NBER WORKING PAPER SERIES

THE NEW DIGITAL DIVIDE

Mayana Pereira
Shane Greenstein
Raffaella Sadun
Prasanna Tambe
Lucia Ronchi Darre
Tammy Glazer
Allen Kim
Rahul Dodhia
Juan Lavista Ferres

Working Paper 32932
http://www.nber.org/papers/w32932

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2024

The first four authors contributed equally to this work. Raffaella Sadun appreciates funding for her research from Harvard University, Microsoft Corporation, NBER, NSF, Sloan Foundation, and Russell Sage Foundation. We thank Dafna Bearson for her helpful comments. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The New Digital Divide
Mayana Pereira, Shane Greenstein, Raffaella Sadun, Prasanna Tambe, Lucia Ronchi Darre, Tammy Glazer, Allen Kim, Rahul Dodhia, and Juan Lavista Ferres
NBER Working Paper No. 32932
September 2024
JEL No. C43, L63, L86

ABSTRACT

We build and analyze new metrics of digital usage that leverage telemetry data collected by Microsoft during operating system updates across forty million Windows devices in U.S. households. These measures of US household digital usage are much more comprehensive than those made available through any existing commercial or government survey. We construct representations of devices in ZIP codes and find evidence of significant variation in usage reflecting an urban-rural divide. We also show the existence of substantial disparities in usage even within narrowly defined Metropolitan Statistical Areas. Income and education correlate with these observed differences. These effects are large and suggest digital literacy gaps that extend beyond the availability of essential IT infrastructure at the local level. These findings call for interventions beyond the traditional focus on infrastructure access and address usage and skills development. The indices are made publicly available to support future research.

Mayana Pereira
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
mayana.wanderley@microsoft.com

Shane Greenstein
Technology Operation and Management
Morgan Hall 439
Harvard Business School
Soldiers Field
Boston, MA 02163
and NBER
sgreenstein@hbs.edu

Raffaella Sadun
Harvard Business School
Morgan Hall 215
Soldiers Field
Boston, MA 02163
and NBER
rsadun@hbs.edu

Prasanna Tambe
Wharton School
University of Pennsylvania
3730 Walnut Street
558 Jon M. Huntsman Hall
Philadelphia, PA 19104
tambe@wharton.upenn.edu

Lucia Ronchi Darre
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
lronchidarre@microsoft.com

Tammy Glazer
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
tammy.glazer@microsoft.com

Allen Kim
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
Allen.Kim@microsoft.com

Rahul Dodhia
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
rahul.dodhia@microsoft.com

Juan Lavista Ferres
Microsoft Corportation
One Microsoft Way
Redmond, WA 98052
jlavista@microsoft.com

Introduction

Digital literacy encompasses a broad range of skills and knowledge related to using digital technologies and information in various contexts, including but not limited to computers, smartphones, and the Internet. Digital literacy plays a fundamental role in several social outcomes, including enabling economic opportunity, social inclusion, and participation in civic activities (*1–4*). Nonetheless, digital literacy may not be equitably distributed, implying that groups or communities with lower endowments of digital skills may not be able to fully reap the benefits of new technologies.

While closing this gap has been elevated to a major policy goal both in the U.S. and internationally (*5*), the measurement of digital literacy still largely relies on crude indicators of technology adoption, such as computing ownership and broadband availability, which suffer from limited samples and insufficient measurement of the depth and variety of usage.[1] Designing interventions to close the digital divide requires a more detailed and comprehensive understanding of persisting differences in how individuals use digital technologies and the underlying factors driving these differences (*7*).

Our research provides novel data and methodology to measure a proxy of digital literacy–digital usage–for over 28,000 US ZIP codes. These metrics are based on telemetry data collected by Microsoft in 2023 during operating system updates for forty million Windows devices across U.S. households that agreed to share these data, using frontier differential privacy practices.[2] This dataset, the most comprehensive measure of US household computing usage ever assembled, surpasses any existing commercial or government survey in its scope and enables granular measurement of the use of digital applications at a massive scale, which, as we show below, captures a different dimension of digital access than traditional metrics based on computing and network infrastructure.

We construct two indices reflecting distinct digital usage components: the *Media and Information Composite Index* (MCI) and the *Content Creation and Computation Composite Index* (CCI).

[1] Existing comprehensive surveys are designed to offer yardsticks that characterize adoption across the US population and explore only the most essential features of usage, such as whether a household uses email (*6*). Similarly, the Pew Research Center (2023) has collected statistics about samples of broadband and smartphone adopters for many years, and it occasionally conduct surveys of usage by adopters in one-off surveys, but rarely the same in-depth survey twice.

[2] Our data assumes that device usage is a proxy for households laptop and desktop usage. The devices are not part of a business contract.

The MCI captures usage related to media and information consumption and general computing usage across various applications, such as word processing, spreadsheets, and presentations. The CCI captures usage about content creation and specialized digital applications, like image manipulation (e.g., Photoshop) or software developer tools.

We examine variance in digital usage across and within U.S. ZIP codes and counties and provide descriptive evidence on correlates of variation in digital literacy beyond local IT infrastructure availability, focusing on local income and education levels. We find evidence of significant variation in digital usage across and within U.S. regions. These differences document an urban-rural divide, but we also show the existence of substantial disparities in usage in both indices, even *within* narrowly defined Metropolitan Statistical Areas. Finally, we show that these differences are correlated with differences in income and education, even once we consider local differences in the availability of essential IT infrastructure.

Our paper brings a fresh perspective to the field of digital literacy. The research most closely related to our goals has concentrated on variance in the skills and conceptual understanding required to navigate the internet successfully (*1, 8, 9*). No research has followed related principles in characterizing computing usage with this large a population. To address literacy gaps, the findings call for long-term, comprehensive strategies that go beyond the traditional focus on infrastructure access and simultaneously address access, usage, and skills development.

Previous works The measurement of digital literacy has traditionally relied on measures of hardware adoption, such as household use of computing, which has steadily grown since the introduction of the personal computer in the 1970s. The first census of computing ownership in 1984 found approximately 10% of US households had a laptop at home. The most recent censuses find that ownership of computing is widespread today.[3] However, widespread access to computing does not imply that all digital divide issues have been resolved. Many digital literacy studies of online navigation have concluded access to computing does not guarantee its use to its full potential or similar types of use. Users may need more knowledge, skills, and experience to use it effectively

[3] US Census of Computer and Internet Use in 2018 indicates that 92% of households had at least one computing device, including a smartphone. Despite various devices, 78% of households owned a laptop or desktop, with 85% having broadband internet subscriptions (*6*). A 2021 survey estimated it at 92% (Pew, 2023).

and for different purposes (*1–4*).

This study contributes to the literature on the digital divide. To date, the literature on the digital divide has mainly focused on measuring the regional variance in the geographic supply of access to and the adoption of the infrastructure supporting the commercial internet (*6, 10*). Such studies have documented geographic variance in the supply and demographics of users. Periodic surveys of different scales by the Pew Charitable Trust (Pew, 2023) and the US government (*6*) have documented the changing adoption of broadband by US households, with age, income, and race playing critical roles in the adoption of frontier broadband, with related patterns arising in smartphones, tablets, and cord-cutting.

Our paper also makes several contributions to the field of digital literacy. The research most closely related to our goals has concentrated on variance in the skills and conceptual understanding required to navigate the internet successfully (*1, 8, 9*). No research has followed related principles in characterizing computing usage.

Measuring Digital Literacy

Telemetry Data Like other software firms that aspire to patch their software frequently, Microsoft uses online connections with devices to update their software. Our data are collected from anonymized Windows devices that share diagnostic data with Microsoft. This "telemetry data" includes logs of Windows devices' interactions with various applications and indicates time spent by sub-routines supporting applications. Our sample covers the period from October 2022 to March 2023. The data does not include personal information.[4]

Microsoft Windows installations are a subset of all household computing infrastructure. Although Microsoft Windows commands almost 73% of desktop OS share worldwide (64% in North America), our sample includes laptops and, tablets and other devices where Windows represents a smaller share of the market.[5] This means that our usage measures, although national in scope

[4] We only include information from consumer devices, such as PCs and tablets. We excluded virtual, test devices, and all business devices. In addition, to ensure that each device counts towards a single postal code, a device is affiliated with only the postal code with the most activity, presumably the owner's residence.

[5] See statistics reported at https://gs.statcounter.com/os-market-share/desktop/worldwideStatCounter, viewed on July 22, 2024.

and consistently collected across ZIP codes, may be less accurate in ZIP codes where there is a particularly large usage share of other operating systems.

Classifying Digital Usage We direct attention to the distinction between using a PC for information-processing purposes, such as spreadsheets and word processing, and using it for content-creation and computational activities, such as graphics and developer tools. Information-processing tasks prioritize data manipulation, calculation, and business operations. In contrast, content creation or computational activities may require hardware-software tools for graphic design or specialized tools. Each demands different software and skills.

We grouped applications into 12 categories based on their publishers, as noted in the Windows Store. We divide these categories into two large areas of usage based on the IEEE standard 3527.1 for digital literacy (*5, 11*), which states that digital literacy is one of the competencies for digital intelligence, which is further divided into three dimensions: 1) media and information literacy, 2) content creation and computational literacy, and 3) data and AI literacy. Our indices capture the first two dimensions; we leave the third for future research while noting that research suggests that digital literacy, as might be represented by the first two categories, is foundational for data and AI literacy (*12*).

Table 1 provides a description of all variables present in the Windows data. Note that the data provides device-level information regarding how many minutes a device spent on each application category over the period of a month.

Index construction An ideal index for measuring variance in the usage of computing devices across the US population should be constructed generally. It allows the index to be regularly updated and regionally disaggregated and lends itself to longitudinal analysis. A pragmatic index should, however, also find a point that trades off details with satisfying objectives for maintaining anonymity.

To achieve these conflicting objectives, we construct a measure of usage at the device level and then build weighted averages of these usage measures at the ZIP code level. By aggregating the data through privacy-enhancing technologies, our approach prevents the revelation of any private information.

We compute, for each device, device-level indices, which are weighted sums of the time spent

Table 1: **Description of the activity features present in the Windows data.** The table provides description of monthly activity data features for different application types in Windows, and examples. In addition to the activity features presented in the data, the telemetry system includes device location in ZIP code format.

Feature	Description	Applications
Media and Information Composite Index (MCI) features		
ss	Minutes of activity in software for visually organizing and analyzing tabular data	Excel, WPS Spreadsheets
dw	Minutes of activity in software for creating written content, such as reports and letters	Adobe Acrobat, Word, Wordpad
em	Minutes of activity in software for sending and receiving electronic messages	Thunderbird, Outlook, Mailbird
pt	Minutes of activity in software for organizing creating visual slideshows	PowerPoint, Canva, Prezi
md	Minutes of activity in software for playing and editing multimedia content	VLC player, Spotify, Windows Media Player
bw	Minutes of activity in software for browsing the internet	Firefox, Chrome, Opera, Edge
cc	Minutes of activity in software for communicating with others remotely	Zoom, Discord, Teams, Telegram
Content creation and computation composite index (CCI) features		
dv	Minutes of activity in software for creating, testing, and debugging code	Unity engine, Visual Studio, IntelliJ IDE
cd	Minutes of activity in software for remote file storage	One Drive, Dropbox, iCloud, Google drive
ut	Minutes of activity in software for system management and optimization	Remote desktop, printer management
sd	Minutes of activity in software that protects systems from cyberthreats	McAfee agent, Norton Security, Bitdefender
ct	Minutes of activity in software for creating and editing digital media	Adobe Photoshop, Virtual DJ, Blender

(in minute, during one month) in the applications corresponding to the two types of digital literacy defined in (*11*). The weights used in each sum are computed via principal component analysis (PCA). PCA, which has been previously leveraged in the index construction literature (*13–15*), is a natural choice for computing index weights as it assumes that information is captured in the variance of the data features. We publish the average of the device-level indices at a ZIP code level. Please refer to the supplementary materials for a detailed view on index construction and interpretation.

Demographic Correlates We matched the indices with information from the 2020 Census and 2021 American Community Survey (*16*) for each ZIP code. This information covers the demographic features of households in each ZIP code, including average income, education, population, area density, and broadband availability.

Exploring the New Digital Divide

We now turn to illustrating the variation in these indices.

Variation across U.S. counties The left panels of Figure 1 (A) and (B) show maps of MCI and CCI by US county, while broadband infrastructure is shown in Figure 1 (C). Across all measures, the counties with levels in the lowest (highest) decile of the indices are light (dark).

These maps show considerable variation in the MCI and CCI measures across US counties. A visual comparison shows that the two indices are not perfectly correlated and yield similar but different insights about usage. Interestingly, both are darker in urban counties but not to the same degree everywhere.

Visually comparing broadband with MCI and CCI maps provides further insight. It would have been hard to capture usage disparities by relying solely on the diffusion of broadband infrastructure, which is broadly distributed across geographies.

The data also confirm the well-documented existence of an urban vs. rural digital divide. The average MCI (CCI) index is 0.19 (0.29) in urban ZIP codes and -0.27 (-0.41) in rural ZIP codes. The standard deviation for MCI (CCI) urban area ZIP codes is smaller than for rural areas, 0.87 (0.74) and 1.05 (1.06), respectively. That implies a large part of the variance is reflected in urban-rural differences.

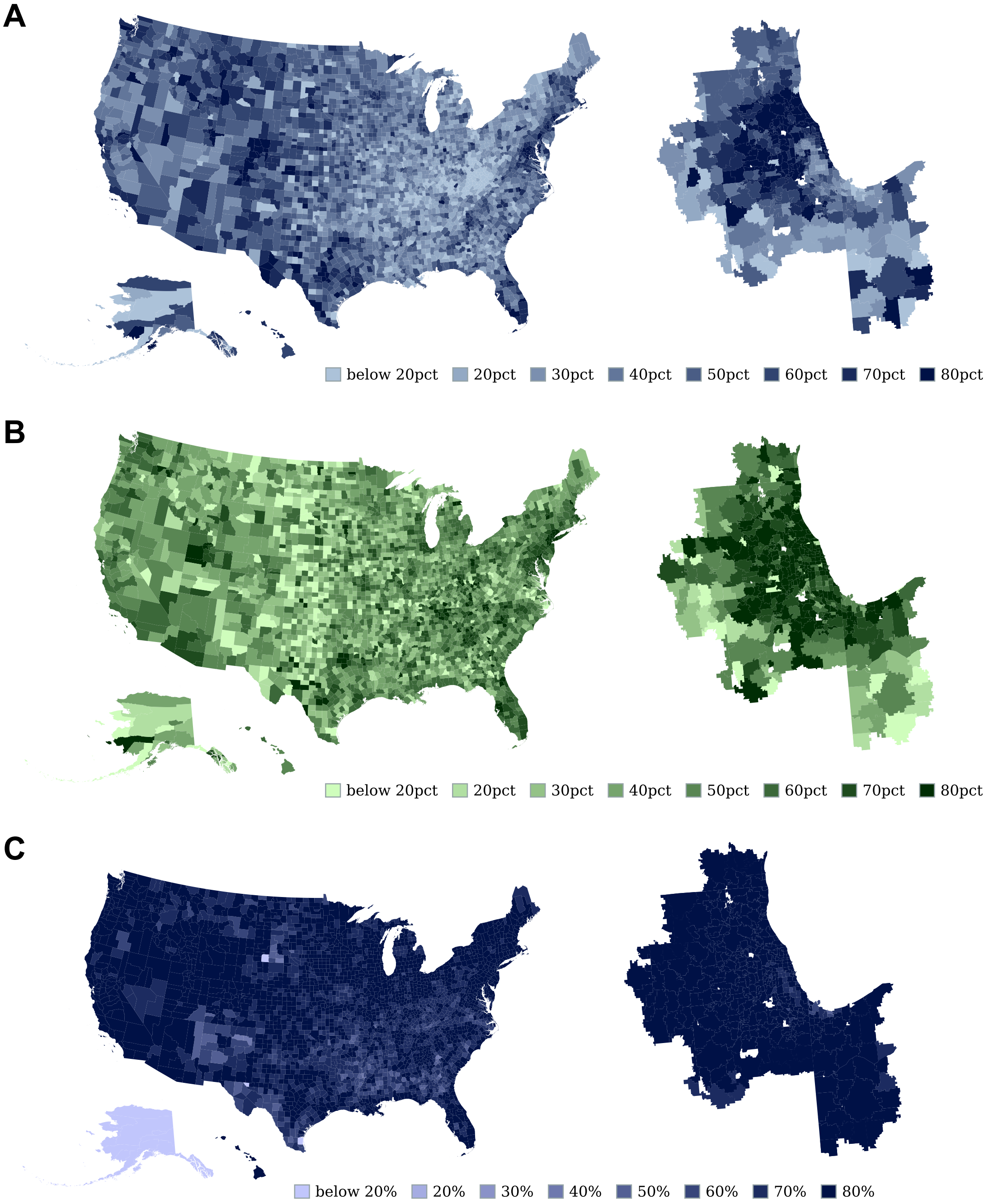


Figure 1: **Spatial distributions of media and information composite index, content creation and computation composite index, and broadband availability.** (**A**) Media and information composite index across US counties (left) and Chicago-Naperville-Elgin, IL-IN ZIP codes (right). (**B**) Content creation and computation index across US counties (left) and Chicago-Naperville-Elgin, IL-IN ZIP codes (right). (**C**) Broadband availability across US counties (left) and Chicago-Naperville-Elgin, IL-IN ZIP codes (right).

The means and variances define the distribution of experience in urban and rural regions. Urban and rural devices share but have limited overlapping experiences with usage. As an indication, the level that marks the upper quartile for the MCI (CCI) indices among urban ZIP codes is equivalent to the 37.0 (53.5) percentile reading for rural ZIP codes.

Those statistics indicate that the most extreme situations at the low (high) end disproportionately occur in rural (urban) settings. Notably, the wide variance of experiences in rural areas may also be due to more heterogeneity in what it means to be rural. Rural designations can include agricultural and mountainous regions and wealthy vacation areas in low-density settings like the Rocky Mountains.

These statistics also show wide variation in digital usage *within* urban areas, which would be hidden using less granular indices of usage. As a further illustration of this point, the right panels of Figures 1 (A) and (B) show variation in MCI, CCI, and broadband by ZIP code in the Chicago MSA. Similar to the county-level maps, the two indices show wide variations within the MSA. This variation would be missed entirely, focusing only on broadband indicators, which are broadly homogeneous within the MSA.

Socio-Economic Correlates of the New Digital Divide

We also examine correlations between the MCI and CCI indices and variables describing socio-economic and infrastructural differences across U.S. geographies, including educational attainment, per capita income, computing ownership, and broadband availability measures. We examine these correlations at the county level and then turn to a more detailed exploration at the ZIP code level.

County-level correlations Figure 2 illustrates the sensitivity of MCI and CCI to income (upper panel) and the share of residents obtaining a bachelor's degree or more (lower panel). Both are positively sloped, broadly illustrating that higher income is correlated with greater usage as measured by MCI and CCI. Contrasting the two figures, MCI is more sensitive to income and education than CCI.

Zip code level correlations Figure 3 examines these relationships at the ZIP code level. We report coefficients from univariate correlations with each variable in the upper panels and multivariate

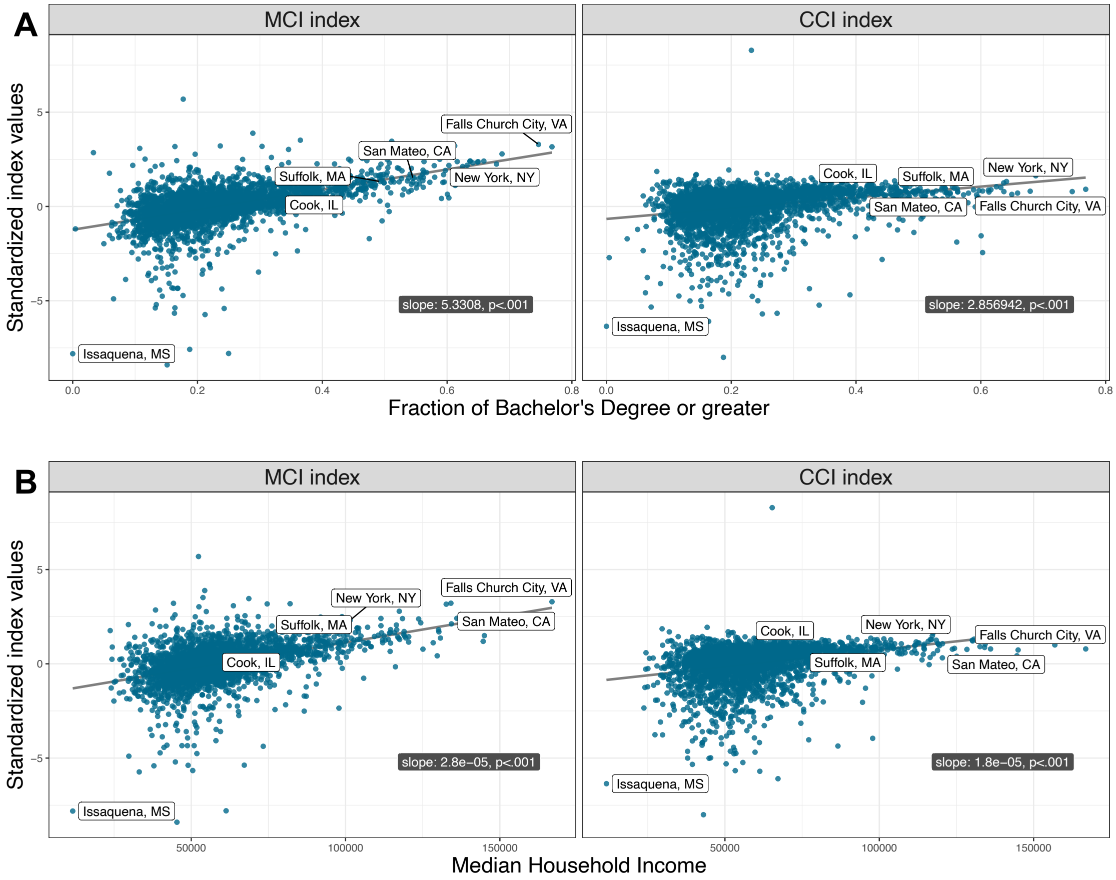


Figure 2: **Household income, educational attainment, and county-level usage indices.** This figure plots educational attainment (**A**) and median household income (**B**) against MCI and CCI index at the county level. County-level measures are generated from the ZIP-code-level data. Each ZIP code in the data is associated with a primary county. County-level measures are created by computing weighted sums of income, MCI, and CCI for each ZIP code, which lists that county as its primary county. The weights are the share of the total households in the county that are in that ZIP code. The plotted points are a randomly sampled 20 percent of all counties with the restriction that the labeled points must be included in the sample.

correlations that simultaneously include all controls, including fixed effects at the state level, in the lower panel of the figure. The inclusion of state-fixed effects controls for unobserved state-level factors that determine outcomes across ZIP codes within a state and thus identifies coefficients from variations across ZIP codes within states.

For ease of interpretation, in these analyses, the MCI and CCI indices and all exogenous variables were transformed into variables with a mean of zero and a standard deviation of one. This transformation yields coefficient estimates that provide a comparable quantitative estimate of how variance in each variable affects the endogenous variable.

The plot of the univariate correlations shows that indices are positively correlated with the log of per capita income and the share of the population with a bachelor's degree, even at more granular levels of analysis. The positive relationship between education and digital usage extends to more specific metrics of educational attainment, such as the share of the population with a bachelor's degree or with a STEM or business degree. Conversely, the indices negatively correlate with the share of the ZIP code population with some college.

Digital usage is lower in areas with an older population (as measured by the log of the median age), a higher share of males, and a higher share of Caucasians. The digital usage indices are also significantly higher in more populated and densely populated ZIP codes. We interpret these estimates as a proxy for thriving urban locations with amenities that attract heavy computer users. As expected for univariate correlations, we also see higher digital usage indices in areas with greater household broadband availability and computing.

When examining these relationships in multivariate correlations accounting for state-fixed effects and the broad socio-demographic and infrastructure variables discussed above, we see similar results for MCI. The relationship between the digital usage indices and income continues to be positive and significant for MCI and CCI, even in this more stringent specification. However, the magnitude of the coefficient drops to 0.09 (0.06) for MCI (CCI). Income is strongly correlated with digital usage even after netting out the role of possible confounding factors, such as the quality of broadband and computing adoption in the ZIP code, which are accounted for in the multivariate regressions.

Education continues to play a significant role in understanding the ZIP-code level MCI index (the coefficients on education variables are relative to the omitted category, which is the share with a high school degree or less). Whether a ZIP code has a high fraction of households with a bachelor's degree is essential for MCI (the coefficient is 0.22 for MCI, which is large). We see similar positive (albeit smaller) coefficients for the share of STEM graduates and the share of business graduates (0.02). In contrast, we see weaker (if not negative) correlations between educational attainment measures and CCI. The fraction who obtained some college has a minimal (0.02) effect on CCI compared to the omitted category, the share with a high school degree or less. In contrast, the share of STEM and business graduates show a small *negative* correlation with CCI at the ZIP code level, which may be due to the diversity of signals beyond college education that employers use for professions like cybersecurity that use CCI applications heavily (*17*).

The two indices also show opposite correlations for median age and the share of Caucasians (both positive for MCI and negative for CCI).

The most crucial variable in these multivariate correlations is the ZIP code's population size. A change in the log of population size by one standard deviation raises the MCI (CCI) by 0.24 (0.62) of a standard deviation. Log population density also increases CCI, with a small coefficient of 0.07. After controlling for available infrastructure, these act as proxies for other amenities in urban locations.

Finally, in the multivariate regressions, the presence of broadband infrastructure does not predict MCI with statistically significant coefficient estimates, nor does being in a neighborhood with more PC adoption. Broadband makes a small (0.04) contribution to CCI.

Summary and Implications

By examining the digital usage of over forty million devices, this study provides new evidence on variance in computing usage across and within US geographies. It suggests a significant digital divide in the broad use of computing despite the considerable progress made in the diffusion of broadband infrastructure and computing hardware.

We uncover factors correlating with the variance in computing usage, most notably household income and education. Because the usage levels are linked to such demographic features, these findings call for comprehensive strategies that address access, usage, and skills development.

We also show that city population correlates with these indices and that this effect reflects more

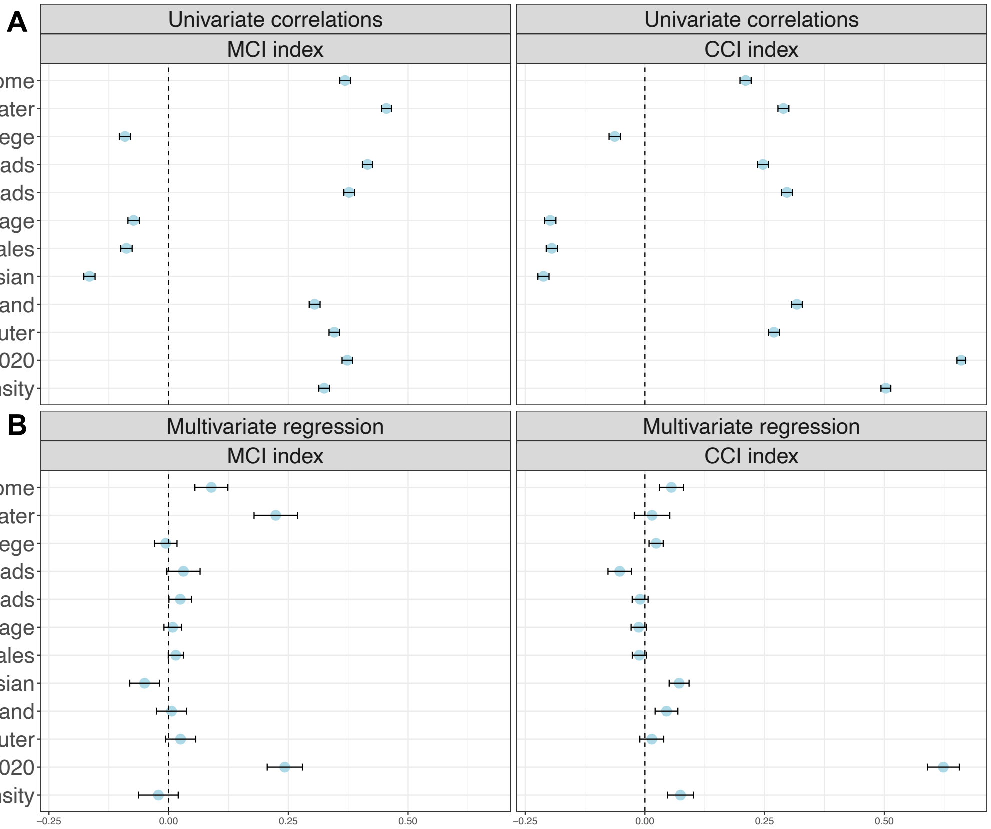


Figure 3: **Univariate correlations and multivariate regression coefficients of standardized MCI and CCI indices.** The top panel (**A**) plots univariate correlations between standardized index measures and standardized income, education, demographic, technology, and population variables. The bottom panel (**B**) plots the coefficient estimates from multivariate regressions using the same variables. The multivariate regression also includes *State* fixed-effects (not shown), which are dummy variables indicating the state where the ZIP code is located. 95% confidence intervals are shown on each point estimate.

than the effects of more infrastructure. Hence, more than simply closing the divide in broadband access will be required to alter the variance in usage.

Our goal has been to provide new measurements and evidence. We do this in a way consistent with internationally recognized definitions of digital usage to make them actionable from a policy standpoint. We will also make these indices available on GitHub for the research community.[6] We look forward to more research on digital usage.

[6] Data set availability is conditioned on paper acceptance. Link to the data will be added here once paper gets accepted.

References and Notes

1. E. Hargittai, M. Micheli, Internet skills and why they matter. *Society and the internet: How networks of information and communication are changing our lives* **109**, 109–124 (2019).

2. L. Hitt, P. Tambe, Broadband adoption and content consumption. *Information Economics and Policy* **19** (3-4), 362–378 (2007).

3. A. Goldfarb, J. Prince, Internet adoption and usage patterns are different: Implications for the digital divide. *Information Economics and Policy* **20** (1), 2–15 (2008).

4. A. Boik, S. Greenstein, J. Prince, The persistence of broadband user behavior: Implications for universal service and competition policy. *Telecommunications Policy* **43** (8), 101820 (2019).

5. A. Knowles, A. Hampton, J. Gooddell, White Paper-Concepts for Classification of Adaptive Instructional Systems. *Concepts for Classification of Adaptive Instructional Systems* pp. 1–14 (2023).

6. M. Martin, *et al.*, Computer and internet use in the United States: 2018. *American Community Survey Reports* (2021).

7. P. Dine, Measuring Digital Literacy Gaps Is the First Step to Closing Them (2024), `https://itif.org/publications/2024/04/26/measuring-digital-literacy-gaps-is-the-first-step-to-closing-them/`.

8. L. Robinson, *et al.*, Digital inequalities and why they matter. *Information, communication & society* **18** (5), 569–582 (2015).

9. A. Scheerder, A. Van Deursen, J. Van Dijk, Determinants of Internet skills, uses and outcomes. A systematic review of the second-and third-level digital divide. *Telematics and informatics* **34** (8), 1607–1624 (2017).

10. C. Forman, A. Goldfarb, S. Greenstein, Geographic inequality and the internet. *Handbook of digital inequality* pp. 31–45 (2021).

11. IEEE Association, IEEE standard for digital intelligence (DQ)–Framework for digital literacy, skills, and readiness (2020).

12. D. Long, B. Magerko, What is AI literacy? Competencies and design considerations, in *Proceedings of the 2020 CHI conference on human factors in computing systems* (2020), pp. 1–16.

13. J. Blumenstock, G. Cadamuro, R. On, Predicting poverty and wealth from mobile phone metadata. *Science* **350** (6264), 1073–1076 (2015).

14. L.-M. Asselin, Composite indicator of multidimensional poverty. *Multidimensional Poverty Theory* (2002).

15. D. Filmer, L. H. Pritchett, Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. *Demography* **38** (1), 115–132 (2001).

16. S. Manson, J. Schroeder, D. Van Riper, T. Kugler, S. Ruggles, IPUMS National Historical Geographic Information System: Version 17.0 (2022), `http://doi.org/10.18128/D050.V17.0`.

17. J. Marquardson, A. Elnoshokaty, Skills, Certifications, or Degrees: What Companies Demand for Entry-Level Cybersecurity Jobs. *Information Systems Education Journal* **18** (1), 22–28 (2020).

18. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in *Theory of cryptography conference* (Springer) (2006), pp. 265–284.

19. C. Dwork, A. Roth, *et al.*, The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* **9** (3-4), 211–407 (2014).

20. F. D. McSherry, Privacy integrated queries: an extensible platform for privacy-preserving data analysis, in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (2009), pp. 19–30.

21. H. Imtiaz, A. D. Sarwate, Symmetric matrix perturbation for differentially-private principal component analysis, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE) (2016), pp. 2339–2343.

Author contributions: Conceptualization: MP, SG, RS, PT, LR, TG, RD, AK, JLF
Methodology: MP, SG, RS, PT, LRD
Investigation: MP, SG, RS, PT
Visualization: MP, PT
Writing – original draft: MP, SG, RS, PT
Writing – review & editing: MP, SG, RS, PT, LRD, TG, RD, AK, JLF

Competing interests: MP, LRD, TG, AK, RD, JLF are current employees of Microsoft.

Data and materials availability: The data used in this study, which measures digital usage for over twenty eight thousand ZIP codes in the United States, is made publicly available for further research purposes. The dataset, collected by Microsoft during operating system updates for 40 million Windows devices across U.S. households, represents the most comprehensive measure of US household computing usage ever assembled.

To facilitate further research, we have made the ZIP code level indices, namely the Media and Information Composite Index (MCI) and the Content Creation and Computation Composite Index (CCI), publicly available on GitHub.

Supplementary materials

Materials and Methods
Supplementary Text
Figure S1
Tables S1 to S4

Supplementary Materials for

The New Digital Divide

Mayana Pereira[1*†], Shane Greenstein[2†], Raffaella Sadun[2†], Prasanna Tambe[3†]

Lucia Ronchi Darre[1], Tammy Glazer[1], Allen Kim[1] Rahul Dodhia[1], Juan Lavista Ferres[1]

[*]Corresponding author. Email: mayana.wanderley@microsoft.com

[†]These authors contributed equally to this work.

This PDF file includes:

Materials and Methods

Data Description and Construction

Windows Telemetry Data We used telemetry data collected from anonymized Windows devices that share diagnostic data with Microsoft to construct the data set used in this work. The telemetry data captures the engagement time of each device with different applications. More precisely, the data provides information on the time devices spend on different application categories in the period of one month (in minutes) and the Postal Code the device spent the majority of its time in the given month.

The system collects up to 9,000 minutes of activity per application category per month, and truncates all activity values greater than this threshold. Device-level activity minutes is collected at a monthly basis, and we obtained six months of data, from October 2022 to March 2023. The data set was collected as part of Microsoft's ongoing efforts to enhance the performance and security of its software and services, and includes a data sample from 40 million consumer devices. The set of devices includes desktops, laptops and tablets. Virtual and test devices were not included in the data set.

The Windows devices data set doesn't contain any personally identifiable information (PII), including IP addresses. We used differential privacy (*18*) in every data analysis presented in this paper to provide strong privacy guarantees to all devices in the data set. Differential privacy is the gold standard of privacy-preserving data publication. We privatize all values in our data analysis using OpendDP Python library.

Note that, in the table above, the data contains information about time spent in application categories during one month. The data does not provide information about specific applications utilized by the devices. The application publisher is responsible for defining the category of each application, and such information is available in the Windows Store. The applications categories present in the windows telemetry data are described in Table 1 (Main Text).

Consumption of Digital Applications Data Set We utilized the Windows Telemetry data, described in section , to construct a data set that provides information regarding consumption of digital applications by households within each postal code in the U.S. [7] In our work, the term *consumption* refers to engagement of devices with different categories of applications.

We measure the consumption of digital applications of a population from two different aspects: consumption of media and and information applications and consumption of content creation and computational application. This categorization follows the approach defined by the *IEEE Standard for Digital Intelligence (DQ) — Framework for Digital Literacy, Skills, and Readiness* (*11*), which defines the different types of engagement a user can have with digital devices. Moreover, the IEEE standard for digital intelligence defines dimensions of digital literacy, which includes *media and information literacy* and *content creation and computational literacy*. Media and information literacy involves finding, organizing, analyzing, and evaluating information. Content creation and computational literacy involves understanding the theories, practices, and processes of digital content creation, curation, and computational thinking. Individuals with this literacy possess algorithmic literacy, such as programming and digital modeling. They conceptualize, build, organize, create, adapt, and share knowledge, digital content, and technology.

We compute for each device in the data, a weighted sum of the time spent (in minutes, during one month) in the applications corresponding to the two types of digital literacy defined in (*11*). The weights used in each of the sums are computed via principal component analysis (PCA).

PCA-based weights Principal component analysis, a technique often used to summarize multiple features (dimensionality reduction), has also been leveraged in the index construction literature (*13–15*). Our index construction utilizes principal component analysis to compute weights W_i.

The purpose of our index is to capture the variance in the data in a single metric. Principal component analysis (PCA) rises as a natural choice for index construction since it assumes that information is captured in the variance of the features. Therefore, we use PCA to reduce dimensionality while weighting each variable by its contribution to the data variance. In PCA, the first principal component is the projection that captures the largest variance of the data. With this in mind, our index is a linear combination of the variables in our data, where each variable is weighted by its contribution to the variance of the first principal component. More specifically, the weights of our index represents shares of the variance. We obtain such variance shares by squaring the

[7] the data set includes postal codes that are within the 50 states (plus DC) that contains more than 30 households.

loadings of the first principal component.

We denote by w_m the weight vector for the media and information composite index. The weight vector w_m is computed via PCA over the subset of the data that contains the set of variables $\{ss, dw, em, pt, md, bw, cc\}$. Analogously, we denote by w_c the weight vector for the content creation and computational composite index, which is computed from the subset of the data that contains the set of variables $\{dv, cd, ut, sd, ct\}$.

Construction of consumption of digital applications composite indices

Based on the two dimensions of digital literacy defined by the *Standard for Digital Intelligence*, we measure consumption of two groupings of digital applications: consumption of media and information applications and consumption of content creation and computational applications.

We construct indices for measuring variance in the usage of computing devices across the US population. Our proposed indices capture the consumption applications associated to each type of literacy (*Media and information literacy* and *Content creation and computational literacy*).

The process for computing the MCI and the CCI is almost identical. The difference lies in the set of activity features utilized to compute each index. MCI utilizes features ss, dw, em, pt, md, bw, cc, and CCI utilizes features dv, cd, ut, sd, ct.

Index calculation Consider the first category of digital literacy, media and information literacy. The Media and information composite index (MCI) is defined as:

$$\mathrm{MCI_i} = w_m^\intercal x_i, \tag{S1}$$

where $\mathrm{MCI_i}$ denotes the MCI index for machine i, w_m is the vector of weights, $x_i = \log_2(1+m_i)$ is the element-wise logarithm of the vector m_i, where $m_i = (ss_i, dw_i, em_i, pt_i, md_i, bw_i, cc_i)$ represents the usage (in minutes) of device i across the different digital applications related to media and information literacy. The reason for the choice of logarithm is to capture the scale of usage of each application, making the index less sensitive to small variations in usage.

The *Content creation and computational composite index* CCI is defined as:

$$\mathrm{CCI_i} = w_c^\intercal y_i, \tag{S2}$$

where $\mathrm{CCI_i}$ denotes the CCI index for machine i, w_c is the vector of weights, $y_i = \log_2(1 + c_i)$ is the element-wise logarithm of the vector c_i, where $c_i = (dv_i, cd_i, ut_i, sd_i, ct_i)$ represents the usage (in minutes) of device i across the different digital applications related to content creation and computational literacy.

ZIP code level aggregations To generate the average index for each ZIP code z, we then calculate a weighted average where the weights for the average are the total time each machine i spends online within a ZIP code.

The $\overline{\mathrm{MCI_z}}$ and $\overline{\mathrm{CCI_z}}$ average for a ZIP code z are computed as:

$$\overline{\mathrm{MCI_z}} = \frac{\sum_{i \in z}(\mathrm{MCI}_i \cdot T_{m_i})}{\sum_{i \in z} T_{m_i}}, \tag{S3}$$

and,

$$\overline{\mathrm{CCI_z}} = \frac{\sum_{i \in z}(\mathrm{CCI}_i \cdot T_{c_i})}{\sum_{i \in z} T_{c_i}}, \tag{S4}$$

where MCI_i denotes the media and information composite index for device i, CCI_i denotes the content creation and computational composite index for device i, $T_{m_i} = 1^\intercal m_i$ denotes the total time device i spends on media and information applications, and $T_{c_i} = 1^\intercal c_i$ denotes the total time device i spends on media and information applications.

Differentially Private MCI and CCI

The computations of MCI and CCI are done via a privacy-preserving technique called differential privacy (*18*).

Differential privacy is a rigorous privacy notion used to protect an individual's data in a dataset disclosure. We present in this section notation and definitions that we will use to describe our privatization approach. We refer the reader to (*19*), (*20*) and (*18*) for detailed explanations of these definitions and theorems.

Pure Differential Privacy. A randomized mechanism $\mathcal{M} : \mathcal{D} \longrightarrow \mathcal{A}$ with data base domain $\mathcal{D}$ and output set $\mathcal{A}$ is ϵ-differentially private if, for any output $A \subseteq \mathcal{Y}$ and neighboring databases

$D, D' \in \mathcal{D}$ (i.e., D and D' differ in at most one entry), we have

$$\Pr[\mathcal{M}(D) \in A] \leq e^{\epsilon}\Pr[\mathcal{M}(D') \in A] \tag{S5}$$

The definition of neighboring databases used in this work is device-level privacy. Device-level privacy defines neighboring to be the addition or deletion of a single device in the data and all possible records of that device. Informally, the definition above states that the addition or removal of a single device in the database does not provoke significant changes in the probability of any differentially private output. Therefore, differential privacy limits the amount of information that the output reveals about any device.

A function f (also called query) from a dataset $D \in \mathcal{D}$ to a result set $A \subseteq \mathcal{A}$ can be made differentially private by injecting random noise to its output. The amount of noise depends on the sensitivity of the query (*18*).

To compute the differentially private $\overline{\text{MCI}}$ and $\overline{\text{CCI}}$, we perform a two-step computation. We start by computing, the differentially private weights of the first component of principal component analysis done at the national level data. After we obtain the dp-weights, we utilize additive noise mechanisms to compute MCI and CCI, as given in equations S3 and S4.

Differentially private ZIP code level index calculation

The differentially private index calculation is done on two steps. The first step we run a differentially private principal component analysis to compute the weights used in the index calculation. Once the weights are calculated, we proceed by computing the device-level MCI followed by the computation of differentially private aggregates $\overline{\text{MCI}}$ at ZIP code level. The differentially private aggregates are computed using the Laplace mechanism.

Differentially private PCA To compute the differentially private eigenvector and eigenvalue of the covariance matrix, we utilize the differentially private PCA from OpenDP,[8] which follows the algorithm proposed in (*21*) and preserves ϵ-differential privacy.

[8] https://docs.opendp.org/en/stable/user/measurements/pca.html

Laplace mechanism Let $\mathsf{Lap}(\lambda)$ be the Laplace distribution with 0 mean and scale λ. The Laplace distribution has a probability density function $\mathsf{Lap}(x|\lambda) = \frac{1}{2\lambda}e^{-\frac{x}{\lambda}}$, and can be used to obtain an ϵ-differentially private answers numeric queries (*18*).

Let $f : \mathcal{D} \longrightarrow \mathbb{R}^n$ be a numeric query. Let x be the query input and ϵ the privacy parameter. The Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + (\eta_1, \ldots, \eta_n) \tag{S6}$$

where η_i are drawn from the Laplace distribution $\mathsf{Lap}(\frac{\Delta f}{\epsilon})$.

The Laplace mechanism is an additive noise mechanism and preserves ϵ-differential privacy (*18*).

The differentially private ZIP code level index aggregations are calculated by applying the Laplace mechanism to the sums found in the numerators and denominators of Equations S3 and S4.

Privacy Loss The privacy loss computation is a straightforward application of the parallel and sequential composition properties of differential privacy mechanisms (*20*).

The privacy-loss results from the generate the indices results from weight calculations (PCA-based) and device index aggregations at ZIP code level.

The privacy-loss resulting from the PCA analysis is of $\epsilon = 0.5$ for each index. The total privacy-loss incurred from PCA analysis is of $\epsilon = 1.0$. The weights calculation were done over a baseline month (March 2023).

The aggregation process for the index resulted in a privacy-loss of $\epsilon = 3.0$.

The histograms in Figure S1 are differentially private. Each histogram was privatized using a Laplace mechanism, with privacy-loss of $\epsilon = 0.05$.

The total privacy-loss of the data release is $\epsilon = 4.1$.

Implementation Details All Laplace mechanism and PCA implementations utilized in this project were developed by the OpenDP library. The OpenDP library includes a comprehensive set of differential privacy mechanisms, algorithms, and validator. The library is open source, and is maintained and vetted by OpenDP community.

We report index values for all 28,199 ZIP codes that part of the Census data. For ZIP codes without any devices, we set the number of devices to zero and minutes of activity to zero before performing differentially private aggregations.

Data Limitations

The data set described in this section offers estimates of usage of applications (media and information applications, and content creation and computational applications) for 28,199 ZIP codes in the U.S. It is one of the most extensive open data sets available for gaining insights into digital device usage. However, there are a few limitations to consider.

Firstly, the data set only covers desktop and laptop devices, and does not include information on application usage on mobile devices. This limitation restricts our understanding of the complete usage landscape across different device types.

Secondly, the estimates provided are based solely on data from Windows devices and do not take into account usage data from other operating systems. This limitation may introduce biases and prevent a comprehensive analysis of application usage across different platforms.

While acknowledging these limitations, the data set still offers valuable insights into application usage for the specified ZIP codes and can serve as a valuable resource for understanding digital device usage trends.

Supplementary Text

Data Exploration

The data set with ZIP code level MCI and CCI aggregates provides a summarization of the usage of digital applications by the ZIP code population. For each device, we collect the log of minutes spent in each device category, and we compute the weighted sum of time (log minutes) spend in each category.

PCA weights

We compute the first component loadings utilizing principal component analysis for each set of variables (ss, dw, em, pt, md, bw, cc) and (dv, cd, ut, sd, ct). By squaring each loading, we obtain

Table S1: **Media and information consumption weights.** The table presents the weights calculated using the PCA algorithm. Note that the feature with highest weight, for media and information applications, is dw, which corresponds to applications for creating written content. For content creation and computational applications, *ut* is the feature with the highest weight, and it correspondes to applications for system management and optimization.

MCI							
Feature	ss	dw	em	pt	md	bw	cc
Weight	0.212	0.263	0.132	0.098	0.082	0.142	0.067
CCI							
Feature		dv	cd	ut	sd	ct	
Weight		0.14	0.21	0.294	0.186	0.17	

the share of variance that each feature represent. The variance shares are used as weights in our index calculation. The resulting weights are the following:

Table S1 shows the weights for the MCI index, and Table S4 shows the weights for the CCI index. One notable thing in Table S1, which shows the MCI weights, is that *time spent in document authoring applications in a month* is the variable with the largest variance share. On the other hand *time spent in communication applications in a month* is the variable with the smallest variance share. When analyzing CCI weights, we note that *time spent in utilities applications in a month* is responsible for the largest share in variance and *time spent in developer applications in a month* is the is the variable with the smallest variance share.

Index interpretation

The indices summarizes the consumption of digital applications of a device by computing the weighted sum of the time spent (log minutes)[9] in each application category. The intuition behind the index is to able to capture intensity of consumption (time spent) as well as variety of consumption, while weighting each category of application according to how much it contributes to the variance in the data. As an example, consider a ZIP codes with 3 devices, d_1, d_2 and d_3, each with the

[9] If time spent is zero, we do not compute the log. *Time spent in minutes* are integer values.

Table S2: **Example of media and information composite index per device.** The table provides examples of three distinct devices, their corresponding features and device-level index.

	Time per category (minutes per month)								
Device	ss	dw	em	pt	md	bw	cc	total time	MCI
d_1	30	200	100	25	5	600	300	1260	4.46
d_2	0	15	120	0	360	1200	45	1740	3.09
d_3	10	20	45	30	20	500	30	655	3.47

following monthly consumption :

Note that, although d_2 has the highest consumption intensity, spending over 1700 minutes across media and information applications, its index (3.09) is the lowest among all devices. This is because the index computation takes into consideration the variety of consumption as well which application category the device spends most time on, given that category weights plays a significant role in the computation. In the ZIP code exemplified in Table S7, the average MCI is computed as:

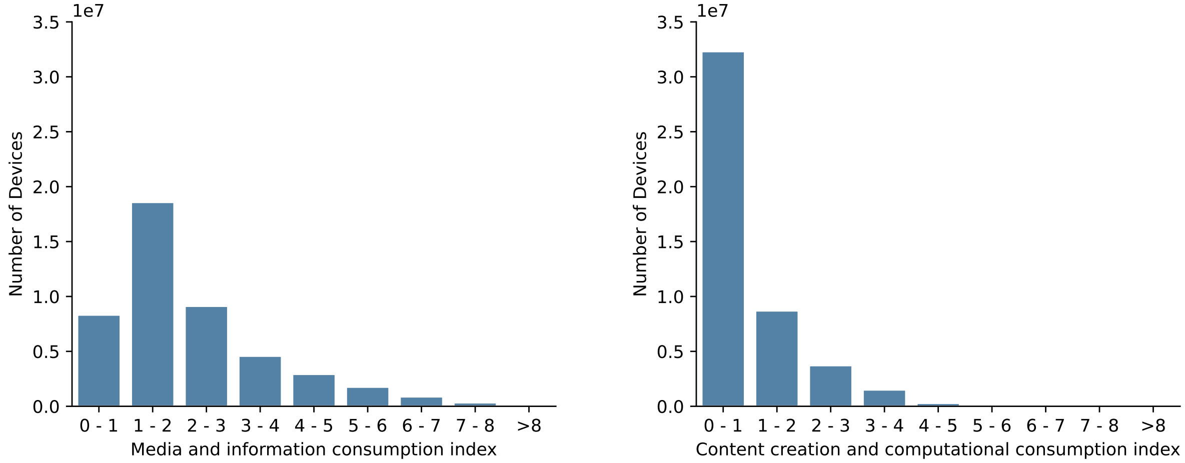
$$\overline{\text{MCI}} = \frac{(1260 \times 4.46) + (1740 \times 3.09) + (3.47 \times 655)}{(1260 + 1740 + 355)} = 3.62 \tag{S7}$$

Device-level index distribution

The distributions of device-level indices in the US are presented in Figure S1. For a device to achieve a MCI or CCI of ≈ 13.135 in a month, it must use all application categories (corresponding to the index) for 9000 minutes (150 hours) or more during that month.

Summary Statistics

Table S3 presents definitions and summary statistics of each ZIP code's indices and key demographic correlates. We first consider whether the MCI and CCI indices contain similar information or measure alternative uses. A good indicator of their differences comes from simple correlations. The mean values are positively correlated but not highly. In the sample of Metropolitan Statistical Areas with populations over five hundred thousand (one million), the correlations of the means are 0.288 (0.330). However, we do see more correlation among the spreads. For example, the coefficient of



(a) MCI device level index

(b) CCI device level index

Figure S1: **Distribution of device-level MCI and CCI in the US in March of 2023**. Although the maximum value that MCI and CCI can achieve is $\log_2(9000) \approx 13.1357$, overwhelming majority of devices have MCI smaller than 8 (**A**), and majority of devices have CCI smaller than 4 (**B**) .

variations for MCI and CCI are positively correlated at 0.578 (0.582), which is moderately high. A similar result is obtained for an alternative measure of spread, the size of the difference between the ninetieth and tenth percentiles, where the correlation is high, at 0.716 (0.684). In other words, the central point for MCI and CCI moderately correlate, but if one is spread out, then it is likely that the other is as well. As we will see, this is because some of the same factors create more variance in both indices.

Interestingly, the indices give similar information about low achievement but differ in their assessment of high achievement. The indication for the tenth percentile for the MCI and CCI index is highly correlated at 0.760 (0.806), but the indices are not correlated at the ninetieth percentile, where the correlation is -0.058 (0.024). Table S4 shows that the pairwise correlations between the indices, household computers and broadband availability.

Table S3: **Summary statistics for the ZIP-code measures.** This table reports summary statistics for the ZIP code level MCI and CCI index variables, as well as for census variables. N=28,199.

Variable	Mean	S.D.	Min	Max
Standardized MCI index	0.000	1.000	-6.612	6.563
Standardized CCI index	0.000	1.000	-4.446	8.201
Logged per capita income	10.297	0.364	7.161	13.125
Share Bachelor's or greater	0.254	0.159	0.000	1.000
Share Some college or Associates	0.301	0.080	0.000	1.000
Share High school or less	0.445	0.155	0.000	1.000
Share STEM grads	0.120	0.090	0.000	0.853
Share Business grads	0.051	0.042	0.000	0.661
Logged median age	3.734	0.194	2.186	4.454
Share Male	0.500	0.029	0.203	0.956
Share Caucasian	0.761	0.222	0.000	1.000
Share houshold broadband	0.827	0.107	0.000	1.000
Share houshold computer	0.741	0.134	0.000	1.000
Logged population (2020)	8.320	1.635	0.693	11.837
Logged population density	5.111	2.265	0.000	12.491
Logged households	7.400	1.552	3.932	10.631

Table S4: **Pairwise correlations between tech indicators.** The table shows pair-wise correlations among the following variables: CCI index, MCI index, household broadband availability and household computer ownership.

	(1)	(2)	(3)	(4)
(1) Standardized CCI Index	1			
(2) Standardized MCI Index	0.51	1		
(3) Household broadband	0.32	0.3	1	
(4) Household computer	0.27	0.35	0.77	1