MANAGEMENT SCIENCE
Articles in Advance, pp. 1–22
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1875

The Effect of Content on Global Internet Adoption and the Global "Digital Divide"

V. Brian Viard
Cheung Kong Graduate School of Business, 100738 Beijing, China, brianviard@ckgsb.edu.cn

Nicholas Economides
Stern School of Business, New York University, New York, New York 10012; and Haas School of Business, University of California, Berkeley, Berkeley, California 94720, economides@stern.nyu.edu

A country's human capital and economic productivity increasingly depend on the Internet as a result of its expanding role in providing information and communications. This has prompted a search for ways to increase Internet adoption and narrow its disparity across countries—the global "digital divide." Previous work has focused on demographic, economic, and infrastructure determinants of Internet access that are difficult to change in the short run. Internet content increases adoption and can be changed more quickly; however, the magnitude of its impact, and therefore its effectiveness as a policy and strategy tool, has until now been unknown. Quantifying the role of content is challenging because of feedback (network effects) between content and adoption: more content stimulates adoption, which in turn increases the incentive to create content. We develop a methodology to overcome this endogeneity problem. We find a statistically and economically significant effect, implying that policies promoting content creation can substantially increase adoption. Because it is ubiquitous, Internet content is also useful to affect social change across countries. Content has a greater effect on adoption in countries with more disparate languages, making it a useful tool to overcome linguistic isolation. Our results offer guidance for policymakers on country characteristics that influence adoption's responsiveness to content and for Internet firms on where to expand internationally and how to quantify content investments.

Keywords: Internet; technology adoption; economic development; two-sided markets; network effects; technology diffusion; digital divide; language
History: Received November 2, 2011; accepted October 5, 2013, by Lorin Hitt, information systems. Published online in *Articles in Advance*.

1. Introduction

The number of Internet users has exploded since its commercialization in the early 1990s. From approximately 10.1 million users in early 1992, the Internet had expanded to almost 1.6 billion by 2009.[1] However, this growth has been very uneven across countries, with penetration rates varying from 90% to nearly 0% (see Figure 1). This global "digital divide" is of concern because Internet access is increasingly important for economic productivity and a well-informed citizenry as more information is accessed online.[2] As a consequence, there is a large literature examining economic and social determinants of cross-country Internet adoption but focuses almost exclusively on factors that are fixed in the short run. We focus on a factor that can be changed quickly: Internet content.

It is well understood that more Internet content in a language will lead to more adopters who use that language. As a United Nations (UN) report asserts, "Availability of *content*, in an appropriate *language* also affects the diffusion of the Internet. After all if you cannot find content in your language and you do not read other languages, how can you use the Internet?" (ITU 1999, p. 3, italics in original). What is not known is the magnitude of the effect of content on adoption. This has important policy implications. Because content is more easily altered than economic, educational, or infrastructure conditions, it offers governments and nongovernmental organizations (NGOs) a means to more quickly influence Internet diffusion. The exact magnitude of the effect is also relevant for Internet firms that rely on a user base for advertising or subscription revenue. It is important in evaluating the trade-off between investing in content creation to build the user base indirectly versus marketing efforts to attract new users directly. Our

[1] World Development Indicators Database (2010b).

[2] For an aggregate study on the link between the Internet and productivity, see Litan and Rivlin (2001); compare with a critique by Gordon (2000). Industry-specific studies include Goolsbee (2002) in health insurance and Scott Morton et al. (2001) in car retail. The International Telecommunication Union (ITU) (1999) provides a policy perspective on its economic and social role.

estimates quantify the effectiveness of this "build content and they will adopt" strategy and allow us to offer some guidance to Internet firms in making such investments.

If content sufficiently stimulates adoption, the ability to target content by language suggests a useful strategy to narrow the global digital divide. The UN has suggested content's role in reducing this divide, stating, "The dominance of European languages has limited the spread of Internet use by excluding those not fully literate in those languages" (Mutume 2006, p. 14). It would also suggest that content production is an effective strategy for firms to expand their user base. However, the question remains how effectively content stimulates adoption.

Content has a statistically and economically significant effect on adoption, implying that it is an effective policy and strategic tool. Our estimates explicitly recognize language as the conduit from content to adoption, confirming that creating content in underserved languages is an effective policy to address the global digital divide. We quantify content's effect on adoption in four different ways, but all indicate a large effect. First, we find an elasticity of adoption with respect to content of 0.31—about three-fourths the price elasticity of adoption. Second, a country one standard deviation above the mean level of relevant content has an adoption rate 2.0 percentage points or 20% higher than the mean adoption rate of 9.9 percentage points in the sample. Third, the magnitude of content's effect is about one-third that of the gross domestic product (GDP) (the most significant driver) and is stronger or of similar strength to that of other economic, infrastructure, and demographic factors that significantly affect adoption. Fourth, the annual rate of content creation in our sample increased adoption by 6.0% to 7.8% annually.

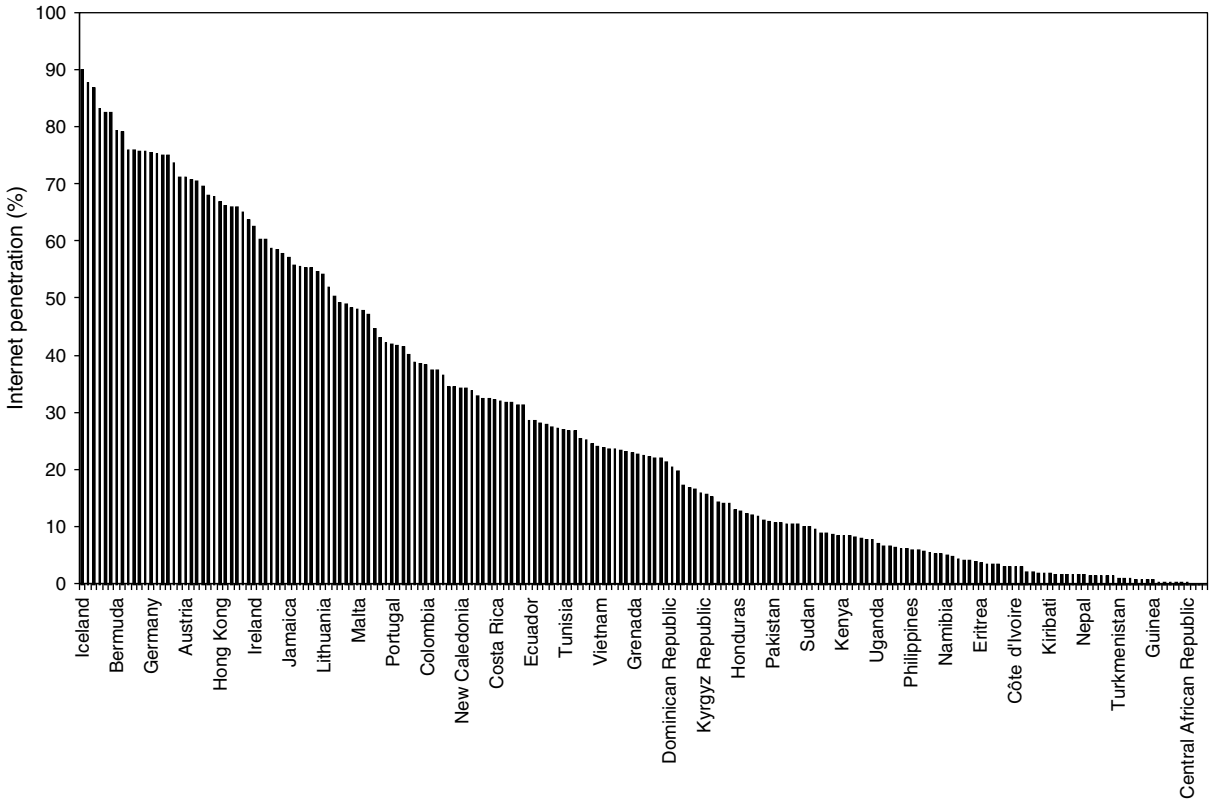
To further inform policy-making and firm strategies, our model can identify country characteristics that affect adoption's sensitivity to content. Content has greater influence in countries with better infrastructure, as measured by the extensiveness of the domestic phone system and international gateway speeds. This suggests Internet content providers wishing to access international markets should target such countries, and infrastructure investment is a means for governments to stimulate adoption. Content also has more influence in countries with weaker intellectual property protection, consistent with less costly and more widely available content. Thus, content providers who can sufficiently protect their content will experience greater uptake in international markets with weaker protections, and governments setting copyright policies face a trade-off between dynamic incentives to create content and its usage once created.

We also identify an important role for the Internet in overcoming linguistic isolation. Content affects adoption more in countries with more disparate languages. This suggests that creating content targeted at populations that speak languages uncommon in their surroundings may reduce their isolation. The predominance of English-language Internet content has been cited as an important dimension of inequality between social and linguistic groups (see DiMaggio et al. 2004). This result parallels that of Sinai and Waldfogel (2004), who find that the Internet helps overcome racial isolation in the United States. This also suggests an opportunity for Internet firms to target such populations.

Internet service is a two-sided market—user adoption depends on content availability, and vice versa. This feedback makes it difficult to empirically isolate content's effect on adoption. Estimating the causal effect of content is further complicated by the likely presence of unobserved country-specific factors that drive both content production and adoption. In particular, populations of countries with a high desire for Internet usage for unobserved reasons may also create more content for the same reasons. We develop a methodology to control for the endogeneity of content with respect to the installed base of Internet users while controlling for a host of factors known to affect adoption. This approach also helps eliminate sources of spurious correlation that explain both content and adoption. To further reduce the possibility of spurious correlation, we include an extensive set of fixed effects in our estimation.

Our identification approach uses "large"-country content as an instrument for relevant content when estimating the effect of content on adoption for "small" countries, where we define *small* and *large* based on the number of potential adopters in a country. We argue and provide empirical evidence that content production by large countries is exogenous to Internet adoption in small countries. We assume that potential adopters value most content in their own language. Therefore, to identify content relevant to a country's potential adopters, we use the distribution of their language usage and measure content based on the storage capacity of computers hosting Internet content in those languages. Previous papers support our use of language to define Internet content relevance. ITU (1999) uses aggregate Web-traffic statistics to show that language determines the relevance of Internet content. Gandal (2006) shows that language usage heavily influences the languages of websites visited during individual-level browsing and provides evidence that English-language dominance in Internet content may continue based on bilingual users' online behavior.

Figure 1 Internet Penetration Across Countries in 2008



Source. World Development Indicators Database (2010a). Internet penetration (fraction of population with Internet access) for 197 countries is sorted from highest to lowest penetration. Not all country names are displayed due to a lack of space.

Using large-country content to instrument small-country-relevant content also helps eliminate sources of spurious correlation that might bias our results. Instrumenting sterilizes the estimates from unobserved factors that drive both content and adoption within each small country. Any remaining spurious correlation must be across small and large countries. We include an extensive set of fixed effects that makes this unlikely. Country fixed effects remove country-specific time-constant unobservables, and year fixed effects eliminate time-specific unobservables operating across the small and large countries. Finally, language fixed effects remove language-specific unobservables that drive adoption in small countries and content production by large countries with which they are instrumented.

Our results have implications for government policies that affect Internet content production. Governments directly create content, so much so that its quantity has raised concerns about effective archiving.[3] Much of this is generated as a part of regular government business, but some is specifically targeted at underserved languages. Qatar's government is developing digital archives of major Arabic texts to increase Arabic content (Pratap 2010). NGOs have also targeted underserved languages. One NGO, Canada's International Development Research Centre (IDRC), described content development efforts in Uganda as "increasingly important and valuable to the market" (Uganda Communications Commission 2005, p. 27). Arab countries working with NGOs have established rewards for high-quality Arabic content and encouraged collaboration between universities and research centers to produce content.[4] Other efforts are targeted at underserved populations. In the United States, the Federal Communications Commission announced in late 2011 a policy to promote job and education information relevant to households that had not yet adopted broadband (Stelter 2011).

Perhaps more important than governments' direct content creation are the indirect effects of their policies. Decisions on Internet technical standards have far-reaching effects on content creation. Originally architected in English, the Internet does not easily accommodate developing or finding content in

[3] See Carrell (2009). Also refer to *National Archives: The Challenge of Electronic Records Management, Before the Subcommittee on Government Management, Information, and Technology, Committee on Government Reform* (1999) (statement of L. Nye Stevens, Director, Federal Management and Workforce Issues, General Government Division).

[4] The Emirates Center for Strategic Studies and Research (2008).

languages using non-Latin characters. In response, the Internet Governance Forum approved a multiyear effort to allow non-Latin characters in website addresses.[5] Similarly, many Internet browsers will not properly display Arabic content because of a lack of agreement among Arab countries on a uniform format.[6] Our results also offer guidance to firms in evaluating their content investments. The estimates of adoption's sensitivity to content can be used to quantify the trade-off between marketing investments to increase usage directly and content investments to increase it indirectly. These are especially useful for firms relying on user-generated content to evaluate investments in customer acquisition.

2. Identification Strategy

Simply relating adoption and content will overstate content's effect because it will conflate content's effect on adoption with the feedback effect of adoption on content. At the same time, unobserved heterogeneity across countries may introduce spurious correlation. Our identification approach addresses both of these issues.

To disentangle content's effect on adoption, we use the subset of content created by large (in terms of number of language users but not necessarily geographic area) countries as an instrument for relevant content when estimating the effect of content on adoption for small countries only.[7]

Identification relies on the assumption that content creation by large countries is exogenous to adoption in small countries. Intuitively, we assume that the number of adopters in small countries is small enough that content creators in the large countries focus only on the number of adopters in the large countries.[8] That is, we assume that content created in large countries is relevant to and therefore consumed by those in small countries who share the same language, even though the latter are typically ignored by the content creators when choosing the profit-maximizing level of content.

We justify this assumption based on two related arguments. First, even if content creators can collect revenues from users in small countries, these represent such a small fraction that they do not affect content creation decisions. Second, it is frequently difficult to collect revenues from users in small countries because of legal impediments, high fixed costs of collecting subscription fees across country boundaries, and difficulty in targeting online advertising to these small groups. Relevancy of content to users in small countries has been reported to create a financial conundrum for major content providers such as Facebook and YouTube, which must provide the additional bandwidth to support these users despite difficulty in collecting revenues.[9] Besides these qualitative arguments, we provide quantitative evidence that this assumption holds when we present our data and results. At the same time, the instrument's inclusion restriction is met because relevant content consumed in small countries is affected by large-country content, given its ubiquity. We must omit the large countries from estimation to maintain exogeneity. Therefore, our results may not extrapolate to large countries; however, the combined population of our small countries is 2.0 billion.

We assume that an Internet user is most interested in content of her primary language and define small and large countries accordingly. We identify countries that comprise a large percentage of the worldwide users of a language as *large*. The remaining countries with small populations using that language we identify as *small*. Identification requires languages with a skewed distribution of users—a few countries represent most of the worldwide users, and many countries have a small percentage of the users. This provides a large number of observations while satisfying the exogeneity assumption.

For each small country, relevant content includes worldwide content (produced by both small and large countries) in the language(s) of its population. Since a small country's population may use a mixture of

[5] See Waters (2006). Methods of using non-Latin characters in website addresses emerged in 2003 but without standardization or official approval.

[6] The Emirates Center for Strategic Studies and Research (2008).

[7] We use the number of language users as a measure of potential adopters in that language. We do not use the actual number of adopters using the language because it is endogenous.

[8] Two previous papers use related identification schemes. Gowrisankaran and Stavins (2004) estimate network effects in adoption of the automated clearinghouse system (ACH) by clusters of U.S. banks. One method to isolate the network effect from a strong local preference for ACH is to examine the effect of adoption by small branches of large banks on the adoption decisions of rival banks in the same local markets. Identification relies on the fact that a bank must implement ACH at all its branches simultaneously. Shriver et al. (2013) examine the effect of online content production on the formation of social ties and Internet usage among surfers. They use exogenous wind speed changes as an instrument to break the feedback loop between social ties and production of user-generated content.

[9] Stone and Helft (2009) note that "Facebook is in a particularly difficult predicament. Seventy percent of its 200 million members live outside the United States, many in regions that do not contribute much to Facebook's bottom line," and quotes the chief executive officer of a San Diego-based video-sharing site who says of its users in Africa, Asia, Latin America, and Eastern Europe: "They sit and they watch and watch and watch. The problem is that they are eating up bandwidth, and it's very difficult to derive revenue from it."

languages, we construct a weighted-average measure of the relevant content based on the fraction using each language. For example, in Belgium, 38% of people speak Dutch, 33% French, 9% Walloon, 9% Flemish, 5% Limburgish, and 2% Italian as their primary language.[10] Relevant content for Belgium would equal 0.38 times the worldwide quantity of Dutch content plus 0.33 times the worldwide quantity of French content and so on. As a by-product, the language usage distributions provide significant cross-sectional variation in relevant content. The instrument for each small country is constructed analogously—a weighted-average of large-country content based on the language distribution of the small country's population.

Identification may also be affected by the presence of country-level unobservables that affect both content production and adoption but are separate from the indirect network feedback loop. If unaccounted for, these will induce correlation between relevant content and the error in our adoption equation and bias the coefficients on relevant content and the control variables. Our instrumenting approach combined with the large number of controls we include makes this unlikely. In our estimation we include year and country fixed effects in addition to a wide range of control variables. This means that any unobserved factors cannot be common to countries within the same year or result from country-specific characteristics. Thus, our estimation approach is robust to, among others, country-specific policies that promote adoption or content production, changes in standards that promote adoption or content production Internet-wide, and secular trends in adoption or content production due to factors such as technological changes in storage or transmission of data.

The content variable, once instrumented, will only be correlated with the adoption error if the unobserved factors drive both large-country content production and small-country adoption. Moreover, our instrumenting approach groups small and large countries based on the distributions of language usage across countries. Bias would require that adoption and content production be correlated within these groupings but in a way such that there is no common correlation across the small countries and no common correlation across the large countries because these would be absorbed by the year fixed effects. Importantly, these languages, and therefore the set of large countries within each group, differ for each small country according to its language distribution, which is exogenous with respect to Internet adoption and content. Since the grouping of a small adopting country with large content producers is mediated through language, a possible way for bias to enter is through language-specific unobservables. To address this, we show that our results are robust to adding language fixed effects.

[10] The remaining 4% use languages that each represents less than 1% of Belgium's population.

3. Econometric Model

We model the simultaneous determination of a country's content production in a language and adoption in that country by people using that language. The fraction of a language's users adopting the Internet in a country is a function of the worldwide content available in that language since Internet content is accessible anywhere.[11] Internet content produced by a country in a language is a function of the worldwide adopters using that language since the content is accessible worldwide.[12]

Let $i = 1, 2, \ldots, I$ index countries, $j = 1, 2, \ldots, J$ languages, and $t = 1, 2, \ldots, T$ years. We model adoption and content production according to the simultaneous system of stochastic equations:

$$\frac{\text{Adopters}_{ijt}}{\text{Users}_{ij}} = \beta^A X_{it}^A + \lambda^A Z_i^A + \rho_t^A + \delta_i^A + \gamma^A \sum_{k=1}^{I} \text{Content}_{kjt} + \tilde{\varepsilon}_{ijt}^A, \quad (1a)$$

$$\text{Content}_{ijt} = \beta^C X_{it}^C + \lambda^C Z_i^C + \rho_t^C + \delta_i^C + \gamma^C \sum_{k=1}^{I} \text{Adopters}_{kjt} + \tilde{\varepsilon}_{ijt}^C, \quad (1b)$$

where Adopters_{ijt} is the number of Internet adopters who use language j in country i at time t, Users_{ij} is the number of users of language j in country i that does not vary over time in our data, and Content_{ijt} is the content available in language j at time t produced by country i. X_{it}^A and X_{it}^C include possibly overlapping sets of time-varying factors affecting Internet adoption and content, whereas Z_i^A and Z_i^C are the same for time-constant factors.

The parameters to be estimated are $\{\beta^A, \lambda^A, \gamma^A, \beta^C, \lambda^C, \gamma^C\}$. The latent year effects, ρ_t^A and ρ_t^C, capture unobserved time-specific factors affecting adoption and content, respectively. The latent country effects, δ_i^A and δ_i^C, are time-invariant random variables that capture unobserved factors affecting adoption and content, respectively. We discuss the statistical properties of these fixed effects below. The error terms,

[11] We control for government restrictions on Internet access in our estimation.

[12] As explained below, the content is not necessarily hosted on a computer located physically within the country.

$\tilde{\varepsilon}^A_{ijt}$ and $\tilde{\varepsilon}^C_{ijt}$, are independently and identically distributed across countries, languages, and time periods. We expect $\gamma^A, \gamma^C > 0$. This specification assumes that content's effect on adoption is the same across languages. Although in theory we could allow the effect to vary by language, in practice there are insufficient data to identify this.

If X^A_{it} and X^C_{it} each contain at least one variable not contained in the other, a system method of estimation for (1a) and (1b) may be feasible. Unfortunately, we do not have available any variables thought to affect content but not adoption. Instead, we estimate (1a) using limited-information estimation methods and use Equation (1b) to inform our search for an appropriate instrument for the content variable in Equation (1a).

For a set of the most frequently used languages, J_F, we divide countries into large ($i \in I_L$) and small ($i \in I_S$) based on the number of language users with $I = \{I_S, I_L\}$. Our identification assumption is that content production by large countries is unaffected by adoption in small countries. More formally, $\sum_{k\in I_L} \text{Adopters}_{kjt} \approx \sum_{k=1}^{I} \text{Adopters}_{kjt}$ so that

$$\text{Content}_{ijt} = \beta^C X^C_{it} + \lambda^C Z^C_i + \rho^C_t + \delta^C_i + \gamma^C \sum_{k\in I_L} \text{Adopters}_{kjt} + \tilde{\varepsilon}^C_{ijt} \quad \forall i \in I_L, \forall j \in J_F. \tag{1b'}$$

If Equation (1b′) holds, then $\sum_{k\in I_L} \text{Content}_{kjt}$ (large-country content) is a valid instrument for $\sum_{k=1}^{I} \text{Content}_{kjt}$ (worldwide relevant content) in Equation (1a) estimated on the set of small countries:

$$\frac{\text{Adopters}_{ijt}}{\text{Users}_{ij}} = \beta^A X^A_{it} + \lambda^A Z^A_i + \rho^A_t + \delta^A_i + \gamma^A \sum_{k=1}^{I} \text{Content}_{kjt} + \tilde{\varepsilon}^A_{ijt} \quad \forall i \in I_S. \tag{1a'}$$

To preserve degrees of freedom, we use only the world's most pervasive languages to construct the instrument. Enlarging this set involves a trade-off between decreasing available data and increasing the instrument's power. Including an additional language reduces the available data because large content producers for that language must be excluded to maintain the exogeneity assumption. On the other hand, it increases the instrument's power since more languages mean the instrument is more highly correlated with the small countries' consumed content. In §5 we empirically assess the exogeneity and relevance conditions for our instrument. Our choice of languages for the instrument is discussed in §4.

Since we observe only the aggregate number of Internet adopters in each country, we transform Equation (1a′) into one which we can estimate. Multiplying through by the number of users of language j and then summing across all languages we obtain

$$\sum_{j=1}^{J} \text{Adopters}_{ijt} = (\beta^A X^A_{it} + \lambda^A Z^A_i + \rho^A_t) \sum_{j=1}^{J} \text{Users}_{ij} + \sum_{j=1}^{J} [\text{Users}_{ij}(\delta^A_i + \tilde{\varepsilon}^A_{ijt})] + \gamma^A \sum_{j=1}^{J} \left[\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt} \right] \quad \forall i \in I_S. \tag{2}$$

Since $\sum_{j=1}^{J} \text{Users}_{ij} = \text{Population}_i$,

$$\frac{\sum_{j=1}^{J} \text{Adopters}_{ijt}}{\sum_{j=1}^{J} \text{Users}_{ij}} = \sum_{j=1}^{J} \frac{\text{Adopters}_{ijt}}{\text{Population}_i} = \text{Penetration}_{it}, \tag{3}$$

where Penetration_{it} is the fraction of country i's population that have adopted the Internet at time t, which we observe. Dividing both sides of Equation (2) by Population_i, we get[13]

$$\text{Penetration}_{it} = \beta^A X^A_{it} + \lambda^A Z^A_i + \rho^A_t + \delta^A_i + \gamma^A \frac{\sum_{j=1}^{J} [\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt}]}{\text{Population}_i} + \varepsilon^A_{it} \quad \forall i \in I_S. \tag{4}$$

We call the weighted-average measure of content in Equation (4) the relevant content for small country i in year t:

$$\text{relcon}_{it} = \frac{\sum_{j=1}^{J} [\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt}]}{\text{Population}_i}, \quad i \in I_S. \tag{5}$$

This includes content produced worldwide in each of the languages used within country i weighted by the proportion of the population using that language. Worldwide content includes content produced in country i as well as content in relevant languages produced outside of it. The instrument for relevant content is defined similarly but includes only content produced by large countries:

$$\text{instrument}_{it} = \frac{\sum_{j=1}^{J} [\text{Users}_{ij} \sum_{k\in I_L} \text{Content}_{kjt}]}{\sum_{j=1}^{J} \text{Users}_{ij}}, \quad i \in I_S. \tag{6}$$

We denote a country-time period unobservable that affects adoption in country i at time t by ε^A_{it}. We

[13] Transforming Equation (1a′) into Equation (4) introduces country-level heteroskedasticity since the distribution of languages varies across countries. This is difficult to accommodate in the Hausman–Taylor estimates (Hausman and Taylor 1981). However, our fixed-effects estimates in Table 5, which are consistent, are robust to general forms of heteroskedasticity and yield similar results to the Hausman–Taylor estimates.

distinguish, on a priori grounds, columns of X and Z that are asymptotically uncorrelated with δ_i^A from those that are not so that our assumptions about the random terms in the model are

$$E(\varepsilon_{it}^A) = E(\delta_i^A \mid X_{1it}^A, Z_{1it}^A) = 0$$

but

$$\begin{aligned} E(\delta_i^A \mid X_{2it}^A, Z_{2it}^A) &\neq 0, \\ \mathrm{Var}(\delta_i^A \mid X_{1it}^A, Z_{1it}^A, X_{2it}^A, Z_{2it}^A) &= \sigma_\delta^2, \\ \mathrm{Cov}(\varepsilon_{it}^A, \delta_i^A \mid X_{1it}^A, Z_{1it}^A, X_{2it}^A, Z_{2it}^A) &= 0, \\ \mathrm{Var}(\varepsilon_{it}^A + \delta_i^A \mid X_{1it}^A, Z_{1it}^A, X_{2it}^A, Z_{2it}^A) &= \sigma^2 = \sigma_\varepsilon^2 + \sigma_\delta^2, \\ \mathrm{Corr}(\varepsilon_{it}^A + \delta_i^A, \varepsilon_{is}^A + \delta_i^A \mid X_{1it}^A, Z_{1it}^A, X_{2it}^A, Z_{2it}^A) & \\ = \rho = \sigma_\delta^2/\sigma^2. & \end{aligned} \tag{7}$$

This error structure allows the Hausman and Taylor (1981; HT hereafter) estimator. HT refer to X_{1it}^A as time-varying exogenous, X_{2it}^A as time-varying endogenous, Z_{1it}^A as time-invariant exogenous, and Z_{2it}^A as time-invariant endogenous variables. We discuss these classifications and justify our use of the HT estimator vis-à-vis a fixed-effects and random-effects estimator in §5.

The exogeneity of large-country content suggests a simpler estimation approach: regress small-country adoption on large-country content. However, this has an important practical limitation. Large countries for all included languages must be dropped from the analysis to meet the exogeneity condition. To maintain a sufficient sample size, not all languages can be included, and therefore it is not possible to include all external (outside the country) content for each small country. This introduces an omitted variable bias, the sign of which depends on the correlation between the included and excluded external content net of the effect of the control variables. Although the sign of this correlation is theoretically indeterminate, it is likely negative since more included external content for a small country implies less excluded content. Our instrumenting strategy frees us from producing inconsistent estimates because now we can include all content (internal and external) for each small country while solving the endogeneity problem. We still must exclude large countries from the analysis, but this can be minimized because we need only enough included languages to adequately satisfy the instrument's inclusion restriction.

Ideally we would estimate adoption's effect of content using a similar strategy—use adoption rates in large countries as an instrument for small-country adoption rates when predicting small-country content production. This is not possible for two reasons—one methodological and the other practical. Large-country adoption rates as an instrument fails the exclusion restriction. Since content is ubiquitous, large and small country content are substitutes. We also face a practical problem: we do not observe language-specific adoption rates. Therefore, only time-series variation would identify adoption's effect on content.

4. Data

Our sample includes data on 176 small countries and 31 large countries from 1998 to 2004.[14] Table 1 contains summary statistics on the main variables. Online Appendix B contains more details on variables and their sources. (Online appendices available at http://english.ckgsb.edu.cn/faculty_content/brian-viard.)

Internet Users: Our dependent variable is the fraction ofcountry i's population with Internet access at time t (see Figure 2 for 2004 small-country adoption rates in the sample). The ITU collects these data and does not distinguish speeds or modes of access. During our sample years, virtually all access was through one of three modes: narrowband (or dial-up) access through a phone line, broadband (or digital subscriber line) access through a phone line, and broadband access through cable lines. The ITU data measure all Internet users regardless of location.[15] Unfortunately, the data do not allow us to control for access speed since content may drive adoption of higher-quality access. During our sample period, most relevant content is textual minimizing this concern;[16] however, if nontextual content were significant, it would bias our

[14] Online Appendix A contains a list of the small countries. These include 12 non-self-governing territories: overseas territories (Bermuda), overseas regions (French Guiana, Guadeloupe, Martinique), overseas collectivities (French Polynesia, Mayotte), sui generic collectivities (New Caledonia), special administrative regions (Hong Kong, Macao), disputed territories (Palestinian West Bank and Gaza), unincorporated organized commonwealths (Puerto Rico), overseas departments (Reunion), and unincorporated organized territories (Guam, U.S. Virgin Islands). We include these because we believe their social and economic conditions differ substantially enough from their governing countries that they represent independent observations. Content measures are not available for Hong Kong, Macao, and Mayotte, so they do not identify the effect of content.

[15] ITU's data distinguish between "estimated Internet users" and "Internet subscribers." Users of Internet cafés, for example, would be included in the former, which is our variable, but not in the latter.

[16] Of file space for publicly available Internet data in 2003, text-related files (Excel, text, Word, Powerpoint, PDF, PHP, and HTM/HTML) represented 41%, image data 23%, and audio and movie files 7%. The remaining 29% were of unknown type or executable files (Lyman and Varian 2003). Since images may also contain text a lower bound for text files as a fraction of known file types is 58% and of all (classifiable and unknown) is 41%. Since

Table 1 Descriptive Statistics, 176 Small Countries, 1998–2004

Variable	*N*	Mean	S.D.	Min	Max
Time-varying covariates					
Internet Users (per 100 people)	1,169	0.099	0.143	0.000	0.755
Per-Capita GDP (US$ thousands)	958	8.392	9.323	0.450	60.249
Telephone Infrastructure	776	0.240	0.209	0.000	0.908
Log Normalized Internet Price (1998)	25	−5.935	0.578	−6.644	−4.496
Log Normalized Internet Price (2000)	25	−6.457	0.601	−7.167	−4.981
Log Normalized Internet Price (2001)	111	−4.562	1.517	−7.279	−1.526
Fraction School Enrollment	653	0.872	0.160	0.278	1.000
Civil Liberties Index	1,055	4.495	1.794	1.000	7.000
Time-constant covariates					
Literacy Rate	753	0.795	0.197	0.240	1.000
Gini Coefficient	688	0.406	0.106	0.247	0.743
Age Below 20	1,078	0.424	0.117	0.196	0.605
Age 20 *to* 39	1,078	0.302	0.034	0.244	0.480
Age 40 *to* 64	1,078	0.208	0.071	0.110	0.341
Age Above 64	1,078	0.066	0.044	0.011	0.182
Fraction Urban Population	1,168	0.546	0.241	0.077	1.000
Household Size	678	4.525	1.413	2.000	10.500
Content measures					
Relevant Content (millions of relevant hosts)	1,114	3.047	13.442	0.000	172.503
Own Content (millions of hosts)	1,114	0.081	0.343	0.000	5.434
Large Country Content (millions of hosts)	926	22.525	45.152	0.000	206.814
Language Herfindahl	926	0.868	0.197	0.378	1.000
Supplementary variables					
Log[*Gateway Capacity*] (gigabits per second)	1,169	0.331	1.154	0.000	8.144
IP Protection	321	5.090	1.973	0.300	8.600
Price instruments					
Government Tax Receipts (% of GDP)	519	16.908	7.153	0.958	43.705
Corporate Tax Rate (%)	499	27.783	9.609	0.000	54.000
Telephone Employees (per 1,000 fixed lines)	922	12.000	20.789	0.068	175.385

Note. See Online Appendix B for a description of the variables and their sources.

estimate of content's effect downward. There would tend to be less content created in languages whose users have slow connections. Since we allocate content based on language usage, we will tend to overallocate content to language users with slow connections and underallocate to those with fast. Since language users with slow connections actually have less content available, they will have lower adoption rates, biasing our results downward.

Content: We measure content by the number of host computers connected to the Internet in each year for each country. Host computers contain accessible content, and the total quantity of content is proportional to the number of computers.[17] This does not measure content quality; however, for our estimates it need only be the case that quality is proportional to storage capacity across different languages. We do not directly observe the language of these computers' content but rather infer it from the registration country, as explained below.

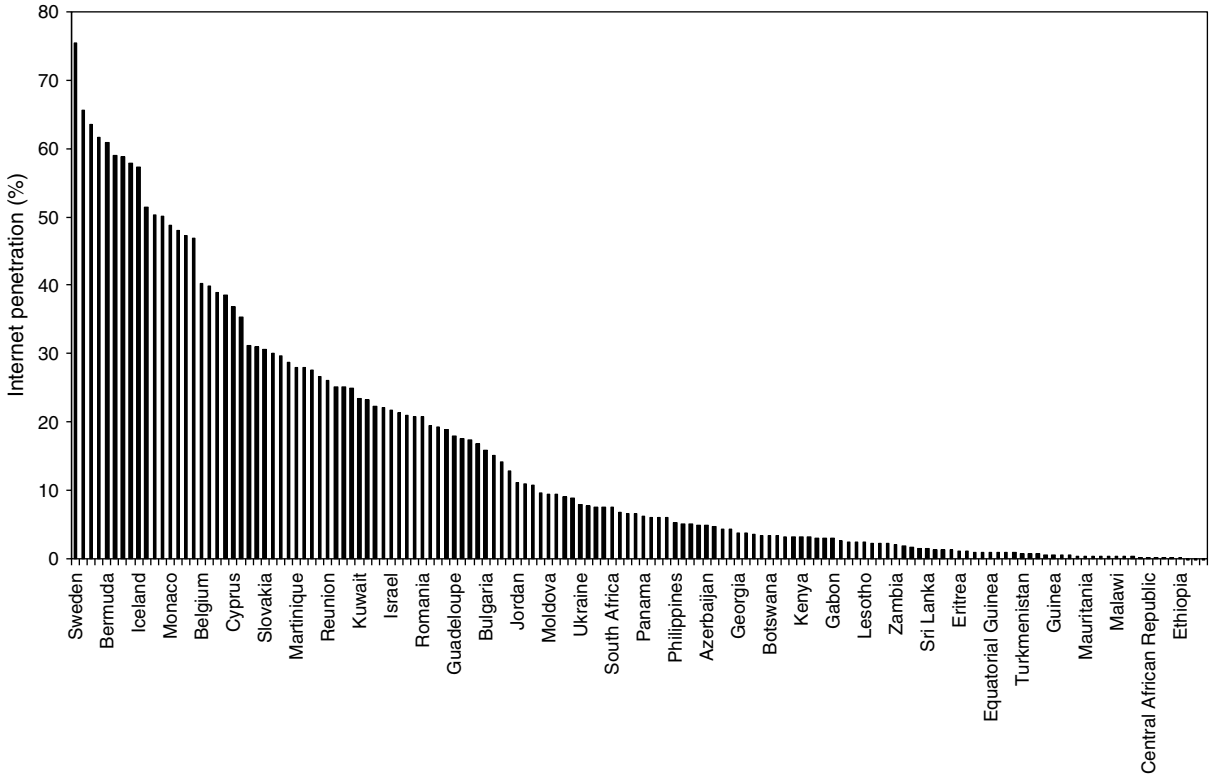
Internet host numbers are based on data from the Internet Systems Consortium, Inc. (ISC). During our sample period, ISC took an annual census of host computers connected to the Internet. ISC maintained the same sampling procedure throughout our sample years, ensuring comparability. However, since computer storage capacity may change over time, we include year dummies in all our estimates and also estimate separate yearly effects as a robustness check.

The ISC data also allocate each host to a country, which allows us to allocate them to languages. Assignment of a host to a country does not necessarily

2003 is near the end of our sample period, text is likely an even higher fraction in earlier years as faster access speeds over time have led to increased use of images and video. These data are for the "surface" Web. There is a large amount of data in the "deep" Web, but most of it is not publicly accessible during our sample period either because it is behind corporate firewalls or because is not indexed by search engines (see Bergman 2001). Our measure of hosts includes the surface but not deep Web. For a later period, Bohn and Short (2009) estimate that in 2008 Internet text comprised 178 hours of usage for the average Internet user while video comprised 2 hours. In terms of storage, they estimate that in 2008 there were 8.0 exabytes of Internet text compared with 0.9 of video. Video would play an even smaller role during our sample period when Internet connections were much slower.

[17] Host computers are connected to the Internet and contain accessible content. There are many more computers connected indirectly to the Internet through local area networks (intranets). Computers on an intranet can access the Internet but cannot host content.

Figure 2 Internet Penetrations in Sample Small Countries in 2004



Source. World Development Indicators Database (2010a). Internet penetration (fraction of population with Internet access) for 176 sample small countries sorted from highest to lowest penetration. Not all country names displayed because of a lack of space.

mean that the computer is physically located within the country; however, our estimation requires only that the computer contains content created within that country. The rules for assigning hosts make this likely. Although the rules differ slightly across countries, most require a local presence requirement such as citizenship, resident address, or local administrative contact.[18]

Since more than one language is used in most countries, we allocate the total hosts to each language based on the fraction of the country's population using each language.[19] This assumes that all content is language-specific. As discussed earlier, Internet content during our sample period is primarily textual; moreover, much nontextual content is language-specific as images often contain text and videos language-specific dialogue. Nonetheless, Online Appendix D shows that our model accommodates non-language-specific content assuming that each country's language-specific content is produced in the same proportion as language users in that country and that language-specific and non-language-specific content affect adoption equally on the margin. The former assumption is the same as that required when we allocate hosts to languages within countries, so only the latter is potentially restrictive. If adoption is more responsive to language-specific than non-language-specific content, then our estimates will understate content's effect. If the opposite is true, we will overstate it. Combining this measure of content for each country in each year with the language data, we construct the relevant content and instrument for each country-year pair based on Equations (5) and (6).

Prior to 1999, if ITU could not find an independent estimate of Internet users in a country, it based its estimate on a multiple of the number of host computers in the country, which would pose problems for our estimation. After this, ITU used only surveys to quantify users (Minges 1999). To determine whether this is a problem, we reestimated our baseline estimates dropping the year 1998 data. The results were virtually identical.

Language Users: Our source for language data is *Ethnologue* (Gordon 2005), which offers the most comprehensive catalogue of the world's languages (for linguistic reviews, see Campbell and Grondona 2008,

[18] Online Appendix C contains more detail on how ISC collects the host data and allocates it to countries.

[19] This is not a major concern for our instrument because the populations of virtually all of the large countries are dominated by a single language. We assume that all the host computers in large countries pertain to that country's dominant language.

Hammarström 2005, Paolillo and Das 2006). *Ethnologue* provides detailed and comprehensive estimates of first-language speakers of each language by country.[20] Its data are not complete enough to estimate using second-language speakers.

Since Internet content was primarily textual during our sample period, we ideally would use the numbers of literate users of each language to create our relevant content measure. Because we do not observe language-specific literacy rates by country, we use the numbers of speakers of each language in a country and include the country's overall literacy rate as a control variable. We combine spoken dialects whose users employ the same written language. For example, we combine speakers of the many Chinese dialects that all utilize simplified Chinese for writing. *Ethnologue* is a thorough accounting of the world's languages. As a result, some are spoken by very few people. To make data entry manageable, for each country we added languages in descending order of the most spoken and kept adding until the next language would contribute less than 1% of the country's population or all languages were exhausted. Across all countries, this comprised 811 languages or spoken dialects.

To choose instrument languages, we apply the two criteria discussed in §2: the language is spoken in many countries and its usage distribution is skewed with a few countries making up a significant fraction of total users. Based on these criteria, we use 14 languages to construct our instrument:[21]

$$J_F = \left\{ \begin{array}{l} \text{Chinese, Spanish, English, Hindi,} \\ \text{Portuguese, Russian, Japanese,} \\ \text{German, French, Hausa, Zulu,} \\ \text{Nyanja, Pulaar, Pular} \end{array} \right\}. \quad (8)$$

The first 8 are among the top 10 most spoken languages in the world based on *Ethnologue*.[22] French is the 17th most spoken language. The usage of the languages between the 10th and 17th (Javanese, Telugu, Marathi, Vietnamese, Korean, and Tamil) is either not widespread or fairly uniformly distributed across countries. The last five languages were chosen to include African languages subject to meeting our two criteria. Each of these five is spoken in at least four countries, and the two most populous countries using the language represent at least 82% of total users. Column 3 of Table 2 shows the total number of users for the 14 languages used to construct our instrument: 2.7 billion people, or 44% of the 6.1 billion world population in 2000.

We use the number of potential adopters (i.e., population using a language) in a country to identify large and small countries. We choose the large countries (the set I_L) for our instrument (Equation (6)) by the following procedure: For each language, sort the countries in descending order by the number of users. Starting at the top, add countries until the last country added brings us above 75% of worldwide users. There were three exceptions when we kept adding above 75%.[23] Column 5 of Table 2 shows the 31 large countries chosen, and columns 6–8 show the number of users in the large countries and as a percentage of worldwide users. Identification relies on the column 8 percentages being large so that these countries are unaffected by adoption in small countries (the 31 large countries will be excluded from our analysis). The large countries represent 80% or more of the world's users of each language.

The last three columns of Table 2 show data for each language's largest small country. Columns 10 and 11 show the number of users in the largest small country and as a fraction of worldwide users. Identification depends on the column 11 percentages being small so that Internet adoption in these countries does not affect large countries' content production. The largest small countries represent eight percent or fewer of the world's users for each language. The percentages for all other small countries are below this.

Control Variables: We include as many control variables from previous studies of Internet adoption as possible so as to isolate content's effect. Therefore, subject to preserving enough degrees of freedom to discern content's effect, our goal is to maximize the variance explained by our regressions rather than the significance of individual coefficients. To identify control variables, we rely on previous papers estimating cross-country Internet adoption.

Time-varying factors include measures of wealth, education, infrastructure, cost of access, and freedom

[20] *Ethnologue* does not distinguish between native and primary first-language speakers. This should be considered in interpreting our results.

[21] To check the robustness to the choice of instrument languages, we randomly divided these into two groups of seven languages each and reestimated our baseline results in column 3 of Table 5, using each of these two sets to construct the instrument. The coefficient on relevant content for both estimates was within 1.6% of our baseline estimate, and the coefficients remained significant at below the 0.01% level.

[22] Arabic (fourth) and Bengali (seventh) were not included because their usage was not skewed enough.

[23] The three exceptions were because there was an obvious large drop between two countries. For Chinese, mainland China alone would bring us above 75%, but we added Taiwan because it had 5.2 times as many Chinese speakers as the next largest country, Malaysia. For English, the United States and the United Kingdom alone would bring us above 75%, but we added Canada and Australia because Australia was 4.9 times as large as the next largest country, New Zealand. For Portuguese, Brazil alone would bring us above 75%, but we added Portugal because it is 15.6 times as large as the next largest country, Paraguay.

Table 2 Profiles of Large and Small Countries for Included Languages

		Worldwide		Large countries				Largest small country		
1 Ranking[a]	2 Language	3 Total no. of users (millions)	4 Total content (1,000s of hosts)[b]	5 Country	6 No. of users (millions)	7 Total no. of users (millions)	8 % of worldwide	9 Country	10 Total no. of users (millions)	11 % of worldwide
1	Chinese	1,204.76	2,674.68	China	1,171.05	1,193.74	99.1	Malaysia	4.39	0.36
				Taiwan	22.69					
2	Spanish	322.30	11,100.00	Mexico	86.21	256.16	79.5	Dominican Republic	6.89	2.14
				Colombia	34.00					
				Argentina	33.00					
				Spain	28.17					
				Venezuela	21.48					
				Peru	20.00					
				Chile	13.80					
				Cuba	10.00					
				Ecuador	9.50					
3	English	309.35	89,300.00	United States	210.00	297.78	96.3	New Zealand	3.21	1.04
				United Kingdom	55.00					
				Canada	17.10					
				Australia	15.68					
5	Hindi	180.77	37.01	India	180.77	180.77	100.0	Nepal	0.11	0.06
6	Portuguese	177.46	2,538.54	Brazil	163.15	173.15	97.6	Paraguay	0.64	0.36
				Portugal	10.00					
8	Russian	145.03	677.62	Russia	145.03	145.03	100.0	Ukraine	11.34	7.82
9	Japanese	122.43	8,370.64	Japan	122.43	122.43	100.0	Singapore	0.02	0.02
10	German	95.39	5,289.15	Germany	75.30	82.80	86.8	Kazakhstan	0.96	1.00
				Austria	7.50					
17	French	64.86	2,953.88	France	64.86	64.86	100.0	Belgium	4.00	6.17
	Hausa	24.16	0.26	Nigeria	18.53	23.53	97.4	Chad	0.10	0.41
				Niger	5.00					
	Zulu	9.56	60.72	South Africa	9.20	9.20	96.2	Lesotho	0.25	2.59
	Nyanja	9.35	0.35	Malawi	7.00	8.60	92.0	Mozambique	0.50	5.32
				Zambia	1.60					
	Pulaar	3.24	0.36	Senegal	2.39	2.65	81.7	Guinea-Bissau	0.25	7.56
				Gambia	0.26					
	Pular	2.92	0.12	Guinea	2.55	2.55	87.4	Sierra Leone	0.18	6.12
		2,671.58	123,003.32		2,563.25	2,563.25	95.9		32.81	1.23
	All languages	6,070.50[c]	138,648.22							

[a]Most spoken languages by first-language speakers according to Gordon (2005). If blank, it is not ranked.
[b]Average number of hosts across six years of data.
[c]Based on year 2000 data from United Nations (2004).

of expression. Per-capita GDP measures a country's wealth, which we expect to positively affect adoption. Internet access is likely more highly valued in countries with more educated populations, so we include the fraction of eligible children enrolled in primary school. We include the fraction of the population with fixed phone lines to measure telecommunications infrastructure quality. Although there are other ways to access the Internet during this time, these were either rare (satellite and Wi-fi) or likely highly correlated with telephone infrastructure (cable television). We include a measure of citizens' freedom to engage in expression, based on a measure used by NGO Freedom House, to control for the degree of government restrictions on content access. The measure ranges from 1 to 7, with 7 being the most free.[24]

[24] Freedom House defines 7 as the least free. We reverse the order for ease in interpretation.

We include average monthly Internet access prices normalized by per-capita GDP to control for cost of access. Unfortunately, prices are available only for three years (1998, 2000, and 2001) and not for all countries. Since each year's data measure a different type and amount of usage, we cannot pool them across years. Internet access prices and adoption may both be higher in countries with higher unobserved access quality, which would bias the price coefficient toward 0. We therefore instrument with variables affecting price but affecting adoption only through price. Since we include fixed effects in our final estimation, we use three time-varying instruments. Corporate tax rates directly affect the cost of providing Internet access. The ratio of government tax receipts to GDP captures the regulatory atmosphere in which the Internet service providers operate. The number of telephone employees per fixed line proxies for the

productivity of or labor-capital ratio in the telecommunications industry.

We also include a number of time-constant controls. These are "time-constant" in that we observe only one year of data, although they likely change slowly. The Gini coefficient of income controls for the wealth distribution within a country, and we expect higher inequality (higher Gini coefficient) to negatively affect adoption. The fraction of a country's population living in urban areas measures infrastructure, demand, or both. More densely populated areas can be served more cheaply on a per-customer basis than more dispersed. At the same time, it may be that Internet access demand by urban residents differs from that by rural residents. Since familiarity with the Internet is likely age dependent, we control for the country's age distribution using the population fraction in four age brackets. Average household size allows for potential economies of scale in adopting Internet access within households. Literacy rate controls for the ability of a country's population to read content.

These control variables are drawn from a variety of papers. Wallsten (2007) explains broadband penetration for Organisation for Economic Co-operation and Development (OECD) countries, and Wallsten (2005) assesses the impact of regulation on developing countries' Internet adoption rates and prices. Ford et al. (2008) produce a broadband performance index for OECD countries based on the predicted values from an adoption regression. Chinn and Fairlie (2006) explain cross-country Internet and computer adoption rates.

We know of only three papers that include language to explain Internet adoption, and all include only a single language (English) and do not control for endogeneity. Hargittai (1999) explains Internet adoption by OECD countries and includes English-language usage as an explanatory variable because of its importance in the media and computing fields. The effect of language is not significant. Kiiski and Pohjola (2002) estimate a diffusion model of Internet adoption by OECD countries and include English-language proficiency for the same reason but estimate a negative effect. Wunnava and Leiter (2009) also estimate a diffusion model of Internet adoption but with more countries. They include English-language proficiency to measure the accessibility of English-language content. They find a positive and significant effect.

5. Main Results

Content availability has a positive and statistically significant effect on adoption. Since content is not directly measurable, there is no single right way to quantify its effect. We provide several ways and find an important role regardless. First, we compute an elasticity of adoption with respect to content and compare it to other markets. Second, we compare content's effect to that of other adoption determinants. Third, we quantify the additional adoption that results from the annual average content production in our sample.

Table 3 **Adoption/Content Correlation Matrix for Sample Countries, 1998–2004** ($N = 779$)

	Internet users	Own content	Relevant content
Own content	0.532 (0.000)		
Relevant content	0.348 (0.000)	0.033 (0.366)	
Large country content (instrument)	0.293 (0.000)	0.083 (0.021)	0.517 (0.000)

Note. Significance levels are in parentheses.

Before we formally estimate the effect of content on adoption, we examine the correlations between adoption and the various measures of content including the instrument. These are shown in Table 3. A country's own content is highly positively correlated with its own Internet adoption, consistent with a two-sided market. Relevant content is also highly positively correlated with adoption, consistent with the ubiquity of Internet content (this content is produced both within and outside the country but in the languages of its population). However, relevant content and own content are not significantly correlated. This is consistent with a country's own content production being determined by two-sided market effects within the country, whereas relevant content is determined by two-sided market effects across many countries sharing common languages.

Finally, large-country content (the instrument) is highly correlated with both adoption and relevant content but is much less correlated with a country's own content. This is consistent with large-country content influencing a country's content production only indirectly through adoption. The low correlation between large-country and own-country content is informal evidence that the exclusion restriction is met, whereas the high correlation between large-country and relevant content is informal evidence of its relevance. We provide more formal tests of the instrument's validity below.

5.1. First-Stage Results

Columns 1–3 of Table 4 show the first-stage results for Internet access prices. Given the small number of observations in each year, the coefficients are noisy. The number of telephone employees has a positive effect and is significant in two of the three years, consistent with higher prices from lower productivity.

Table 4 First-Stage Regressions for Internet Access Prices and Relevant Content

	1 1998 log prices	2 2000 log prices	3 2001 log prices	4 Relevant content	5 Relevant content
Intercept	−4.5397***	−5.1697***	−4.3932***	1.2422***	0.5514
	(0.5632)	(0.7015)	(0.2037)	(0.3084)	(0.4485)
Large country content				0.0266	0.1379**
				(0.0250)	(0.0566)
(*Large country content*)^2				0.0007*	
				(0.0004)	
Government tax receipts (% of GDP)	−0.0359*	−0.0438**	−0.0601***		
	(0.0193)	(0.0216)	(0.0124)		
Corporate tax rate (%)	−0.0219	−0.0138	0.0030		
	(0.0144)	(0.0195)	(0.0707)		
Telephone employees (per fixed line)	0.0692**	0.1493**	0.1371		
	0.0345	0.0588	0.1162		
R^2	0.3000	0.3909	0.2563	0.1946	0.1803
N	44	44	134	926	926

Notes. Standard errors are in parentheses. Standard errors for relevant content regressions are clustered at the country level. Dummy variables for missing values are included for all variables in price regressions.
*Indicates 10% significance; **indicates 5% significance; ***inidcates 1% significance.

Government tax receipts has a significantly negative effect in all three years, consistent with greater subsidies for Internet access in countries with greater government revenues.

We allow for a flexible functional form in the first-stage regression of relevant content. We use a second-order, Taylor-series expansion of the instrument as shown in column 4 of Table 4.[25] Both the linear and quadratic terms are positive, although only the quadratic term is significant. Specification tests indicate the exclusion restriction and the relevance condition are likely met. A Hausman specification test of exogeneity yields a test statistic of 47.1 compared to a critical value of 0.1, and the *F*-value for our first-stage regression is 111, which greatly exceeds the critical value of 10 specified in Staiger and Stock (1997) to rule out weak instruments. To test the sensitivity of our results to the quadratic functional form, we reestimated using the linear first-stage specification shown in column 5 of Table 4. Large-country content has a significantly positive effect on relevant content, and very similar second-stage results were obtained.

5.2. Panel Data Results

Although we control for many factors thought to affect Internet adoption, we also include country fixed effects to control for country-level unobservables. A within-groups estimate of Equation (4) provides consistent estimates of the time-varying variables in the model including content. To compare content's effect to that of as many variables as possible, we also include time-constant variables. Since we believe that we have plausibly exogenous time-invariant factors available, we use an HT estimator.

Of the time-varying variables, telephone infrastructure and civil liberties are likely endogenous in the HT sense (i.e., correlated with country-level unobservables). A country that invests heavily in technology (more than commensurate with its per-capita GDP) likely has high Internet adoption and high fixed-phone line penetration. A society with greater unobserved preferences for Internet access may also have a greater preference for civil liberties. Price and relevant content are exogenous by design. Neither per-capita GDP nor school enrollment is likely affected by unobserved preferences for Internet adoption in the short run.

Of the time-invariant variables, all are likely exogenous in the HT sense except for literacy. The income distribution, age distribution, average household size, and urban density are not likely affected by unobserved preferences for Internet adoption. Measuring literacy is subjective because there are no standard criteria across countries. Countries with low literacy rates may report artificially high rates and also have a low unobserved preference for Internet access.

Column 1 of Table 5 shows the second-stage results of a random-effects specification with standard errors clustered by country and robust to general heteroskedasticity. The table is divided into four panels classifying the variables as time varying versus time invariant and exogenous versus endogenous. We will not discuss the results in detail since this is rejected in favor of a fixed-effects specification, but relevant content has a highly statistically significant effect (below the 0.01% level).

[25] A cubic term was not significant.

Table 5 Effect of Content on Internet Adoption for All Sample Countries, 1998–2004, Second-Stage, Panel Data Estimates

	1 RE	2 FE	3 HT-GLS	4 HT-GLS
Time-varying exogenous				
Per-Capita GDP	0.0101***	0.0115***	0.0104***	0.0105***
	(0.0008)	(0.0014)	(0.0010)	(0.0010)
Log Normalized Internet Price (1998)	0.0131	0.0033	0.0063	0.0108
	(0.0423)	(0.0416)	(0.0417)	(0.0416)
Log Normalized Internet Price (2000)	−0.0363	−0.0447	−0.0413	−0.0399
	(0.0290)	(0.0286)	(0.0286)	(0.0285)
Log Normalized Internet Price (2001)	−0.0012	−0.0016	−0.0013	−0.0014
	(0.0062)	(0.0061)	(0.0061)	(0.0061)
Fraction School Enrollment	0.0051	0.0022	−0.0006	−0.0023
	(0.0218)	(0.0224)	(0.0222)	(0.0222)
Relevant Content	0.0015***	0.0016***	0.0015***	0.0050***
	(0.0004)	(0.0004)	(0.0004)	(0.0011)
(*Language HHI Above Mean*) × *Relevant Content*				−0.0039***
				(0.0011)
Time-varying endogenous				
Telephone Infrastructure	−0.1579***	−0.1723***	−0.1745***	−0.1716***
	(0.0167)	(0.0171)	(0.0169)	(0.0168)
Civil Liberties Index	0.0028	0.0002	−0.0007	−0.0001
	(0.0030)	(0.0039)	(0.0038)	(0.0038)
Time-invariant exogenous				
Gini Coefficient	−0.0247		0.0015	−0.0166
	(0.0790)		(0.1020)	(0.1013)
Fraction Urban Population	0.0582*		0.1018**	0.0920*
	(0.0342)		(0.0499)	(0.0489)
Age Below 20	0.1609		0.9908	0.9698
	(0.4155)		(0.6393)	(0.6319)
Age 20 *to* 39	0.0720		0.2976	0.2820
	(0.3357)		(0.4368)	(0.4361)
Age 40 *to* 64	0.5408		1.7489*	1.6693*
	(0.6158)		(0.9401)	(0.9289)
Household Size	−0.0138*		−0.0070	−0.0056
	(0.0072)		(0.0103)	(0.0102)
Language HHI Above Mean				0.0117
				(0.0178)
Time-invariant endogenous				
Literacy Rate	−0.0549		0.0936	0.1247
	(0.0450)		(0.1279)	(0.1249)
σ_ε	0.045	0.045	0.045	0.044
ρ	0.695	0.827	0.788	0.785
R^2		0.917		
N	1,169	1,169	1,169	1,169
Wald χ^2-statistic	1,785.3		1,719.3	1,743.6
Specification test	69.7		20.3	

Notes. Standard errors are in parentheses. Year dummies and dummy variables for missing values are included for all variables in all regressions. Prices and relevant content were instrumented in all regressions. Standard errors are clustered by country and allow for general heteroskedasticity in the random-effects (RE) and fixed-effects (FE) specifications. The HT estimates use the covariance matrix specified in Hausman and Taylor (1981).

*Indicates 10% significance; **indicates 5% significance; ***indicates 1% significance.

Column 2 of Table 5 shows the second-stage results of a fixed-effects regression with standard errors clustered at the country level and robust to general heteroskedasticity. The regression yields an R^2 of 0.917, consistent with a wide range of control variables. Only a few of the control variables are significant, but there are two reasons why. First, given the country fixed effects identification comes only from time-series variation. Second, we include more control variables than previous studies (conditional on including country fixed effects). Since the results are similar to those obtained in the HT specification, we postpone their discussion. The fixed-effects estimates are consistent even if included variables are correlated with

the country-level unobservables, allowing a Hausman specification test for the consistency of the random-effects estimates. The null hypothesis of consistency is rejected below the 0.01% level with a chi-squared statistic of 69.7, consistent with correlation between unobserved country-level effects and the regressors.

Column 3 of Table 5 contains HT estimates. Since the fixed-effects specification provides consistent estimates and our model is overidentified, we can perform a Hausman specification test of the exogeneity of our HT instruments. The null hypothesis that our instruments are uncorrelated with the country-level unobservables is not rejected (16% significance level with a chi-squared statistic of 20.3). Thus, both the fixed-effects and HT estimators provide consistent estimates of the time-varying factors; however, the HT estimator is more efficient and provides consistent estimates of the time-invariant factors. This is our preferred specification, although content's effect is similar across both.

Per-capita GDP has a positive and highly significant effect on adoption. An additional \$958 in annual per capita GDP is associated with one percentage point higher adoption.[26] A country one standard deviation above the mean per-capita GDP has 9.7 percentage points higher adoption than one at the mean. This is a large effect given the mean adoption level of 9.9% in the sample.

Internet prices for two of the three years are negative but only the year 2000 prices are borderline significant (at the 12% level). The lack of significance is likely due to the lack of data. A country one standard deviation above the year 2000 mean log price has 2.5 percentage points lower adoption (25.0% of the mean adoption level). The estimates imply a price elasticity of -0.42 for Internet adoption. School enrollment and civil liberties are not significant, although there is little time-series variation in these. Telephone infrastructure has a significant negative effect on adoption, inconsistent with prior expectations. This may be because countries with heavily-subsidized and inefficient telephone industries have high Internet access prices and poor telephone infrastructure. Consistent with this, telephone infrastructure and instrumented prices are significantly negatively correlated. We also show below that the time-series impact from this variable is small.

Content has a positive and significant (below the 0.01% level) effect on adoption. A country one standard deviation above the mean in relevant content has 2.0 percentage points higher adoption or 20.0% of the mean adoption level. Countries with users of languages with more worldwide accessible content have higher adoption rates. The unreported coefficients on the year dummies are consistent with higher Internet adoption rates over time (and all but year 1999 are significant); however, this should be interpreted with caution since the content measure is not necessarily consistent over time.[27]

Of the time-invariant variables, only urbanization is significant at the 10% level or better, although the age bracket dummies are jointly significant at the 12% level. Fraction of urban population has a positive and very significant effect, consistent with either easier construction of Internet infrastructure in more densely populated areas, greater demand for access relative to more rural areas, or both. Each additional 1% of population living in urban areas is associated with 0.1 percentage points higher adoption. A country one standard deviation above the mean has 2.5 percentage points higher adoption, or 24.8% of the average adoption rate in the sample. Although the age variables are not highly statistically significant, they have a large economic impact. Countries with a smaller fraction of people above 65 years of age (the omitted age category) have higher adoption levels, with the greatest effect both statistically and economically in the age 40 to 64 category. Increasing the fraction of population in the age 40 to 64 category by one standard deviation and spreading an equivalent decrease equally across the other three categories results in a 9.3 percentage point increase in adoption (93.6% of the average adoption rate in the sample). Running the same experiment (increasing a category by one standard deviation and decreasing the other three categories equally by the same total amount) results in the following: below 20 category, a 36.1% increase; 20–39 category, a 21.2% decrease; and above 64 category, a 73.0% decrease.

5.3. Interpreting the Effect of Content

Content has a large impact on the equilibrium level of adoption. Our estimates imply an elasticity of 0.046 of adoption with respect to relevant content. For an elasticity of adoption with respect to hosts, we need to estimate how much relevant content increases with one additional host as determined by the language distributions across countries. We measure this by the ratio of relevant content to hosts across all countries (small and large) and all years, yielding 6.72. Therefore the elasticity of adoption with respect to hosts

[26] The effects of changes in independent variables are calculated at the mean values of all other variables unless otherwise noted.

[27] If some countries add more hosts earlier when computers have smaller capacity and other countries add more hosts later when computers have greater capacity, this could bias our results. We reestimated Equation (4) using a three-year moving-average of instrumented relevant content. This allows relevant content to depend on both the current and previous stocks of host computers. We tried moving averages of 1/4, 1/3, and 1/2 and found results very similar to our baseline estimates.

is 0.31. This is more powerful than the indirect network effect estimated by Gandal et al. (2000) in the compact disk (CD) market (an elasticity of CD players with respect to the number of CD titles of 0.033) but below that estimated by Dranove and Gandal (2003) for the digital video disc (DVD) market (an elasticity of DVD players with respect to fraction of movie titles released on DVD of 1.13).[28] Using our year 2000 price elasticity of −0.42, adoption is 0.74 times as responsive to content as price—above that in the CD market (the ratio of content and price elasticities is 0.54) but below that in the DVD market (a ratio of 1.2).[29]

We can also compare content's impact to that of other factors. Its effect is below that of GDP and some age-group redistributions but is comparable to the other significant control variables. A country one standard deviation above the mean in relevant content has 20.0% higher adoption. For time-varying factors the effects of a one standard deviation increase are the following: per-capita GDP, a 98.2% increase; year 2000 normalized prices, a 25.0% decrease; and telephone infrastructure, a 36.8% decrease. For time-constant factors the effects are as follows: fraction urban population, a 24.8% increase; and age distribution, a 73.0% decrease to a 93.6% increase, depending on the age category that is increased.

This has important implications for countries wishing to stimulate Internet adoption. Increasing GDP will increase Internet adoption dramatically, but this is difficult. Similarly, short-run changes in the age distribution would require dramatic immigration policy changes. Stimulating relevant content, either directly or indirectly, is easier and less costly. In addition, governments and NGOs can influence adoption in other countries by creating relevant content in the target country's languages.

There are two issues with this comparison. First, moving any of these variables by one standard deviation is a lot. Therefore, it is useful to estimate the effect of "reasonable" changes. Second, it assumes that it is equally easy to move the variables by one standard deviation. Therefore, it is useful to gauge the speed at which these variables change over time. To do so, we compute annual changes in the time-varying factors and the effects such changes would have on adoption. Since we do not have a comparable price measure over time, GDP and telephone infrastructure are the only variables to which we can compare (although we cannot measure yearly changes in the age distribution or fraction urban population, these are likely small, implying small changes in adoption).

The top panel of Table 6 summarizes these changes for the small countries. Adoption increased on average 2.2 percentage points per year in small countries. The two rightmost columns compute the effect that the annual changes in each of the explanatory variables would have on small-country adoption evaluated at the mean of all other variables. For example, per-capita GDP increased \$398 per year on average in the small countries. This would increase adoption by 0.42 percentage points, or 19.1% of the average yearly increase of 2.2 percentage points for the small countries. Similar calculations for telephone infrastructure reveal a minimal 1.0% annual decrease. Relevant content for the small countries increased on average by 885,000 hosts per year. This would increase small-country adoption by 6.0% of the average yearly increase in their adoption.

The bottom panel of Table 6 summarizes annual effects based on the large countries. Per-capita GDP increased \$493 per year on average for these countries. Such an increase would stimulate small-country adoption by 23.6% of the 2.2 percentage point annual increase in adoption for the small countries. A similar calculation for telephone infrastructure yields a 5.9% decrease. The annual increase in large-country content—the content produced by the

Table 6 Estimated Effects of Variables on Adoption by Small Countries

Variable	*N*[a]	Average annual change 1998–2004	Implied increase in adoption for small countries[b]	% of annual increase in Internet usage by small countries[c]
Small countries				
Internet Users	157	0.022		
Per-Capita GDP (US\$ thousands)	136	0.398	0.0042	19.1
Telephone Infrastructure	41	0.001	−0.0002	−1.0
Relevant Content (millions of hosts)	152	0.885	0.0013	6.0
Large countries				
Per-Capita GDP (US\$ thousands)	28	0.493	0.0051	23.6
Telephone Infrastructure	7	0.007	−0.0013	−5.9
Own Content (millions of hosts)	29	1.150	0.0017[4]	7.8[d]

Note. Large countries are identified in Table 2 and small countries in Online Appendix A.

[a]Data are missing for some countries in some years.

[b]Marginal effect evaluated at the means of all other independent variables.

[c]Relative to the average annual increase in Internet users in small countries (0.022).

[d]Assumes all content is "relevant" as defined in the text.

[28] Although these two markets are not directly analogous to the Internet, they are similar in that a CD or DVD title is replicated multiple times just as a host of content is "replicated" by multiple users accessing it.

[29] Other papers estimate the magnitude of two-sided network effects on firms' market shares (Corts and Lederman 2009, Nair et al. 2004, Ohashi 2003, Park 2004). These results are not directly comparable to ours since we estimate the effect on overall demand and they estimate the effect on firms' residual demands.

countries themselves—is 1.2 million hosts. This would increase small-country adoption by 7.8% of the 2.2 percentage points annual change in adoption for the small countries. Whether the top or bottom panel of Table 6 is more appropriate depends on which more accurately predicts annual changes. However, they are similar. In both, content is an important factor in affecting adoption—it has about one-third the impact of GDP.

5.4. Linguistic Isolation

Internet content may act as a substitute for or complement to isolation. If isolated populations use the Internet to access people with similar interests or characteristics, content would have a greater effect on adoption by more isolated groups. On the other hand, if people learn about the Internet through word of mouth, and this is less likely if one is isolated, content would have a smaller effect on adoption by more isolated groups. We distinguish these alternatives using linguistic isolation, as measured by linguistic heterogeneity. We implement this using a Herfindahl index (HHI) of language usage in each small country:

$$\mathrm{HHI}_i = \sum_{j=1}^{J} \left(\frac{\mathrm{Users}_{ij}}{\sum_{j=1}^{J} \mathrm{Users}_{ij}} \right)^2 \quad i \in I_S. \tag{9}$$

A country with an HHI close to 0 is linguistically very heterogeneous, whereas a country with an HHI of 1 is completely homogeneous. To identify content's importance in linguistically homogeneous versus heterogeneous countries, we interact instrumented relevant content with a dummy variable indicating whether a small country has an above-average HHI.

Column 4 of Table 5 shows the results. The baseline effect of language heterogeneity is insignificant. Relevant content has a positive and significant effect, but the effect is lower for countries above the mean language HHI. Content has a smaller effect in countries with more homogeneous language users. A small country one standard deviation above the mean in relevant content has 5.2 percentage points higher adoption if it is below the mean language HHI but only 1.5 percentage points if it is above. This is consistent with the Internet being a tool to overcome linguistic isolation. In contemplating the future of the online encyclopedia *Wikipedia*, its founder, Jimmy Wales, asked in mid-2009: "Is it more important to get to 10 million articles in English, or 10,000 in Wolof?" (Cohen 2009). Our results imply that in terms of adoption—the latter.

5.5. Robustness

To see whether our relevant content measure simply proxies for the small country's own content production, we add a measure of the latter to our estimation:

$$\mathrm{owncontent}_{it} = \frac{\sum_{j=1}^{J} [\mathrm{Users}_{ij} \mathrm{Content}_{ijt}]}{\mathrm{Population}_i}, \quad i \in I_S. \tag{10}$$

This differs from relevant content in Equation (5) in excluding content produced outside the country. Since this variable is endogenous, its coefficient should be interpreted with caution. Columns 1 and 2 of Table 7 show the results. Relevant content's coefficient and significance is very close to that in our baseline results in column 3 of Table 5. This is consistent with instrumented relevant content measuring content that affects but is not affected by small-country adoption. The other coefficients are not greatly affected except that the age variables are more significant. Own content is associated with higher adoption and is highly statistically significant, as would be expected in a two-sided market. The magnitude is not interpretable since it is endogenous, but it exceeds that of instrumented relevant content since it reflects the feedback between adoption and content.

Our main results assume that content's effect on adoption is the same across years. In columns 3 and 4 of Table 7, we relax this assumption and allow for differential effects in each year. The content coefficients are all positive and jointly very significant (at the 1% level). The magnitudes are similar across years (the effect of a one standard deviation increase in content ranges from 2.1 to 4.4 percentage points) and generally greater than that obtained when restricted to be equal in all years (2.2 percentage points).

If there are language-specific unobservables that drive adoption and content production, this may bias our estimates. For example, if users of certain languages have higher preferences for adoption not captured by our control variables, this will lead to higher adoption in small countries whose populations use that language and at the same time lead large countries to produce more content in that language to serve the higher large-country demand. To address this, we add language along with country and year fixed effects to Equation (1a). Once transformed into Equation (4), this is equivalent to including as a regressor the fraction of each small country's population using each language. Since including fixed effects for all languages is infeasible, we include them only for the 14 instrument languages. These are the languages that link small and large countries in our instrumenting approach and are most likely to introduce endogeneity. The results are shown in columns 5

Table 7 Effect of Content on Internet Adoption for All Sample Countries, 1998–2004, Second-Stage Estimates

	1	2	3	4	5	6
	Hausman–Taylor					
	Own content		Year effects		Language fixed effects	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Time-varying exogenous						
Per-Capita GDP	0.0095***	0.0010	0.0104***	0.0010	0.0119***	0.0015
Log Norm. Internet Price (1998)	−0.0102	0.0419	0.0101	0.0414	0.0121	0.0416
Log Norm. Internet Price (2000)	−0.0408	0.0287	−0.0419	0.0281	−0.0387	0.0285
Log Norm. Internet Price (2001)	−0.0028	0.0061	−0.0015	0.0060	−0.0007	0.0061
Fraction School Enrollment	0.0008	0.0222	0.0016	0.0218	0.0031	0.0224
Own Content	0.0311***	0.0074				
Relevant Content	0.0014***	0.0004			0.0018***	0.0004
Relevant Content (1998)			0.0093	0.0079		
Relevant Content (1999)			0.0059	0.0045		
Relevant Content (2000)			0.0033	0.0025		
Relevant Content (2001)			0.0027*	0.0016		
Relevant Content (2002)			0.0030**	0.0014		
Relevant Content (2003)			0.0023***	0.0008		
Relevant Content (2004)			0.0019***	0.0006		
Time-varying endogenous						
Telephone Infrastructure	−0.1720***	0.0169	−0.1721***	0.0166		
Civil Liberties Index	0.0001	0.0038	−0.0010	0.0038		
Time-invariant exogenous						
Gini Coefficient	−0.0018	0.0988	−0.0066	0.1088		
Fraction Urban Population	0.1185**	0.0482	0.0539	0.0510		
Age Below 20	1.1436*	0.6161	0.3947	0.6444		
Age 20 *to* 39	0.4665	0.4194	0.0835	0.4626		
Age 40 *to* 64	1.9370**	0.8979	0.7969	0.9474		
Household Size	−0.0038	0.0100	−0.0087	0.0106		
Time-invariant endogenous						
Literacy Rate	0.0856	0.1247	0.0715	0.1289		
σ_ε	0.044		0.045		0.045	
ρ	0.774		0.817		0.832	
R^2					0.919	
N	1,169		1,169		1,169	
Wald χ^2-statistic	1,757.5		1,747.1			

Notes. Prices and relevant content are instrumented in all regressions. Dummy variables for missing values are included for all variables in all regressions. Estimates in columns 2 and 4 use the covariance matrix specified in Hausman and Taylor (1981). Standard errors in column 6 are clustered by country and allow for general heteroskedasticity. Columns 1–4 also contain country and year fixed effects; columns 5 and 6 also include country, year, and language fixed effects.

*Indicates 10% significance; **indicates 5% significance; ***indicates 1% significance.

and 6 of Table 7 and are similar to our baseline estimates in column 2 of Table 5.[30]

6. Applications

Our model can be used to measure how country characteristics influence adoption's sensitivity to content. Including an interaction between country characteristics and instrumented relevant content in Equation (4) captures whether content plays a smaller or larger role as these characteristics vary. Although some of these results are descriptive, others such as those for international gateways are explicit hypothesis tests because "natural experiments" induce exogenous cross-country differences. Since we worry about the quality of instruments available for these interactions in an HT specification, we use fixed-effects specifications. Also, there are insufficient data to simultaneously identify multiple interactions, so we estimate each effect separately.[31] Thus, the effects are not conditional on the other interaction effects.

[30] An alternative explanation of our results is that countries affect each other's adoption through a direct network effect: a common language between countries leads to increased economic activity and therefore more communication via the Internet such as email or instant messaging, resulting in increased adoption. Adding a trade-weighted measure of trading partners' adoption rates to Equation (4) as a proxy for the economic closeness between country pairs has minimal effect on the estimated effect of content, consistent with indirect and direct network effects being orthogonal.

[31] In a regression combining all of the interaction effects, the coefficients on the interaction terms have similar magnitudes as in the separate regressions, although not all of them are significant.

Table 8 Effect of Content and Interactions Between Content and Country Characteristics on Internet Adoption for All Sample Countries, 1998–2004, Second-Stage, Fixed-Effects Estimates

	1 *Per-Capita GDP interaction*	2 *Telephone Infrastructure interaction*	3 *Gini Coefficient interaction*	4 *IP Protection interaction*	5 *Gateway Capacity interaction*
Per-Capita GDP	0.0134***				
	(0.0015)				
Telephone Infrastructure		−0.2064***			
		(0.0186)			
Log[*Gateway Capacity*]					−0.0028
					(0.0018)
Relevant Content	0.0027***	0.0013***	0.0028***	0.0020***	0.0017***
	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0004)
Per-Capita GDP × *Relevant Content*	−0.0001***				
	(0.0000)				
Telephone Infrastructure × *Relevant Content*		0.0071***			
		(0.0017)			
Gini Coefficient × *Relevant Content*			−0.0078***		
			(0.0019)		
Intellectual Property Protection × *Relevant Content*				−0.0003**	
				(0.0001)	
Log[*Gateway Capacity*] × *Relevant Content*					0.0010**
					(0.0005)
σ_ε	0.045	0.045	0.045	0.045	0.045
ρ	0.825	0.837	0.826	0.826	0.828
R^2	0.918	0.919	0.919	0.917	0.918
N	1,169	1,169	1,169	1,169	1,169

Notes. Standard errors, clustered by country and allow for general heteroscedasticity, are in parentheses. All control variables shown in column 2 of Table 5, year dummies, and dummy variables for missing values for all variables are included in all regressions. Prices and relevant content instrumented in all regressions.
*Indicates 10% significance; **indicates 5% significance; ***indicates 1% significance.

These results have important implications for public policy. Content plays a larger role in driving adoption in poor countries, suggesting that direct network effects may play a larger role in rich countries. The results suggest that lowering income inequality enhances content's effect on adoption. Developing both a ubiquitous domestic telephone network and high-speed international links appears to enhance content access and stimulate adoption.

These results also have important implications for firm strategies. They inform which countries Internet content providers should target in expanding internationally. A country in which adoption is more sensitive to content suggests that its population finds content more appealing. If so, countries with well-developed telecommunications networks and lower income inequality are better targets. Poorer countries are also better targets, although revenue recovery is likely problematic. Our results for intellectual property (IP) protection have interesting policy and strategy implications. Not unexpectedly, weaker IP protection allows content to more heavily influence adoption. Therefore, regulators face a trade-off in strengthening IP protection—while increasing the incentive for content creation, it discourages content dissemination. For firms, targeting countries with weaker protections is a good strategy if the firm can sufficiently protect its own content.

Per-Capita GDP: Column 1 of Table 8 shows that relevant content's effect on adoption declines in a country's wealth. A one standard deviation increase in relevant content increases adoption by 1.9 percentage points less (19.0%)[32] for a country one standard deviation above the mean per-capita GDP than for a country at the mean. Although there are alternative explanations, one possibility is that adoption in poor countries relies more on externally produced content, whereas adoption in rich countries depends more on greater direct network effects within the country.

Telephone Infrastructure: Column 2 of Table 8 shows that relevant content's effect on adoption increases in a country's telephone infrastructure quality (although the direct effect remains negative). A one standard deviation increase in relevant content increases adoption by 2.4 percentage points more (24.7%) for a country one standard deviation above the mean level of telephone main lines in use than for a country at the mean. Because the telephone network is the primary means of Internet access during our sample

[32] All comparisons in this section are to the average adoption level (0.099) in the sample.

period, this is consistent with a more pervasive network allowing widespread content access to thereby stimulate adoption.

Gini Coefficient: Column 3 of Table 8 reveals that greater income inequality in a country dampens content's influence on adoption. A one standard deviation increase in relevant content increases adoption by 1.4 percentage points less (13.8%) for a country one standard deviation above the mean Gini coefficient than for a country at the mean. This is consistent with more evenly distributed wealth leading to a broader desire to access content. Although we find no direct effect of income inequality on adoption, there is an indirect negative effect via content sensitivity.

Intellectual Property Protection: Column 4 of Table 8 investigates the role of a country's IP protection based on the intellectual property rights (IPR) component of the Intellectual Property Rights Index (Horst 2006). The index rates each country's level of IP protection on a 0 to 10 scale, with 10 being the strongest. Greater protection diminishes content's influence. A one standard deviation increase in relevant content increases adoption by 0.9 percentage points less (9.4%) for a country one standard deviation above the mean IPR than for a country at the mean. This is consistent with greater protection, making content less freely available to stimulate adoption. Of course, weaker IP protection reduces the dynamic incentives to create content, but our results suggest that it stimulates usage of extant content.

High-Speed Infrastructure: During our sample period, more than 95% of Internet traffic between countries traveled over submarine cables (Carter et al. 2009). Landing points for these high-speed cables must be in countries adjacent to the ocean. As a result, landlocked countries connect through generally slower terrestrial cables to access external content providing exogenous differences in geographic advantage. This allows us to estimate the causal indirect network effect of international gateway capacity.

We identified the major telecommunications submarine cables, their years of operation, capacity, and landing points (see Online Appendix B for sources). From this we calculated each country's gateway capacity in each year and interacted it with relevant content. The results are shown in column 5 of Table 8. International gateway capacity has an insignificant direct effect on adoption. This is consistent with international gateways being located exogenously—based on geography rather than Internet access demand. However, adoption in a country with greater capacity is more affected by relevant content than is a country with lower capacity. A one standard deviation increase in relevant content increases adoption by 1.9 percentage points more (19.1%) for a country one standard deviation above the mean log capacity than for a country at the mean.[33]

Managerial: Our results provide some rough guidelines for firms evaluating content investments. We estimate an elasticity of 0.31 of adoption with respect to number of hosts. This is the effect on the extensive margin (an increased number of adopters); assuming that the usage of these additional adopters is spread across websites in proportion to their content, this translates into an elasticity of usage on a particular website.[34] If traffic is the goal (as it is for most Internet firms), then increasing content by a given percentage has about one-third the effect of increasing adoption by the same percentage. Therefore, investments can be evaluated by comparing the marketing cost of increasing the user base by a certain percentage to the cost of increasing content by the same percentage. If this ratio exceeds about three, then the firm should focus on content production—otherwise, adoption. For firms relying on user-generated content, our estimates provide a means for adjusting a user's lifetime value to the company. On average, each percentage increase in the user base will ultimately yield roughly 1.3 times that because of the increased adoption from content that these users create. Of course, not all content is created equal. Higher-quality or more-targeted content will lead to a greater elasticity and lower-quality or less-targeted content to a smaller elasticity.

It would be preferable to have firm-level estimates of the effect of content on usage. This is possible given firm-level data during episodes in which a firm adds discrete chunks of content but does not otherwise alter its marketing efforts to encourage usage. Such estimates will reflect not only net increases in aggregate usage but also business-stealing effects (usage diverted from other Internet sites).

7. Conclusion

Internet content plays a significant role in stimulating Internet adoption. Its effect is on par with many other important social, demographic, and economic factors. Thus, content can play a crucial policy role in encouraging Internet diffusion even in the short run, and some countries are already taking action. ITU, the UN body responsible for information technologies, reports that "some countries are launching

[33] An alternative explanation is that countries adjacent to the ocean were easier to colonize and gained closer associations with large countries, sharing the same language. We estimated our results in column 3 of Table 5 excluding small countries adjacent to the ocean. The results were virtually identical.

[34] The effect on the intensive margin—increased usage by preexisting adopters—means that this will understate the elasticity of usage with respect to content. A firm-level elasticity of usage would be still greater because it includes business-stealing effects (shifting usage from other sites without increasing aggregate usage).

initiatives to subsidize the production of local content in its initial stages. Several of them are also revising and upgrading key legal instruments that would allow them to protect and promote the production of local content" (ITU 1999, p. 121).

Governments and NGOs can influence adoption, and thereby encourage social change, in other countries through this mechanism. In fact, this is implicit in our estimation strategy. Policymakers can use content targeted at particular countries and in the appropriate language to stimulate adoption in countries adversely affected by the global digital divide. Internet content can also play an important role in overcoming social isolation. Countries with more disparate language usage are more affected by content than are those with more homogeneous. More targeted Internet content is likely to have even greater effects than we find since we treat all content in a given language as equally relevant.

For Internet firms wishing to predict Internet adoption at the country level, we provide estimates for a more comprehensive list of factors driving adoption—factors varying rapidly over time as well as those more slowly changing. For firms attempting to target countries with high Internet adoption rates, our results suggest that content will influence adoption more heavily in countries with lower income inequality, better telephone infrastructure, weaker IP protection, and larger gateways connecting the country to externally produced content. This last effect is likely to increase in importance over time as Internet information includes more video and audio. Our results also suggest that targeting linguistically isolated populations offer higher expected usage of a firm's content.

Because of the need to ensure exogenous changes in content production, we are unable to estimate the effects of content on large-country adoption. This also prevents us from distinguishing the effect of content produced within a country from that produced externally. To estimate this would require different data than are available to us. It would require an exogenous change in content or its availability in large countries that affects adoption only via content. We can speculate on a few possibilities. The official use of non-Latin characters in Web addresses became feasible in 2010 because of a regulatory change. This potentially provides an abrupt exogenous change in the availability of preexisting Internet content within large countries such as China, Russia, India, and Japan. At the same time, it does not directly affect the incentive to adopt Internet access. Discrete changes in countries' intellectual property laws or their enforcement may suddenly increase or decrease availability of preexisting content within the country without otherwise changing incentives for Internet adoption. Government subsidies to produce content or bring it online might provide exogenous geographic variation if they target local populations (such as schools or local governments). These and other possibilities will have to await future research.

Acknowledgments

The authors thank Steve Berry, Avi Goldfarb, Guido Meyerhans, Hongbin Cai, Yuxin Chen, Li Gan, Fiona Scott Morton, Stéphane Straub, Noam Yuchtman, the editor, and two anonymous referees for helpful comments. They also thank seminar participants at the Cheung Kong Graduate School of Business, Peking University, Yale University, Southwestern University of Finance and Economics, Zhejiang University, and University of California, San Diego, as well as participants at the Insitut d'Economie Industrielle Conference on the Economics of the Software and Internet Industries, the Second Annual Internet Search and Innovation Conference, and the 2009 International Industrial Organization Conference. They thank Wang Xin and Qin Mian for excellent research assistance. All errors are their own.

References

Bergman MK (2001) White paper: The deep Web: Surfacing hidden value. *J. Electronic Publishing* 7(1), http://dx.doi.org/10.3998/3336451.0007.104.

Bohn RE, Short JE (2009) How much information? 2009 report on American consumers. Report, Global Information Industry Center, University of San Diego, La Jolla, CA. http://hmi.ucsd.edu/howmuchinfo_research_report_consum.php.

Campbell L, Grondona V (2008) *Ethnologue: Languages of the World* (review). *Language* 84:636–641.

Carrell S (2009) Website archives to be fast-tracked. *Guardian* (December 27), http://www.theguardian.com/books/2009/dec/27/libraries-internet.

Carter L, Burnett D, Drew S, Marle G, Hagadorn L, Bartlett-McNeil D, Irvine N (2009) Submarine cables and oceans—Connecting the world. UNEP-WCMC Biodiversity Series Report 31, ICPC/UNEP/UNEP-WCMC, Lymington, UK.

Chinn MD, Fairlie RW (2006) The determinants of the global digital divide: A cross-country analysis of computer and Internet penetration. *Oxford Econom. Papers* 59:16–44.

Cohen N (2009) Wikipedia looks hard at its culture. *New York Times* (August 30), http://www.nytimes.com/2009/08/31/business/media/31link.html.

Corts K, Lederman M (2009) Software exclusivity and the scope of indirect network effects in the U.S. home video game market. *Internat. J. Indust. Organ.* 27:121–136.

DiMaggio P, Hargittai E, Celeste C, Shafer S (2004) Digital inequality: From unequal access to differentiated use. Neckerman K, ed. *Social Inequality* (Russell Sage Foundation, New York), 355–400.

Dranove D, Gandal N (2003) The DVD-vs.-DIVX standard war: Empirical evidence of network effects and preannouncement effects. *J. Econom. Management Strategy* 12:363–386.

Ford GS, Koutsky T, Spiwak LJ (2008) The broadband efficiency index: What really drives broadband adoption across the OECD? Phoenix Center Policy Paper No. 33 (May), http://www.phoenix-center.org/pcpp/PCPP33Final.pdf.

Gandal N (2006) Native language and Internet use. *Internat. J. Sociol. Language* 182:25–40.

Gandal N, Kende M, Rob R (2000) The dynamics of technological adoption in hardware/software systems: The case of compact disc players. *RAND J. Econom.* 31:43–61.

Goolsbee A (2002) Does the internet make markets more competitive? Evidence from the life insurance industry. *J. Political Econom.* 110:481–507.

Gordon RG, ed. (2005) *Ethnologue*: *Languages of the World*, 15th ed. (SIL International, Dallas).

Gordon RJ (2000) Does the "new economy" measure up to the great inventions of the past? *J. Econom. Perspect.* 14:49–74.

Gowrisankaran G, Stavins J (2004) Network externalities and technology adoption: Lessons from electronic payments. *RAND J. Econom.* 35:260–276.

Hammarström H (2005) Review: General linguistics: Gordon (2005)—*Ethnologue: Languages of the World*, 15th edition. *Linguist List* 16 (September 12) Article 2637.

Hargittai E (1999) Weaving the Western Web: Explaining differences in Internet connectivity among OECD countries. *Telecomm. Policy* 23:701–718.

Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. *Econometrica* 49:1377–1398.

Horst AC (2006) Intellectual Property Rights Index (IPRI): 2007 Report. Study, Americans for Tax Reform Foundation, Washington, DC. http://www.americansfortaxreformfoundation.org/userfiles/2007_IPRI.pdf.

International Telecommunication Union (ITU) (1999) Challenges to the network: Internet for development, 1999. Report, ITU, Geneva.

Kiiski S, Pohjola M (2002) Cross-country diffusion of the Internet. *Inform. Econom. Policy* 14:297–310.

Litan RE, Rivlin AM (2001) Projecting the economic impact of the Internet. *Amer. Econom. Rev.* 91:313–317.

Lyman P, Varian HR (2003) How much information? Study, School of Information Management and Systems, University of California, Berkeley, Berkeley. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Minges M (1999) Measuring the diffusion of the Internet. Presentation at INET'99: Dimensions of Internet Diffusion, June 23, Internet Society, Reston, VA. http://www.itu.int/ITU-D/ict/papers/1999/MM-Inet99-Jun99.ppt.

Mutume G (2006) Harnessing the Internet for development: African countries seek to widen access, produce content. *Africa Renewal* 20(2):14–15.

Nair H, Chintagunta P, Dubé J-P (2004) Empirical analysis of indirect network effects in the market for personal digital assistants. *Quant. Marketing Econom.* 2:23–58.

Ohashi H (2003) The role of network effects in the US VCR market. *J. Econom. Management Strategy* 12:447–494.

Park S (2004) Quantitative analysis of network externalities in competing technologies: The VCR case. *Rev. Econom. Statist.* 86:937–945.

Paolillo JC, Das A (2006) Evaluating language statistics: The ethnologue and beyond. Contract report for UNESCO Institute for Statistics.

Pratap J (2010) Qatar initiative to increase Arabic content on Internet. *Gulf Times* (February 10), http://www.gulf-times.com/site/topics/printArticle.asp?cu_no=2&item.

Scott Morton F, Zettelmeyer F, Silva-Risso J (2001) Internet car retailing. *J. Indust. Econom.* 49:501–519.

Shriver SK, Nair HS, Hofstetter R (2013) Social ties and user-generated content: Evidence from an online social network. *Management Sci.* 59:1425–1443.

Sinai T, Waldfogel J (2004) Geography and the internet: Is the internet a substitute or a complement for cities? *J. Urban Econom.* 56:1–24.

Staiger D, Stock JH (1997) Instrumental variables regressions with weak instruments. *Econometrica* 65:557–586.

Stelter B (2011) F.C.C. push to expand net access gains help. *New York Times* (November 9), http://www.nytimes.com/2011/11/09/business/media/fcc-and-cable-companies-push-to-close-digital-divide.html.

Stone B, Helft M (2009) In developing countries, Web gores without profit. *New York Times* (April 26), http://www.nytimes.com/2009/04/27/technology/start-ups/27global.html.

The Emirates Center for Strategic Studies and Research (2008) Arabic content on Internet…Obstacles and solutions. Accessed February 10, 2010, http://www.ecssr.ae/CDA/en/FeaturedTopics/PFFeaturedTopics/0,176.

Uganda Communications Commission (2005) *Funding and Implementing Universal Access: Innovation and Experience from Uganda* (Fountain Publishers, Kampala, Uganda).

United Nations (2004) World population to 2300. Report, United Nations, New York.

Wallsten SJ (2005) Regulation and Internet use in developing countries. *Econom. Development Cultural Change* 53:501–523.

Wallsten SJ (2007) Whence competition in network industries? Broadband and unbundling regulations in OECD countries. Study, Technology Policy Institute, Washington, DC, https://www.techpolicyinstitute.org/files/s8.pdf.

Waters D (2006) International net domains 'risky.' *BBC News* (October 30), http://news.bbc.co.uk/2/hi/technology/6099370.stm.

World Development Indicators Database (2010a) The World Bank: Internet users (per 100 people): International Telecommunications Union. Accessed July 10, 2010, http://databank.worldbank.org/data/views/variableselection/selectvariables.aspx?source=world-development-indicators#s_i.

World Development Indicators Database (2010b) The World Bank: Internet users (per 100 people): International Telecommunications Union; and The World Bank: Population (total). Accessed July 10, 2010, http://databank.worldbank.org/data/views/variableselection/selectvariables.aspx?source=world-development-indicators#s_i.

Wunnava PV, Leiter DB (2009) Determinants of intercountry Internet diffusion rates. *Amer. J. Econom. Sociol.* 68:413–426.