to sell online, which weighs against a purely entrepreneurial reading and toward the information-processing interpretation—while leaving room for an entrepreneurship component that the small self-employed sample cannot resolve.

A natural concern is that the same unobserved ability and motivation that inflate the cross-sectional access premium could also sort more capable users into online learning. These estimates are best interpreted as robust conditional associations rather than causal effects, for three reasons. First, the Oster (2019) coefficient-stability bounds (Section 4.6) imply that selection on unobservables would have to be 1.5–3 times as strong as selection on our rich observables to nullify the learning and shopping coefficients—a higher bar than the access effect clears ($\delta^* = 0.83$). Second, selection appears concentrated at a different margin: regressing each treatment on the full set of observables, individual characteristics explain about twice as much of the access decision ($R^2 = 0.23$) as of the within-user choice to learn online ($R^2 = 0.12$), consistent with the extensive-margin decision of whether to go online being far more bound up with ability and resources than the intensive-margin allocation of an already-established connection. Third, a within-person specification is feasible only with the pre-2020 learning

module: over 2014–2018 (a consistent importance measure) it is too imprecise to adjudicate (95% CI ≈ ± 0.18), and extending the window to 2016–2020—which adds the larger 2020 user sample but breaks the measure, since CFPS rebuilt the module as a binary that year—yields a positive but still imprecise estimate (+ 0.036; 95% CI [− 0.05, + 0.13]) that comfortably includes the cross-sectional coefficient yet cannot reject zero. A powered, clean within-person test of use type is therefore beyond what CFPS currently supports, and the conditional-association reading remains the appropriate one.

Table 4
Internet Use Type Regression (Internet Users Only, n = 5,783)

Variable	**(1)** **Learning**	**(2)** **Shopping**	**(3)** **Gaming**	**(4)** **Joint**
Online learning	0.081***			0.076***
	(0.026)			(0.026)
Online shopping		0.081***		0.077***
		(0.024)		(0.024)
Online gaming			−0.002	−0.008
			(0.027)	(0.027)
Control variables	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes
R^2	0.251	0.251	0.250	0.252
Adj. R^2	0.245	0.245	0.244	0.246
Wald: Learn = Shop				F = 0.02, p = 0.89

*Notes: Sample restricted to internet users. Dependent variable is log annual labour income. These specifications are identified by OLS and should be interpreted as conditional correlations. HC3-robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.*

4.4 Heterogeneity Analysis

Table 5 reports subgroup estimates. Two overarching patterns emerge.

First, a directional but largely statistically insignificant Matthew pattern in the cross-sectional access association. The point estimates favour the advantaged—urban residents (0.110) over rural (0.087), highly educated workers (0.227) over less-educated (0.151), and men (0.140) over women (0.054, insignificant)—but, as the interaction tests below show, the urban–rural and gender gaps are not statistically distinguishable, and the aggregate access effect is itself largely selection (Section 4.2). We therefore read these gradients as suggestive cross-sectional patterns rather than causal differences in the return to connectivity.

Second, and more strikingly, the hierarchy reverses for online learning. Rural learners obtain a coefficient of 0.132—approximately 2.2 times the urban coefficient of 0.059. Women obtain 0.125 ($p < 0.01$) from online learning versus 0.044 (insignificant) for men. Workers aged 46–60 exhibit the largest coefficient of any cohort (0.233, $p < 0.01$), compared to 0.077 for workers aged 31–45 and − 0.019 (insignificant) for young workers aged 16–30. Because the subgroup regressions for the smallest cells (e.g. the 1,517 older users) are imprecise, we treat the full-sample interaction terms as the primary evidence for the reversal. These pooled specifications, interacting treatment with each group dummy, show that the reversal is statistically significant for the gender and age dimensions—the Learning × Female interaction is + 0.205 (SE = 0.048) ($p < 0.01$) and the Learning × Older(46–60) interaction is + 0.173 (SE = 0.056, $p < 0.01$)—whereas the urban–rural learning gap, though pointing in the same direction, is not statistically distinguishable (Learning × Rural interaction, $p = 0.69$). Symmetrically, the Matthew-effect gradients in access returns (Internet × Urban, Internet × Male) are directionally present but not individually significant (interaction $p = 0.22$), so the aggregate-access hierarchy is in fact weaker, not stronger, than the use-quality reversal. The subgroup online-learning premiums for rural, female, and older workers are each individually significant and survive sharpened false-discovery-rate correction ($q < 0.05$; Anderson, 2008).

Notably, the reversal pattern applies to the urban–rural, gender, and age dimensions but not to education. Highly educated workers obtain a larger online-learning coefficient (0.170) than less-educated workers (0.046, insignificant), so the second-level digital divide along educational lines remains in the conventional direction. This asymmetry suggests that education functions differently from other disadvantage dimensions: while rural, female, and older workers face information-access barriers that productive internet use can offset, low-educated workers also lack the prerequisite literacy and cognitive skills to fully extract value from online learning content.

Several mechanisms explain the older-worker finding. Older workers face higher barriers to conventional skill upgrading—fewer employer-sponsored training opportunities, age discrimination in formal education, and higher opportunity costs of time-intensive retraining—making internet-based learning a higher-value marginal substitute. Additionally, older workers typically possess deeper domain-specific human capital that may be more complementary to the targeted, applied knowledge available online. We note that only 65% of workers aged 46–60 use the internet (versus 97% of those aged 16–30), suggesting that the online-learning coefficient for this cohort reflects returns among those who have already overcome the adoption barrier—likely a more motivated and self-selected group.

The reversal pattern implies that, while the first-level digital divide (access) advantages privileged groups, the second-level divide (use quality) operates differently. Precisely because disadvantaged groups historically face greater information market failures—rural information poverty, gender discrimination in training access, age-based exclusion from firm-sponsored training—productive internet use yields higher marginal returns for them. The internet, when used productively, may thus function as a partial equalizing technology, contingent on overcoming the access barrier.

Table 5
Heterogeneity Analysis

Panel A: Urban–Rural			
	Urban	**Rural**	
Internet adoption	0.110***	0.087**	
	(0.041)	(0.044)	
Online learning	0.059*	0.132***	
	(0.030)	(0.049)	
N (full / users)	3,986 / 3,475	2,962 / 2,308	
Panel B: Education (threshold: 12 years)			
	High (≥ 12 yrs)	Low (< 12 yrs)	
Internet adoption	0.227***	0.151***	
	(0.059)	(0.035)	
Online learning	0.170***	0.046	
	(0.031)	(0.049)	
N (full / users)	3,385 / 3,231	3,563 / 2,552	
Panel C: Gender			
	Male	Female	
Internet adoption	0.140***	0.054	
	(0.035)	(0.055)	
Online learning	0.044	0.125***	
	(0.034)	(0.040)	
N (full / users)	4,042 / 3,324	2,906 / 2,459	
Panel D: Age Cohort			
	Young (16–30)	Middle (31–45)	Older (46–60)

Notes: Each subgroup is estimated independently with the same controls as Table 2 *(excluding the grouping variable). Internet user rate (Panel D) is the fraction of each age cohort that reports any internet use. Pooled interaction estimates with community-clustered standard errors: Learning × Female = 0.205 (0.048) and Learning × Older(46–60) = 0.173 (0.056), both p < 0.01; Learning × Rural = − 0.021 (0.053), p = 0.69; Internet × Urban = 0.070 (0.057) and Internet × Male = − 0.071 (0.058), both p = 0.22. Subgroup learning premiums for rural, female, and older workers survive sharpened FDR correction with q = 0.011, 0.004, and below 0.001, respectively (*Anderson, 2008*). For the subgroup estimates in the table, HC3-robust standard errors in parentheses; inferences are robust to community-clustered standard errors. * p < 0.10, ** p < 0.05, *** p < 0.01.*

Panel A: Urban–Rural			
	Urban	**Rural**	
Internet adoption	0.117	0.029	0.138***
	(0.111)	(0.052)	(0.040)
Online learning	−0.019	0.077**	0.233***
	(0.050)	(0.035)	(0.054)
Internet user rate	97%	83%	65%
N (full / users)	1,727 / 1,673	2,886 / 2,593	2,335 / 1,517

Notes: Each subgroup is estimated independently with the same controls as Table 2 *(excluding the grouping variable). Internet user rate (Panel D) is the fraction of each age cohort that reports any internet use. Pooled interaction estimates with community-clustered standard errors: Learning × Female = 0.205 (0.048) and Learning × Older(46–60) = 0.173 (0.056), both p < 0.01; Learning × Rural = − 0.021 (0.053), p = 0.69; Internet × Urban = 0.070 (0.057) and Internet × Male = − 0.071 (0.058), both p = 0.22. Subgroup learning premiums for rural, female, and older workers survive sharpened FDR correction with q = 0.011, 0.004, and below 0.001, respectively (*Anderson, 2008*). For the subgroup estimates in the table, HC3-robust standard errors in parentheses; inferences are robust to community-clustered standard errors. * p < 0.10, ** p < 0.05, *** p < 0.01.*

4.5 Robustness Checks

Table 6 presents robustness checks across three dimensions.

Alternative income measure. Replacing emp_income with the CFPS primary-job income variable (incomeb) yields virtually identical coefficients: internet adoption 0.086, online learning 0.081, online shopping 0.086 (all p < 0.01). Results are stable across both income definitions.

Extreme-value treatment. Trimming the top and bottom 1% of the income distribution gives internet adoption 0.118 and online learning 0.094. Winsorising at 1% yields 0.106 and 0.078. Neither modification materially changes the inference, confirming that outliers do not drive the results.

Device-type decomposition. Splitting access by device (mobile-only 0.098; PC-only 0.262; jointly 0.079 and 0.258) shows the expected descriptive gradient, but device ownership is itself selected, so these rows of Table 6 are reported for completeness and play no role in our inference.

Table 6
Robustness Checks

Specification	Internet adoption	Online learning	Online shopping	N
Baseline (OLS)	0.103***	0.081***	0.081***	6,948
	(0.030)	(0.026)	(0.024)	
Alternative income (incomeb)	0.086***	0.081***	0.086***	7,044
	(0.031)	(0.026)	(0.024)	
Drop top/bottom 1%	0.118***	0.094***	0.071***	6,820
	(0.027)	(0.023)	(0.021)	
Winsorise at 1%	0.106***	0.078***	0.080***	6,948
	(0.029)	(0.025)	(0.022)	
Mobile internet only	0.098***	—	—	6,948
	(0.030)			
PC internet only	0.262***	—	—	6,948
	(0.026)			
Mobile + PC (joint)	Mob: 0.079***	—	—	6,948
	(0.030)			
	PC: 0.258***			
	(0.026)			

*Notes: All specifications include the full control vector and province FE. — indicates the coefficient was not separately estimated for this specification. The alternative-income (incomeb) row uses 7,044 observations rather than 6,948 because the primary-job income variable has a slightly different missingness pattern from emp_income. The device-type specifications are observational (OLS) and should be interpreted as descriptive evidence. HC3-robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.*

4.6 Identification Robustness: Alternative Instruments, Selection, and Oster Bounds

The community-level instrument, despite its strong first stage, may partly capture community-level shocks rather than a pure peer effect, and its exclusion restriction is untestable. We first probe the instrument directly and then subject the internet-adoption effect to three further checks that do not rely on it; Table 7 collects the results.

Falsification and local-conditions checks. If community penetration affects income only through own adoption, it should not predict the income of respondents who remain offline: among the 1,165 never-online respondents, the coefficient on iv_comm is small and statistically indistinguishable from zero

(0.080, SE = 0.138, p = 0.56)—supportive of the exclusion restriction, though the selected nature of this subsample makes the test suggestive rather than decisive. More telling is a local-conditions control: augmenting the 2SLS with the leave-one-out community mean of log income—a direct proxy for local economic development—cuts the coefficient nearly in half, from 1.037 to 0.548 (p = 0.026), while the first stage remains strong. The instrument thus partly reflects local economic conditions; accordingly, the community IV-2SLS serves as supplementary evidence, with identification resting on the panel design.

Lewbel (2012) instrument. Exploiting heteroskedasticity in the first-stage residuals, we construct instruments internally from the included controls. A Breusch–Pagan test strongly rejects homoskedasticity (F = 418, p < 0.001), confirming the approach is applicable. The resulting estimate of 0.119 (p = 0.01; ≈12.7%) is close to the OLS and PSM figures and an order of magnitude below the community-IV LATE, reinforcing our reading of the latter as a complier-specific local effect rather than the average return.

Selection into employment. Because the wage sample conditions on positive labour income, we estimate a Heckman (1979) two-step model on the full working-age sample (n = 16,190; participation rate 51.3%), using the number of co-resident children under 16 as the selection-equation exclusion restriction—it strongly predicts participation (p < 0.001) but is plausibly excludable from the own-wage equation. The wage equation is estimated on the employed subsample, which is slightly larger than the 6,948-person main analysis sample because the selection model does not require the community identifier used to construct the instrument. The inverse Mills ratio is marginally significant (p = 0.07), indicating modest positive selection; nonetheless the selection-corrected internet coefficient (0.100; ≈10.6%) is essentially unchanged from the baseline, so employment selection does not drive our result.

Oster (2019) bounds for use type. For the use-type coefficients—identified by OLS and thus most exposed to omitted-variable bias—we compute coefficient-stability bounds. Using Oster's recommended benchmark ($R_{max} = 1.3\tilde{R}$), the selection on unobservables relative to observables that would be required to nullify the effect is $\delta^* = 1.47$ for online learning and $\delta^* = 3.23$ for online shopping; both exceed the conventional robustness threshold of 1, and the bias-adjusted coefficients remain positive (0.026 and 0.056). By the same metric the binary-access effect is far less robust ($\delta^* = 0.83 < 1$), in line with the panel fixed-effects result (Section 4.2) that this premium is largely selection and with our broader argument that use quality is more sharply identified than mere access and should carry the weight of the contribution.

Table 7
Additional Identification Checks

Panel A: Internet-adoption effect under alternative identification				
	Estimate	**Std. err.**	**Implied %**	**Diagnostic**
OLS (baseline)	0.103***	(0.030)	+ 10.9%	—
PSM (ATT)	0.109***	(0.037)	+ 11.6%	977 pairs
Lewbel (2012) IV	0.119**	(0.046)	+ 12.7%	BP F = 418
Heckman selection	0.100***	(0.031)	+ 10.6%	λ p = 0.07
Community IV-2SLS (LATE)	1.037***	(0.253)	(complier)	1st-stage F = 116
Panel B: Oster (2019**) bounds for use-type effects (R_max = 1.3·R̃)**				
	β (controlled)	δ*	β* (adj.)	Robust?
Online learning	0.081	1.47	0.026	Yes (δ* > 1)
Online shopping	0.081	3.23	0.056	Yes (δ* > 1)
Internet adoption	0.103	0.83	−0.022	Relies on IV/PSM

Notes: Panel A reports the internet-adoption coefficient under five identification strategies; all but the community IV-2SLS LATE cluster tightly in the 0.10–0.12 range. BP F is the Breusch–Pagan heteroskedasticity statistic underpinning the Lewbel instrument; λ is the inverse Mills ratio from the Heckman model. Panel B reports Oster (2019*) statistics for the OLS-identified use-type effects: δ* is the relative degree of selection on unobservables (vs. observables) that would drive the coefficient to zero, and β* is the bias-adjusted coefficient, both at R_max = 1.3·R̃. δ* > 1 indicates robustness to plausible omitted-variable bias. * p < 0.10, ** p < 0.05, *** p < 0.01.*

5. Discussion and Conclusion

This paper examined how the purpose of internet use shapes individual labour income returns in China, using CFPS 2014–2020 data and a research design that pairs cross-sectional estimators with a multi-wave (2014–2020) individual fixed-effects panel. Three principal findings emerge. First, the aggregate internet-adoption premium—near 11% in every cross-sectional estimator (OLS, PSM, Lewbel, Heckman)—is largely a selection artifact: it collapses to essentially zero under individual fixed effects, so merely being online does not robustly raise income once who goes online is held fixed. Second, this association masks sharp use-type heterogeneity: conditional on being online, productive (online learning) and mixed (online shopping) uses are each associated with income premiums of approximately 8.4–8.5%, while consumptive use (online gaming) shows none—robust conditional correlations, given the limits of the data, rather than causal effects. Third, while the cross-sectional returns to simple internet access weakly favour the advantaged (urban, highly educated, male)—a Matthew pattern that is directional but, for the urban–rural and gender margins, not statistically significant—the returns to online learning specifically exhibit a striking reversal: rural residents, women, and older workers (46–60) each obtain substantially

higher marginal returns. Formal pooled interaction tests confirm this reversal for the gender and age dimensions ($p < 0.01$) and show the same directional pattern—though not a statistically significant difference—across the urban–rural divide; the subgroup learning premiums for rural, female, and older workers each survive false-discovery-rate correction.

These findings carry important theoretical implications. They validate the second-level digital divide framework: the income consequences of internet connectivity depend fundamentally on use quality, not access per se. The zero gaming effect aligns with human capital theory—leisure consumption does not raise productivity. The comparable shopping and learning premiums suggest that income returns require, at minimum, information-processing engagement with the internet, regardless of whether that engagement takes a traditional educational form. The reversal in learning returns suggests a compensating mechanism: disadvantaged groups face higher barriers to conventional skill acquisition channels, making online learning a higher-value marginal substitute. The mediation analysis (Appendix A) points to two plausible channels—information acquisition and, more robustly, employment formalisation—consistent with income operating through both skill accumulation and labour-market matching, though both channels are sensitive to unobserved confounding and are read as suggestive.

Several policy implications follow. First, digital policy should shift emphasis from broadband rollout (access) to digital literacy (use quality). Our results suggest that investment in online learning content, digital skills training, and labour-market-relevant online platforms would yield high economic dividends, particularly in rural areas and for older workers. The finding that online learning returns are highest precisely among those currently least likely to go online—older and rural workers—implies large potential gains from complementary adoption-promotion policies. Second, the employment formalisation channel highlights the complementary role of labour market institutions: policies strengthening formal employment and enforcing labour contracts would amplify the income gains from productive internet use. Third, the heterogeneity reversal we document suggests that targeted interventions promoting productive internet use among disadvantaged groups could actively reduce income inequality, countering the Matthew pattern observed at the access level.

Several limitations warrant acknowledgment. Our four-wave (2014–2020) panel removes time-invariant selection and rests on 2,168 status-switchers, but binary-treatment measurement error and a mid-decade change in the CFPS income variable both bias differenced estimates downward, so the slightly negative within point estimate is best read as far below the cross-sectional premium and not robustly positive, rather than as a true penalty; the 2010 and 2012 waves cannot be added because they lack, respectively, a comparable labour-income variable and the internet module. Our headline use-type results are cross-sectional: the binary 2020 indicators (Table 4) cannot be differenced within individuals because CFPS used a different use-type format—frequency and importance—in the surrounding waves. We exploit that earlier format for a within-person robustness check (Section 4.3), but it is underpowered; extending it to later waves that retain the same module—including the recently released 2022 wave—is the natural next step toward a powered within-person test of use type. The exclusion restriction of the community-level IV is plausible but untestable, although the Lewbel instrument and Heckman correction

(Section 4.6) provide convergent estimates that do not depend on it. The use-type regressions (Table 4) are identified by OLS; separate instruments for each use type would be needed for strict causal identification, though the Oster (2019) bounds indicate the learning and shopping effects are robust to plausible selection on unobservables. The mediation analysis relies on the no-unmeasured-confounding assumption for each mediator-outcome relationship. The CFPS use-type categories (learning, shopping, gaming) are broad; finer-grained measures—specific platforms, time spent, skill assessments—would allow richer characterisation.

In summary, this paper contributes to the literature on digital technology and labour market inequality by demonstrating that what matters for income returns is not whether one is online but how that connectivity is used. Expanding access is necessary but not sufficient; ensuring that broadened access translates into productive use—particularly for disadvantaged populations—is, on this evidence, the more promising policy margin for inclusive digital dividends.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study analyses de-identified secondary data from the China Family Panel Studies (CFPS), collected by the Institute of Social Science Survey at Peking University under its own ethical approval and with informed consent obtained from all respondents by the data provider. No new human-participant data were collected and no additional ethical approval was required.

Informed consent

Not applicable: the study uses de-identified secondary survey data; informed consent was obtained from participants by the original data collector.

Author Contribution

Q.X.: conceptualisation, methodology, formal analysis, data curation, writing—original draft, review and editing. D.Q.: methodology, validation, writing—review and editing.

Acknowledgements

The authors thank the Institute of Social Science Survey, Peking University, for making the China Family Panel Studies data available. This research received no specific grant from any funding agency in the

public, commercial or not-for-profit sectors.

Data Availability

The raw China Family Panel Studies (CFPS) micro-data analysed in this study (2014, 2016, 2018 and 2020 waves) are owned by the Institute of Social Science Survey, Peking University, and are freely available to registered users from the Peking University Open Research Data platform (https://www.isss.pku.edu.cn/cfps/). Under the CFPS data-use agreement the raw files may not be redistributed by third parties and are therefore not included here; the Supplementary script build_analysis_sample.py reconstructs the analysis sample from them in a single step. All other materials underpinning the analyses are provided as Supplementary Material in the archive Supplementary_Data_and_Code.zip: (i) the processed analysis sample used throughout the paper (cfps2020_analysis_sample.dta and cfps2020_analysis_sample.csv; n = 6,948); (ii) a codebook giving the definition, source variable, coding and transformation of every variable (codebook.txt); (iii) the Stata do-file and Python scripts that reproduce every table, figure and robustness check (replication.do and the accompanying .py files), together with their saved outputs; and (iv) a README documenting the reproduction steps. The same archive has also been deposited in the Figshare repository under the reserved DOI https://doi.org/10.6084/m9.figshare.33314916; the record will be made publicly available upon acceptance, and a private access link has been provided to the editorial office for the purposes of peer review.

References

1. Aker JC (2010) Information from markets near and far: Mobile phones and agricultural markets in Niger. Am Economic Journal: Appl Econ 2(3):46–59
2. Anderson ML (2008) Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. J Am Stat Assoc 103(484):1481–1495
3. Autor DH (2001) Wiring the labor market. J Economic Perspect 15(1):25–40
4. Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J Personal Soc Psychol 51(6):1173–1182
5. Becker GS (1965) A theory of the allocation of time. Econ J 75(299):493–517
6. Belo R, Ferreira P, Telang R (2014) Broadband in school: Impact on student performance. Manage Sci 60(2):265–282
7. CNNIC (2025) The 55th Statistical Report on China’s Internet Development. China Internet Network Information Center, Beijing
8. Deng X, Guo M, Liu Y (2023) Digital economy development and the urban–rural income gap: Evidence from Chinese cities. PLoS ONE 18(2):e0280225.

https://doi.org/10.1371/journal.pone.0280225

9. DiNardo JE, Pischke JS (1997) The returns to computer use revisited: Have pencils changed the wage structure too? Quart J Econ 112(1):291–303
10. Fairlie RW (2004) Race and the digital divide. Contrib Economic Anal Policy 3(1):Article 15
11. Galperin H, Fernández-Ardèvol M (2019) Digital inequality and mobile phones in rural Guatemala: A field experiment. Inf Technol Int Dev 15:1–16
12. Hargittai E (2002) Second-level digital divide: Differences in people's online skills. First Monday 7(4). https://doi.org/10.5210/fm.v7i4.942
13. Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47(1):153–161
14. Imai K, Keele L, Yamamoto T (2010) Identification, inference and sensitivity analysis for causal mediation effects. Stat Sci 25(1):51–71. https://doi.org/10.1214/10-STS321
15. Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62(2):467–475
16. Kuhn P, Mansour H (2014) Is internet job search still ineffective? Econ J 124(581):1213–1233
17. Lewbel A (2012) Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. J Bus Economic Stat 30(1):67–80
18. Li H, Liu PW, Ma N, Zhang J (2012) Does education pay in urban China? Estimating returns to education using twins. Economica 79(314):383–412
19. Ma W, Nie P, Zhang P, Renwick A (2020) Impact of internet use on economic well-being of rural households: Evidence from China. Rev Dev Econ 24(2):503–523. https://doi.org/10.1111/rode.12645
20. Mincer J (1974) Schooling, Experience, and Earnings. National Bureau of Economic Research / Columbia University
21. Oster E (2019) Unobservable selection and coefficient stability: Theory and evidence. J Bus Economic Stat 37(2):187–204
22. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 40(3):879–891
23. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
24. Selwyn N (2004) Reconsidering political and popular understandings of the digital divide. New Media Soc 6(3):341–362
25. Si X, Li M (2023) Impact of internet use on informal workers' wages: Evidence from China. PLoS ONE 18(5):e0285973. https://doi.org/10.1371/journal.pone.0285973
26. Stock JH, Yogo M (2005) Testing for weak instruments in linear IV regression. In: Andrews DWK, Stock JH (eds) Identification and Inference for Econometric Models. Cambridge University Press, pp 80–108

27. van Dijk JAGM (2006) Digital divide research, achievements and shortcomings. Poetics 34(4–5):221–235
28. Warschauer M (2003) Technology and Social Inclusion: Rethinking the Digital Divide. MIT Press
29. Weng J, Li W, Zheng X (2024) Internet use, income mobility and the expansion of the rural middle-income group: Evidence from China. Heliyon 10(16):e36070. https://doi.org/10.1016/j.heliyon.2024.e36070
30. Yuan K, Zhang X (2024) The impact of internet usage preferences on labor income: Evidence from China. PLoS ONE 19(8):e0308287. https://doi.org/10.1371/journal.pone.0308287
31. Zhan Y, Yang S (2024) Does internet use improve employment? Empirical evidence from China. PLoS ONE 19(4):e0301465. https://doi.org/10.1371/journal.pone.0301465
32. Zhang A, Chandio AA, Yang T, Ding Z, Liu Y (2023) Examining how internet use and non-farm employment affect rural households' income gap: Evidence from China. Front Sustainable Food Syst 7:1173158. https://doi.org/10.3389/fsufs.2023.1173158
33. Endnotes 1 CFPS 2022 had been released at the time of writing. Its internet-use module returns to the frequency-and-importance format used in 2014–2018 rather than the binary indicators introduced in 2020, so it is not comparable with the 2020 use-type measures underlying our cross-sectional decomposition and cannot extend that cross-section. Because it shares the earlier format, the 2022 wave could instead extend the within-person use-type check of Section 4.3, which we leave to future work

Endnotes

1. CFPS 2022 had been released at the time of writing. Its internet-use module returns to the frequency-and-importance format used in 2014–2018 rather than the binary indicators introduced in 2020, so it is not comparable with the 2020 use-type measures underlying our cross-sectional decomposition and cannot extend that cross-section. Because it shares the earlier format, the 2022 wave could instead extend the within-person use-type check of Section 4.3, which we leave to future work.

Figures

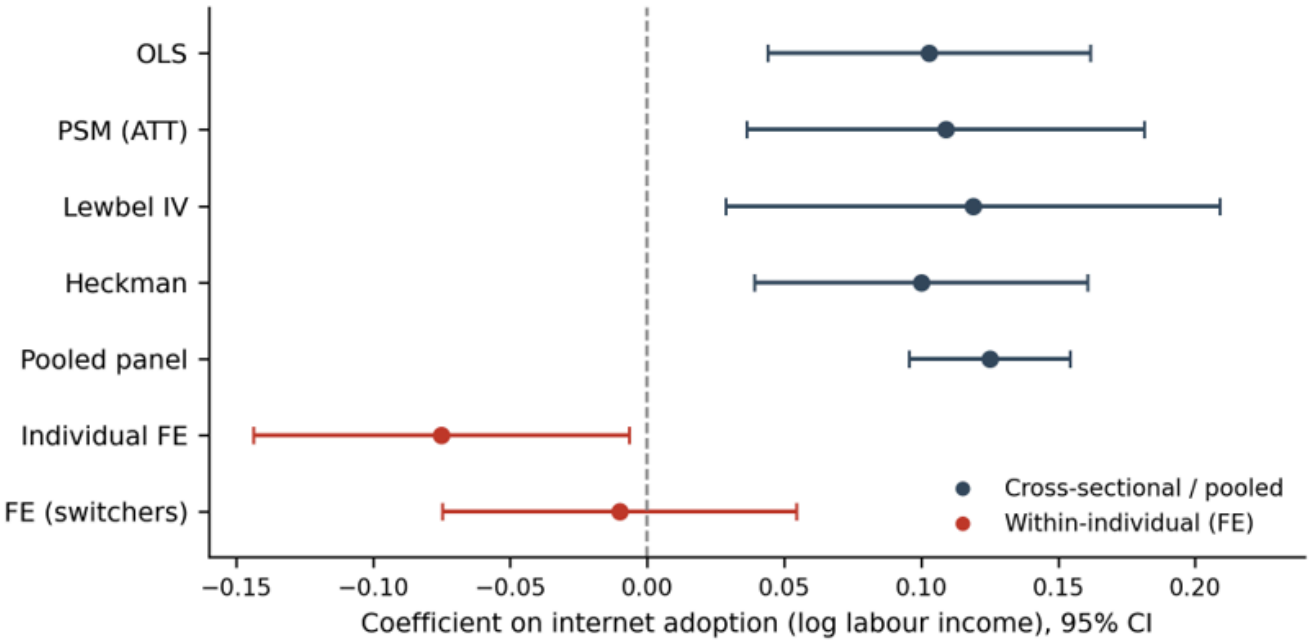


Figure 1

The internet–income coefficient across identification strategies. Point estimates of the internet-adoption coefficient on log labour income, with 95% confidence intervals. The five cross-sectional or pooled estimators (OLS, PSM, Lewbel IV, Heckman, and the pooled four-wave panel) cluster near +0.10 to +0.13; the two within-individual estimators (individual fixed effects, full sample and switcher subsample) are statistically indistinguishable from zero. The community IV-2SLS LATE (1.037) is omitted for scale. The contrast is the paper's central identification result: the aggregate access premium is largely selection.

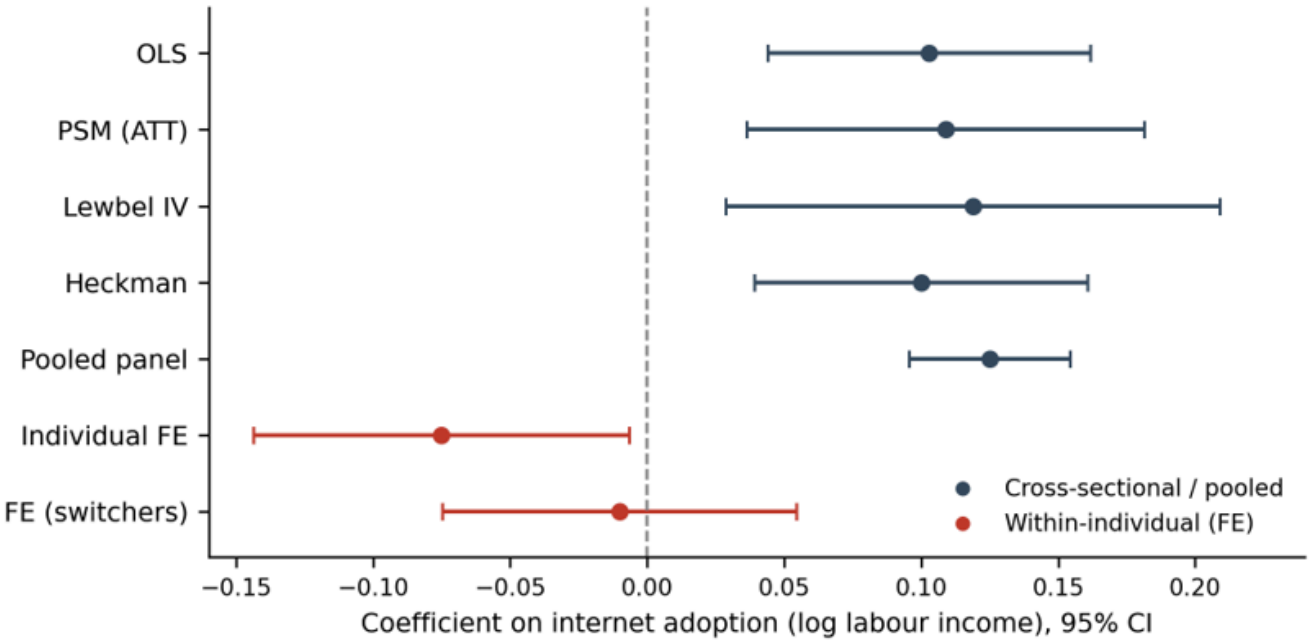


Figure 1

The internet–income coefficient across identification strategies. Point estimates of the internet-adoption coefficient on log labour income, with 95% confidence intervals. The five cross-sectional or pooled estimators (OLS, PSM, Lewbel IV, Heckman, and the pooled four-wave panel) cluster near +0.10 to +0.13; the two within-individual estimators (individual fixed effects, full sample and switcher subsample) are statistically indistinguishable from zero. The community IV-2SLS LATE (1.037) is omitted for scale. The contrast is the paper's central identification result: the aggregate access premium is largely selection.

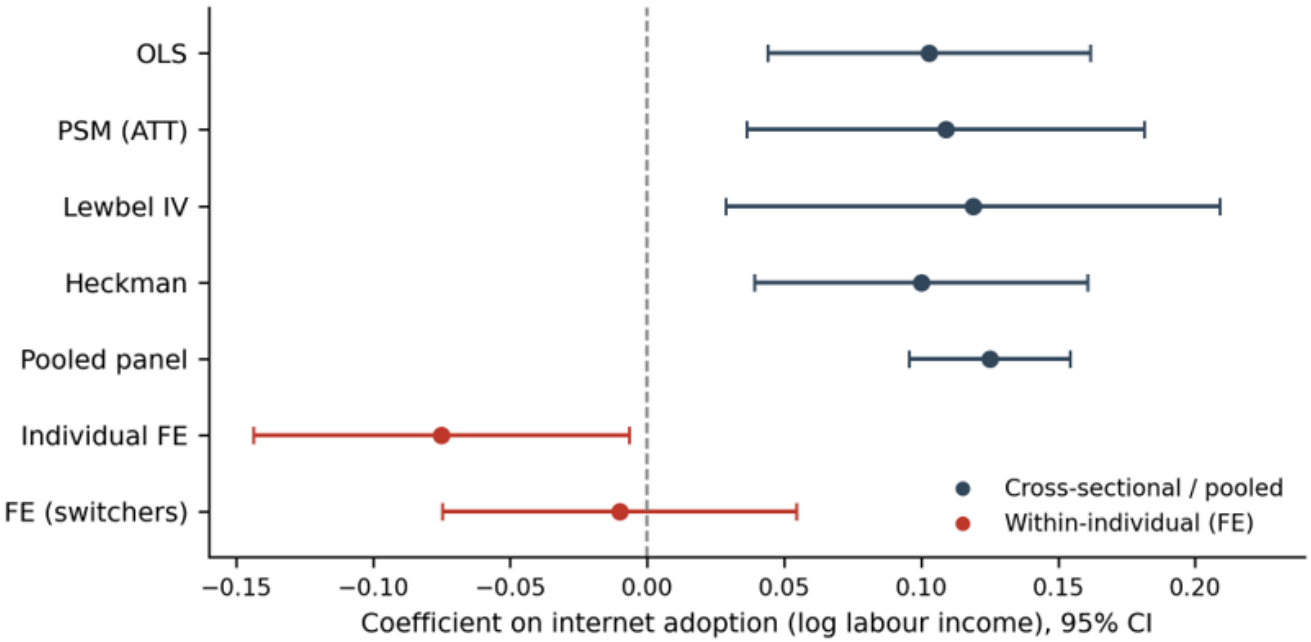


Figure 1

The internet–income coefficient across identification strategies. Point estimates of the internet-adoption coefficient on log labour income, with 95% confidence intervals. The five cross-sectional or pooled estimators (OLS, PSM, Lewbel IV, Heckman, and the pooled four-wave panel) cluster near +0.10 to +0.13; the two within-individual estimators (individual fixed effects, full sample and switcher subsample) are statistically indistinguishable from zero. The community IV-2SLS LATE (1.037) is omitted for scale. The contrast is the paper's central identification result: the aggregate access premium is largely selection.

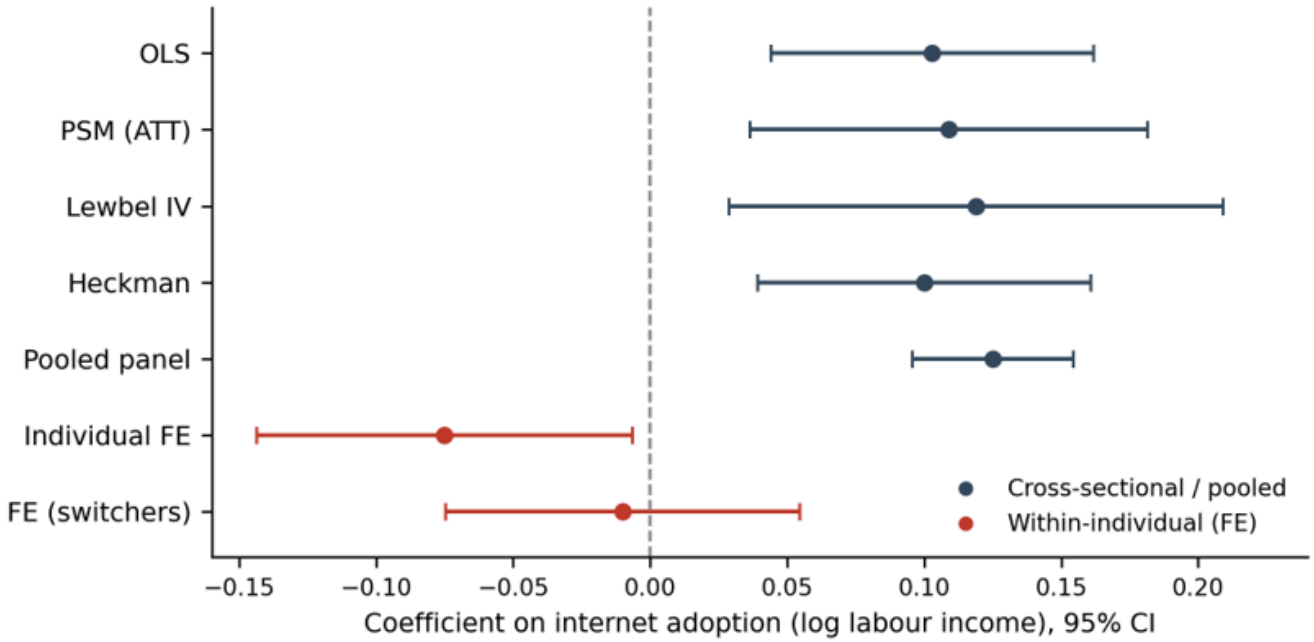


Figure 1

The internet–income coefficient across identification strategies. Point estimates of the internet-adoption coefficient on log labour income, with 95% confidence intervals. The five cross-sectional or pooled estimators (OLS, PSM, Lewbel IV, Heckman, and the pooled four-wave panel) cluster near +0.10 to +0.13; the two within-individual estimators (individual fixed effects, full sample and switcher subsample) are statistically indistinguishable from zero. The community IV-2SLS LATE (1.037) is omitted for scale. The contrast is the paper's central identification result: the aggregate access premium is largely selection.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationHSSC.docx
- SupplementaryInformationHSSC.docx
- SupplementaryInformationHSSC.docx
- SupplementaryInformationHSSC.docx
- SupplementaryDataandCode.zip
- SupplementaryDataandCode.zip
- SupplementaryDataandCode.zip
- SupplementaryDataandCode.zip