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

Abstract

A country's human capital and economic productivity increasingly depend on the Internet due to its expanding role in providing information and communications. This has led to a search for ways to increase levels of Internet access 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 on adoption and therefore its effectiveness as a policy tool is previously unknown.

Quantifying content's role is challenging because there is a positive feedback loop (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 and accurately measure content's impact. We find a statistically and economically significant effect, implying that policies promoting content creation can substantially increase Internet adoption even in the short run. Because it is ubiquitous, Internet content is also a useful tool 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, and in countries with international Internet gateways, underlining the importance of high-speed infrastructure in delivering content.

Keywords: Internet, technology adoption, economic development, two-sided markets, network effects, technology diffusion, digital divide, language.

JEL Classification: O30, O57, L86, L96.

V. Brian Viard
Cheung Kong Graduate School of Business
Beijing 100738 China
brianviard@ckgsb.edu.cn
Tel: 86-10-8518-8858

Nicholas Economides
Stern School of Business
New York University
New York, NY 10012
neconomi@stern.nyu.edu
Tel: 1-212-998-0864

This Draft: 10/27/2011

* We would like to thank Steve Berry, Avi Goldfarb, Guido Meyerhans, Hongbin Cai, Yuxin Chen, Li Gan, Fiona Scott Morton, Stéphane Straub, Noam Yuchtman for helpful comments as well as seminar and conference participants at CKGSB, Peking University, Yale University, Southwestern University of Finance and Economics, Zhejiang University, University of California, San Diego, the IDEI 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. We thank Wang Xin and Qin Mian for excellent research assistance. All errors are our own.

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 focusing 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?"[3] What is not known is the magnitude of content's effect on adoption. This has important policy implications. Because content is more easily altered than economic, educational, or infrastructure conditions, it offers governments and non-governmental organizations (NGOs) a means to more quickly influence Internet diffusion. Our estimates quantify the effectiveness of this "build content and they will adopt" strategy.

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."[4] However, the question remains how effectively content stimulates adoption.

Content has a statistically and economically significant effect on Internet adoption, implying that it is an effective policy tool. Since our estimates explicitly recognize language as the conduit from content to adoption, they also confirm that creating content in underserved languages is an effective policy to address the global "digital divide." We quantify content's effect on adoption in three different ways but all indicate a large effect. First, a country one standard deviation above the mean level of relevant content has an Internet adoption rate 2.0 percentage points or 20% higher than the mean adoption rate of 9.9 percentage points in our sample. Second, the magnitude of content's effect is about one-third that of GDP (the most significant driver of adoption) and stronger or of similar strength to that of other economic, infrastructure, and demographic factors that significantly affect

---

[1] International Telecommunications Union in *World Development Indicators*, World Bank.
[2] For an aggregate study on the link between the Internet and productivity see Litan and Rivlin (2001) but a critique by Gordon (2000). Industry-specific studies include Goolsbee (2002) in health insurance and Scott Morton, *et al*. (2001) in car retail. ITU (1999) provides a policy perspective on the Internet's economic and social role.
[3] ITU (1999), page 4, italics in original.
[4] "Harnessing the Internet for Development: African Countries Seek to Widen Access, Produce Content," *Africa Renewal*, United Nations, Vol. 20, No. 2, July 2006, page 14.

adoption. Third, the annual rate of content creation in our sample led to an annual increase of 6.0 to 7.8% in adoption.

Because of its ubiquity, Internet content can potentially influence adoption and therefore public opinion across political jurisdictions. This is difficult or impossible for most other factors affecting adoption. Our findings quantify the effectiveness of efforts such as the U.S. State Department's "Public Diplomacy 2.0" initiative in which it uses the Internet for diplomacy.[5]

We 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.

We provide evidence that infrastructure for delivering high-speed data is important in facilitating access to Internet content. Content affects adoption more in countries with international Internet gateways than in land-locked countries which must access content over slower data lines. Finally, we offer weak evidence that direct network effects play a role in Internet adoption across countries – adoption in a country is significantly influenced by adoption in its most important trading partners. This direct network effect operates independently of content's indirect network effect.

Primarily because of endogeneity issues involved in estimation, there is no previous work properly assessing content's role in Internet adoption. Internet service is a two-sided market – user adoption depends on content availability and vice-versa. This feedback makes it difficult to empirically isolate the effect of content on adoption. 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.

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. ITU (1999) uses aggregate web-traffic statistics to show that language determines Internet content's relevance. Gandal (2006) shows that language usage heavily influences the languages of websites visited during individual-level browsing and provides evidence that English-

[5] "U.S. Public Diplomacy: Key Issues for Congressional Oversight," United States Government Accountability Office, Report GAO-09-679SP, May 2009 describes the initiative's activities.

language dominance in Internet content may continue based on the online behavior of bilingual users. Although used for a different objective, Gandal's results confirm our use of language to define Internet content relevance.

Various government policies affect Internet content production. Governments directly create content, so much so that its quantity has raised concerns about effective archiving.[6] Much of this content is generated as 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.[7] NGOs have also targeted underserved languages. One NGO described content development efforts in Uganda as, ". . . increasingly important and valuable to the market."[8] Arab countries working with NGOs have established rewards for high-quality, Arabic content and encouraged collaboration between universities and research centers to produce content.[9]

Perhaps more important than governments' direct content creation are their policies indirectly affecting creation. 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 languages using non-Latin characters. In response to this, the Internet Governance Forum approved a multi-year effort to allow non-Latin characters in website addresses.[10] Similarly, many Internet browsers will not properly display Arabic content due to a lack of agreement among Arab countries on a uniform format.[11]

Governments and international organizations also affect copyright policies which impact access to Internet information. Copyright issues have loomed large in "Google Book Search," a private-sector effort to make millions of books in different languages available online.[12] While the copyright issues involved are complex, our results suggest that stimulating adoption is one factor to consider in making these books easily accessible and priced cheaply.

Our paper also provides a more complete picture of factors affecting cross-country Internet adoption. Many studies have examined a variety of factors; however these usually rely on cross-sectional data and do not control for country-specific unobserved factors. Our panel-data estimates reveal only a few significant factors after controlling for country-level unobservables, one of these

---

[6] "Website Archives to be Fast-Tracked," *The Guardian*, December 27, 2009 and "National Archives: The Challenge of Electronic Records Management," General Accounting Office, Report #T-GGD-00-24, October 20, 1999.
[7] "Qatar Initiative to Increase Arabic Content on Internet," *Gulf Times*, February 10, 2010.
[8] Canada's International Development Research Centre (IDRC) described in *Funding and Implementing Universal Access: Innovation and Experience from Uganda*, Uganda Communications Commission, International Development Research Centre, Ottawa, Ontario (Chapter 3).
[9] "Arabic Content on Internet . . . Obstacles and Solutions," The Emirates Center for Strategic Studies and Research, April 22, 2008.
[10] "International Net Domains 'Risky,'" *BBC News*, October 30, 2006. Methods of using non-Latin characters in website addresses emerged in 2003 but without standardization or official approval.
[11] "Arabic Content on Internet . . . Obstacles and Solutions," The Emirates Center for Strategic Studies and Research, April 22, 2008.
[12] As of March, 2009 Google had scanned over seven million books for online searching and was placing newspaper advertisements in more than seventy languages to alert authors of a court settlement over copyrights to the books ("A Google Search of a Distinctly Retro Kind," *New York Times*, March 3, 2009 and "Google Hopes to Open a Trove of Little-Seen Books," *New York Times*, January 5, 2009).

being content. Only a few papers include content as an explanatory variable and these do not properly control for the endogeneity of content production.

2. Identification Strategy

Internet service is a two-sided market as formalized by Rochet & Tirole (2003). In a two-sided market, network effects for two products interact in a common platform so that "hardware" depends on "software" adoption and vice-versa. In Internet services, access is "hardware" and content is "software." Greater content supply drives adoption and a larger installed base drives content creation. Empirically, this feedback loop creates a difficult identification problem. Simply relating adoption and content will overstate content's effect as it will conflate the effect of content on adoption with the feedback effect of adoption on content.

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.[13] Identification relies on the assumption that content created by the "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. When we present our data and results, we provide empirical evidence that this is so. At the same time, relevant content consumed in "small" countries is affected by (and includes) content produced by "large" countries because Internet content is ubiquitous. We omit the "large" countries from estimation. 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 a small population 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 while a large number of 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 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% Vlaams, 5% Limburgisch, and 2% Italian as their primary

[13] We use 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 this is endogenous.

language.[14] 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 byproduct, the distributions of language usage 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 its population.

Our instrumentation strategy can also be described as follows. Imagine adding one unit of content in a particular language in a "large" country. As a result, Internet access is more valuable to potential adopters using that language in both "large" and "small" countries because there is some probability they will want to view the extra content. The resulting aggregate increase in adoption from this increased value stimulates further content production by the "large" country. However, by including only "small" countries in our estimation this feedback is trivial so that we capture only the initial effect of content on adoption. This argument applies regardless of how and whether the content provider charges for the content. A profit-maximizing producer will add an additional unit of content if its fixed production costs will be covered by the revenues from adopters in both "large" and "small" countries. This includes revenues from existing as well as new adopters that the increased content generates. Stimulation of new adopters leads to the creation of still more content as more producers are able to cover their fixed costs, which in turn creates more adoption and so on. Again, we eliminate this feedback because the additional adopters in "small" countries, and therefore revenues from them, are trivial in the aggregate.

Our identification strategy is related to that in Gowrisankaran and Stavins (2004), who estimate network effects in the adoption of the automated clearinghouse system (ACH) by U.S. banks. The authors face an identification issue similar to ours. Clusters of banks adopting ACH may be due to network effects but may also be due to a strong local preference for ACH. To isolate the network effect, one method the authors use is the effect of adoption by small branches of large banks on the adoption decisions of rival banks located in the same local markets. Identification is based on the fact that a bank must implement ACH at all its branches simultaneously. Our identification strategy differs in that we use the distribution of languages across countries as exogenous variation in the content relevant to each adopting country. It is exogenous since people do not move to access Internet content, which is ubiquitous in the absence of government interference.

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 users of a language in a country adopting the Internet is a function of the worldwide content available in that language since

[14] The remaining 4% use languages that represent less than 1% of Belgium's population.

Internet content is accessible anywhere.[15] 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.[16]

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

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

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

where Adopters_{ijt} is the number of Internet adopters among users of language j in country i at time t, Users_{ij} is the number of users of language j in country i which 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, while 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, $\tilde{\varepsilon}_{ijt}^A$ and $\tilde{\varepsilon}_{ijt}^C$, are independently and identically distributed across countries, languages, and time periods. We expect $\gamma^A, \gamma^C > 0$. This specification assumes that the effect of content on adoption is the same across languages. While in theory we could allow the effect to vary by language, in practice there is insufficient data to identify these separate effects.

If X_{it}^A and X_{it}^C 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

[15] We control for government restrictions on Internet access in our estimation.
[16] As explained below, the content is not necessarily hosted on a computer located physically within the country.

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:

(1b') $$\text{Content}_{ijt} = \beta^C X_{it}^C + \lambda^C Z_i^C + \rho_t^C + \delta_i^C + \gamma^C \sum_{k\in I_L} \text{Adopters}_{kjt} + \tilde{\varepsilon}_{ijt}^C, \forall i \in I_L, \forall j \in J_F.$$

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:

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

Put differently, the set of "small" countries are interdependent due to the two-sided nature of the market, while "large" country content production is independent of adoption in "small" countries. This allows an instrument which affects worldwide relevant content consumed by the "small" countries but does not affect their adoption rates except via content. By excluding "large" countries from the analysis and using their content production as an instrument we break the feedback loop between content and adoption for the "small" countries. The exclusion restriction is met based on content production by "large" countries being unaffected by adoption in "small" countries. The inclusion restriction is met because "large" country content affects content consumed by "small" countries because Internet content is ubiquitous.

To preserve degrees of freedom we use only the world's most pervasive languages to construct the instrument. Enlarging this set involves a tradeoff 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 means the instrument is more highly correlated with the "small" countries' consumed content. In Section 5 we provide empirical tests to assess the exogeneity and relevance conditions for our instrument. Our choice of languages for the instrument is discussed in Section 4.

Since we observe only the aggregate number of Internet adopters in each county, 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:

(2) $$\sum_{j=1}^{J} \text{Adopters}_{ijt} = \left(\beta^A X_{it}^A + \lambda^A Z_i^A + \rho_t^A\right) \sum_{j=1}^{J} \text{Users}_{ij} + \sum_{j=1}^{J} \left[\text{Users}_{ij} \left(\delta_i^A + \tilde{\varepsilon}_{ijt}^A \right) \right] + \gamma^A \sum_{j=1}^{J} \left[\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt} \right], \forall i \in I_S.$$

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

$$(3) \qquad \sum_{j=1}^{J} \text{Adopters}_{ijt} \Big/ \sum_{j=1}^{J} \text{Users}_{ij} = \sum_{j=1}^{J} \text{Adopters}_{ijt} \Big/ \text{Population}_i = \text{Penetration}_{it},$$

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:[17]

$$(4) \qquad \text{Penetration}_{it} = \beta^A X_{it}^A + \lambda^A Z_i^A + \rho_t^A + \delta_i^A + \gamma^A \frac{\sum_{j=1}^{J} \left[\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt} \right]}{\text{Population}_i} + \varepsilon_{it}^A, \forall i \in I_S.$$

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

$$(5) \qquad relcon_{it} = \frac{\sum_{j=1}^{J} \left[\text{Users}_{ij} \sum_{k=1}^{I} \text{Content}_{kjt} \right]}{\text{Population}_i}, i \in I_S.$$

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

$$(6) \qquad instrument_{it} = \frac{\sum_{j=1}^{J} \left[\text{Users}_{ij} \sum_{k \in I_L} \text{Content}_{kjt} \right]}{\sum_{j=1}^{J} \text{Users}_{ij}}, i \in I_S.$$

The presence of unobservable factors that affect both content and adoption but are separate from the indirect network feedback loop could bias our estimates. Our instrumenting approach combined with the large number of controls we include in estimation is effective in handling all but the most highly idiosyncratic sources of unobserved factors. In Equation (4) the content variable, once instrumented, will only be correlated with the error if the unobserved factors drive both "large" country content production and "small" country adoption. Moreover, 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 that might bias our results cannot be common to countries within the same year or result from country-specific characteristics. This means that our estimation approach is robust to among other things: country-specific policies that promote adoption or content production; changes in

[17] Transforming Equation (1a') into Equation (4) introduces heteroskedasticity at the country level since the distribution of languages varies across countries. This is difficult to accommodate in the Hausman-Taylor estimates. 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.

standards that promote adoption or content production Internet-wide; secular trends in adoption or content production; and unobserved heterogeneity across countries in adoption or content production.

Our instrumenting approach creates groupings of "small" and "large" countries based on the exogenous distributions of language usage across countries. To introduce bias in our estimates 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 after controlling for observables and country-specific characteristics. Importantly, these languages, and therefore the set of "large" countries within each group, differ for each "small" country and the distribution of languages 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 be introduced is through language-specific unobservables. To address this, we incorporate language fixed-effects (in addition to country and year fixed-effects) as a robustness check when we discuss our results.

ε_{it}^{A} is a country-time period unobservable that affects adoption in country i at time t. We 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:

$$(7)\quad E\left(\varepsilon_{it}^{A}\right)=E\left(\delta_i^A\left|X_{1it}^{A},Z_{1it}^{A}\right.\right)=0 \text{ but } E\left(\delta_i^A\left|X_{2it}^{A},Z_{2it}^{A}\right.\right)\neq 0, Var\left(\delta_i^A\left|X_{1it}^{A},Z_{1it}^{A},X_{2it}^{A},Z_{2it}^{A}\right.\right)=\sigma_\delta^2,$$
$$Cov\left(\varepsilon_{it}^{A},\delta_i^A\left|X_{1it}^{A},Z_{1it}^{A},X_{2it}^{A},Z_{2it}^{A}\right.\right)=0,\ Var\left(\varepsilon_{it}^{A}+\delta_i^A\left|X_{1it}^{A},Z_{1it}^{A},X_{2it}^{A},Z_{2it}^{A}\right.\right)=\sigma^2=\sigma_\varepsilon^2+\sigma_\delta^2,$$
$$Corr\left(\varepsilon_{it}^{A}+\delta_i^A,\varepsilon_{is}^{A}+\delta_i^A\left|X_{1it}^{A},Z_{1it}^{A},X_{2it}^{A},Z_{2it}^{A}\right.\right)=\rho=\sigma_\delta^2/\sigma^2.$$

This error structure allows the Hausman and Taylor (1981) (HT) 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 of our independent variables and justify our use of the HT estimator vis-à-vis a fixed-effects and random-effects estimator in Section 5.

Ideally we would estimate the effect of Internet adoption on content using a similar strategy – use adoption rates in "large" countries as an instrument for adoption rates in "small" countries. 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" county content are substitutes. This problem does not arise in the adoption model because adoption by users outside a country is not a substitute for adopters inside. We also face a practical problem; we do not observe language-specific adoption rates. Therefore, only time-series variation would identify the effect of adoption on content. This problem does not arise in explaining adoption because the distribution of language-specific content adds significant cross-sectional variation.

4. Data

Our sample includes data on 164 "small" countries and 31 "large" countries from 1998 to 2004.[18] We include non-self-governing territories as "countries."[19] We include these because we believe the social and economic conditions in these territories differ substantially from their governing countries so that they represent independent observations. Table 1 contains summary statistics on all variables.[20]

Internet Users: Our dependent variable is the fraction of the country's population with Internet access in country i at time t (see Figure 2 for adoption rates in 2004 for the "small" countries in our sample). The International Telecommunications Union (ITU) collects this data and does not distinguish speeds or modes of Internet access. During our sample years, virtually all Internet 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 measures all Internet users regardless of their location.[21] Unfortunately, the data do not allow us to control for access speed since content may drive adoption of higher-quality access. However, during our sample period most relevant content is text minimizing this concern.[22]

Content: We measure content by the number of host computers connected to the Internet in each year for each country.[23] Host computers contain accessible content and the total quantity of content is proportional to the number of computers. This does not measure content quality; however, for our purposes it need only be the case that quality is proportional to storage capacity across different languages. We do not directly observe the language of the content on these computers but rather infer it from the country of registration as explained below.

The number of Internet host computers is 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,

[18] Online Appendix A contains a list of the "small" countries.
[19] The non-self-governing territories include 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). Content measures are not available for Hong Kong, Macao, and Mayotte so they are not used in identifying the effect of content.
[20] Online Appendix B contains more details on all the variables and their sources.
[21] ITU's data distinguishes between "Estimated Internet Users" and "Internet Subscribers." Users of Internet cafes, for example, would be included in the former, which is our variable, but not in the latter.
[22] In 2003, image data represented 23% of all file space for publicly-available data on the Internet (Lyman and Varian, 2003). Since images may also contain text this is an upper bound for 2003. Image data has taken a higher proportion of file space over time due to faster Internet access speeds. Since 2003 is near the end of our sample period, image data in the earlier years is likely an even smaller fraction. Bohn and Short (2008) estimate that in 2008 Internet text comprised 178 hours of usage for the average Internet user while video comprised two hours. In terms of storage, they estimate that in 2008 there were 8.0 exabytes of Internet text compared to 0.9 of video. Video would play an even smaller role during our sample period when Internet connections were much slower.
[23] Host computers are connected to the Internet and hold accessible content. There are many more computers connected indirectly to the Internet through local area networks (intranets). Content is only accessible on the Internet if stored on a host computer. Computers attached to an intranet can access the Internet but cannot host content.

making the measure of hosts comparable across years. However, since computer storage capacity changed over time, we include year dummies in all our estimates and also estimate year-by-year effects as a robustness check.

The ISC data also allow us to allocate hosts (content) to each country and thereby to languages. Assignment of a host to a country does not necessarily mean that the computer is physically located within the country; however, this is fine for our purposes as long as 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.[24]

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.[25] Using this measure of content for each country in each year combined with the language data we construct the relevant content and instrument for each country-year pair based on Equations (5) and (6).

It is necessary to discuss one issue with ITU's estimates of Internet usage. Prior to 1999, if ITU could not find an independent estimate of the number of 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 measure the number of Internet users.[26] To see if this is a problem, we regress the number of Internet users on the number of hosts for the countries in our sample. Although the number of hosts and users should be related due to the two-sided nature of the market, they should not move in lockstep. This regression yields an R^2 of 0.21,[27] which is virtually identical to the R^2 of 0.22 obtained from regressing the number of Internet users on the number of telephone lines for the countries in our sample. Thus, the number of Internet users is no more closely related to hosts than to the number of telephone lines, data which is collected through a completely separate process.

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), Hammarström (2005), and Paolillo and Das (2006)). *Ethnologue* provides detailed and comprehensive estimates of the number of first-language speakers of each of the world's languages by country.[28] Its data is not complete enough to estimate using second-language speakers.

Since Internet content was primarily textual during our sample period, we ideally would use the number of literate users of each language in creating our relevant content measure. Since we do

[24] Online Appendix C contains more detail on how ISC collects the host data and allocates it to countries.
[25] This is not a major concern for our instrument as 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.
[26] ITU report "Measuring the Diffusion of the Internet" at: www.itu.int/ITU-D/ict/papers/1999/MM-Inet99-Jun99.ppt.
[27] The R^2 using only 1998 data is 0.36 indicating that there may be fewer independent estimates of Internet users in that year.
[28] *Ethnologue* does not distinguish between native and primary first-language speakers. This should be considered in interpreting our results.

not observe language-specific literacy rates by country we use the number of speakers of each language in the 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 only a small number of 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 one percent of the country's population or all languages were exhausted. Across all countries this required including 811 languages or spoken dialects.

To choose the languages for our instrument, we apply the two criteria discussed in Sections 2 and 3: it is spoken in many countries and its usage distribution is skewed with a few countries comprising a significant fraction of total users. Such languages simultaneously generate significant data, while maintaining the exogeneity assumption necessary for identification. Based on these criteria we use fourteen languages to construct our instrument:

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

The first eight are among the top ten most-spoken languages in the world based on *Ethnologue*.[29] French is the seventeenth most-spoken language. The usage of the languages between the tenth and seventeenth (Javanese, Telugu, Marathi, Vietnamese, Korean, and Tamil) is either not widespread or is 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 fourteen 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. To choose the "large" countries (the set I_L) for our instrument (Equation (6)) we used the following procedure. For each language, sort the countries in descending order according to 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 to this procedure when we kept adding above 75%.[30] Column 5 of Table 2 shows the 31 "large" countries chosen, while Columns 6 through 8 show the number of users in the "large" countries and as a percentage of worldwide users. Identification relies on the percentages in Column 8 being large so that these countries are unaffected

[29] Arabic (fourth-ranked) and Bengali (seventh-ranked) were not included because their usage was not skewed enough.
[30] 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 U.S. and the U.K. 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.

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 the largest "small" country for each language. 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 percentages in Column 11 being small so that Internet adoption by these countries does not affect "large" countries' content production. The largest "small" countries represent eight percent or less 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, infrastructure, prices, and freedom of expression. Per-capita gross domestic product (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. While 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 *Freedom House*'s measure of the freedom of citizens in each country to engage in expression. This measure of civil liberties controls for the degree of government restrictions on content access and ranges from one to seven with seven being the most free.[31]

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 measures a different type and amount of usage, we cannot pool it across years. Internet access prices and adoption may both be higher in countries with higher unobserved access quality, which would bias the price coefficient upward. 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.[32]

[31] *Freedom House* defines seven as the least free. We reverse the order for ease in interpretation.
[32] We experimented with individual tax rates and ratio of government expenditures to GDP but these were highly collinear with the other instruments and did not significantly increase overall power.

We also include a number of time-constant controls. These are "time-constant" in that we observe only one year of data, although they are likely to change slowly. The Gini coefficient of income controls for the distribution of wealth 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 or 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. Since familiarity with the Internet is likely age-dependent, we control for the age distribution of a country's population using the fraction of the population in each of 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 (2006) explains broadband penetration for OECD countries, while Wallsten (2005) assesses the impact of regulation on developing countries' Internet adoption rates and prices. Ford, *et al*. (2007) 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 the effect of language on Internet adoption all of which include only a single language (English) and do not address 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 (2008) 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, although they do not address endogeneity.

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 quantify its effect in several ways and find an important role regardless. First, we compare content's effect to that of other determinants of adoption. While below that of per-capita GDP, content's effect is greater than or similar to that of other economic, demographic, and infrastructure variables. Second, we calibrate the additional adoption that would result from a country being richer in content availability (not necessarily in production). A country one standard deviation above the mean in relevant content has a 2.0 percentage point higher adoption rate (20.0% of the average adoption rate of 9.9 percentage points

in the sample). Third, we quantify the additional adoption that results from the annual content production in our sample: 6.0 to 7.8% per annum.

These findings are robust to different specifications and to including a rough measure of direct network effects. Content has a greater effect in countries with more disparate languages, consistent with it helping overcome linguistic isolation. Content also has a greater effect in countries with international Internet gateways, consistent with high-speed infrastructure providing better content access.

Before we estimate formally 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 while 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 county'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, while 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.

First-Stage Results: Columns 1 through 3 of Table 4 show the first-stage results for Internet access prices. Given the small number of observations in each year, the first-stage coefficients are somewhat noisy. Number of telephone employees has a positive effect and is significant in two of the three years consistent with higher prices from lower productivity. 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 for the first-stage regression relating relevant content to the instrument ("large" country content). We use a second-order, Taylor-series expansion of the instrument as shown in Column 4 of Table 4.[33] 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

[33] A cubic term was not significant.

10 specified in Staiger and Stock (1997) to rule out weak instruments. To make sure that our results were not sensitive to the quadratic functional form, we re-estimated 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.

Panel Data Results: Although we control for many factors thought to affect Internet adoption, country-level unobservables likely remain. Therefore, we include country fixed-effects. 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 other variables as possible, we would like to 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, all but telephone infrastructure and civil liberties are likely exogenous in the HT sense (*i.e.*, uncorrelated with the unobserved country-level effects). 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 the literacy rate. The income distribution, age distribution, average household size, and urban density are not likely affected by unobserved preferences for Internet adoption. We allow for the possibility that the literacy rate is endogenous in the HT sense (*i.e.*, correlated with the country-level unobservables). Measuring literacy is subjective as 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 containing the variables classified 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 (at the 0.0% level).

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 a high 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 at the 0.0% 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 regardless of correlations between the regressors and the country-level unobservables and since our model is over-identified 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 on adoption is similar across both specifications.

Per-capita GDP has a positive and highly significant effect on adoption. An additional $958 in annual per capita income is associated with a one percentage point higher adoption.[34] 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 effects of school enrollment and civil liberties are not significant although there is little time-series variation in these. Telephone infrastructure has a very 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 will also see below that the time-series impact from this variable is small.

Content has a positive and significant (at the 0.0% 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 content accessible 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 very significant); however this should be interpreted with caution since the content measure is not necessarily consistent over time.[35]

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

[35] If some countries add more hosts when hosts have smaller capacity while other countries add more hosts later when hosts have greater capacity, this could bias our results. We re-estimated Equation (4) using a three-year

Of the time-invariant variables, only the urbanization variable is significant at the 10% level or better; although the age 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 or greater demand for Internet access in these areas relative to more rural or both. Each additional one percent 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: below 20 category a 36.1% increase, 20 – 39 category a 21.2% decrease, and above 64 category a 73.0% decrease.

Interpretation of Content's Effect: Content's impact 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: 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: 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 is important for countries who wish to stimulate Internet adoption. Increasing GDP will increase Internet access dramatically, but this is difficult. Similarly, short-run changes in the age distribution would require dramatic changes in immigration policies. Stimulating relevant Internet content, either directly or indirectly, is easier and less costly. In addition, governments and NGOs can influence Internet adoption in other countries by creating relevant content in the languages of the target country.

There are two issues with the above comparison. First, moving any of these variables by one standard deviation is a large change. Therefore, it is useful to estimate the effect of "reasonable" changes. Second, this comparison assumes that it is equally easy to move any of the variables by one standard deviation. Therefore, it is useful to gauge the speed with which these variables change over time. To do so, we compute changes in the time-varying factors over our sample period and compute

---

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 ¼, 1/3, and ½ and found results very similar to our original estimates. These results are available upon request.

the effects such changes would imply for adoption. Since we do not have a consistent price measure over time, per-capita 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 extremely small implying very 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 for the "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 the 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 thousand 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 content for "large" countries – the content produced by the 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 rates of change over time. However, they are similar. In either case, content is an important factor in affecting adoption – it has about one-third the impact of GDP.

Linguistic Isolation: Internet content may act as a substitute for or complement to isolation. On the one hand, isolated populations may use the Internet as a means to access people with similar interests or characteristics. If so, content would have a greater effect on adoption by more isolated groups. On the other hand, people may learn about the Internet's usefulness through word-of-mouth and this is more likely if they are less isolated. If so, 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 measure linguistic heterogeneity using a Herfindahl index (HHI) of languages used in each of the "small" countries:

$$\text{(9)} \qquad \text{HHI}_i = \sum_{j=1}^{J} \left(s_{ij}\right)^2 \; i \in I_S,$$

where $s_{ij} = \text{Users}_{ij} \Big/ \sum_{j=1}^{J} \text{Users}_{ij}$. A country with an HHI close to zero is very heterogeneous linguistically while a county with an HHI of one 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 the "small" country has an above-average HHI.

Column 4 of Table 5 shows the results. The baseline effect of language heterogeneity is not significant. Relevant content has a positive and very 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 higher adoption if it is above.

This result is consistent with people using the Internet as a tool to overcome linguistic isolation and complements that of Sinai and Waldfogel (2004). Using individual-level data from the U.S., they find that the Internet is used to overcome racial isolation; blacks are more likely to adopt the Internet if they are a smaller fraction of the local population. 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?"[36] Our results imply that in terms of adoption – the latter.

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

$$(10) \qquad owncontent_{it} = \frac{\sum_{j=1}^{J} \left[\text{Users}_{ij} \text{Content}_{ijt} \right]}{\text{Population}_i}, i \in I_S .$$

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. A country's own content is associated with higher adoption and is highly statistically significant as would be expected in a two-sided market. The magnitude of this variable's effect is not interpretable since it is endogenous, but it exceeds that of instrumented relevant content since it is subject to the feedback between adoption and content.

[36] "Wikipedia Looks Hard at its Culture," *New York Times*, August 31, 2009.

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 a low of 2.1 percentage points in 2000 to a high of 4.4 in 2003) 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 would bias our estimates because instrumented relevant content would be correlated with the error term in the adoption equation (Equation (4)). For example, if users of a language have higher preferences for Internet 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 to "large" countries producing more content in that language to serve the higher demand. To address this, we add language fixed-effects in addition to country and year fixed-effects to Equation (1a). Once transformed into Equation (4) this is equivalent to including the fraction of each "small" country's population using each language as a regressor. Since including fixed-effects for all languages is infeasible, we include them only for the 14 languages used in constructing our instrument $\left(J_F\right)$. Since these are the languages that link "small" and "large" countries in our instrumenting approach, they are most likely to introduce endogeneity. The results are shown in Columns 5 and 6 of Table 7 and are similar to our baseline estimates in Column 2 of Table 5.

6. Alternative and Supplementary Explanations

Direct Network Effect: Our adoption model assumes that countries affect each other through an indirect network effect: content production by a country drives Internet adoption in other countries through the shared platform of the Internet infrastructure. Content's network effect is indirect because more adopters lead to a greater variety of content due to economies of scale in content production and it is this increase in content that leads to even greater adoption. An alternative 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, and increased adoption.

To test this, we estimate Equation (4) but add a trade-weighted measure of trading partners' adoption rates in addition to the instrumented relevant content measure. We use trade as a proxy for the degree of economic closeness between pairs of countries. We thus estimate:

(11) $\text{Penetration}_{it} = \beta^A X_{it}^A + \lambda Z_i^A + \rho_t^A + \delta_i^A + \gamma^A relcon + \theta^A directne + \varepsilon_{it}^A, \forall i \in I_S$, where:

$$(12) \qquad directne_{it} = \frac{\sum_{k=1}^{I} \text{Trade}_{ikt} * \text{Adopters}_{kt}}{\sum_{k=1}^{I} \text{Trade}_{ikt}}, i \in I_S,$$

and Trade_{ikt} is the imports from country k to country i in year t. Trade data is taken from United Nations (1999 – 2005), which includes up to the top fifteen trading partners for each country. If direct network effects alone drive the relationship between countries' adoption levels then we expect θ_A to be significant and positive and to reduce γ^A to insignificance. Columns 1 and 2 of Table 8 show the results. Relevant content's effect is almost identical to that without including trade effects. This is consistent with the indirect network effect via content being orthogonal to the direct network effect.

This measure of direct network effects is endogenous since adoption by a "small" country is affected by adoption in trading partners who are also "small" countries.[37] Therefore, we need to be cautious in interpreting the magnitude of its coefficient. In addition, it only captures economic and not cultural exchanges taking place via the Internet. However, there is evidence of direct network effects in Internet adoption. The trade-weighted adoption measure is positive and highly significant (at the 0.0% level). An additional one million dollars of imports is associated with 0.06 percentage points higher adoption. A country one standard deviation above the mean in direct network effects has 1.3 percentage points higher adoption or 12.7% of the average adoption level in the sample.

A potential issue with the trade data is that for countries with dispersed trading partners, United Nations (1999 – 2005) may not capture a significant portion of its trade since it lists only the top fifteen trading partners. Columns 3 and 4 of Table 8 show results using only those countries with 95% or more of their total imports represented.[38] The results are virtually identical to those obtained using all data. This provides weak evidence of a separate influence on Internet adoption across countries – that of a direct network effect – but it is separate from content's indirect network effect.

Effect of High-Speed Infrastructure: During our sample period, over 95% of Internet traffic between countries traveled over submarine cables.[39] Landing points for these high-speed cables must be in countries adjacent to the ocean. As a result, land-locked countries must connect through generally slower terrestrial cables to access content outside the country. Taking advantage of this exogenous difference in geographic advantage, we estimate the effect of international gateway capacity on the indirect network effect.

Through OECD (2009) and the websites of the International Cable Protection Committee and major submarine cable consortia, we identified the major telecommunications submarine cables and

[37] Endogeneity is also possible if greater Internet adoption by two countries increases their bilateral trade contemporaneously.
[38] In the original regression, 734 of the country-year observations have import data. In the regression containing more comprehensive import data this drops to 580.
[39] "Submarine Cables and the Oceans: Connecting the World," The United Nations Environment Programme World Conservation Monitoring Center, 2009.

their years of operation, capacity, and landing points.[40] From this we calculated each country's gateway capacity in each year. We then added the log capacity as well as an interaction between it and our measure of relevant content in estimating Equation (4). The results are shown in Columns 5 and 6 of Table 8. International gateway capacity has an insignificant effect on Internet adoption. This is consistent with countries housing international gateways exogenously – based on their geography rather than Internet access demand.

The interacted term is positive and significant. That is, adoption in a country with a greater international gateway capacity is more affected by relevant content than a country with lower capacity. A one standard deviation increase in relevant content increases adoption by 1.6 percentage points more (16.3% of the average adoption level in the sample) for a country one standard deviation above the mean log capacity (3.2 gigabit per second increase) than for a country at the mean log capacity.

6. 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 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."[41]

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. 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 in our estimation. Policymakers can also use content targeted at particular countries and in the appropriate language to stimulate adoption in countries adversely affected by the global "digital divide" – the disparity across countries in the opportunity to access the Internet and reap the resulting benefits.

Countries with more disparate language usage are more affected by content than are those with more homogeneous. Thus, Internet content can play an important policy role in overcoming social isolation. We also find evidence that infrastructure, in the form of faster international data connections, amplifies the effect of content on adoption. This effect is likely to be of increasing importance as Internet information is increasingly composed of video and audio.

[40] International Cable Protection Committee's website is http://www.iscpc.org/. The gateways data is available upon request.
[41] ITU (1999), page 121.

Bibliography

Bohn, R. E. and J. E. Short (2009). "How Much Information? 2009 Report on American Consumers," manuscript available at http://hmi.ucsd.edu/howmuchinfo_research_report_consum.php.

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

Chinn, M. D. and R. W. Fairlie (2006). "The Determinants of the Global Digital Divide: A Cross-Country Analysis of Computer and Internet Penetration," *Oxford Economic Papers*, 59, 16 – 44.

DiMaggio, P., E. *et al.* (2004). "From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality," in *Social Inequality*, Kathryn Neckerman ed., Russell Sage Foundation, New York.

Ford, G., Koutsky T. and L. Spiwak (2007). "The Broadband Performance Index: A Policy-Relevant Method of Comparing Broadband Adoption among Countries," working paper.

Gandal, N. (2006). "Native Language and Internet Use," *International Journal of the Sociology of Language*, 182, 25 – 40.

Goolsbee, A. (2002). "Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry," *Journal of Political Economy*, 110, 481 – 507.

Gordon, R., editor (2005). *Ethnologue: Languages of the World*, 15th edition, SIL International, Dallas, Texas. Online version: http://www.ethnologue.com/ accessed in May 2009.

Gordon, R. J. (2000). "Does the 'New Economy' Measure up to the Great Inventions of the Past?" *Journal of Economic Perspectives*, 14, 49 – 74.

Gowrisankaran, G. and J. Stavins (2004). "Network Externalities and Technology Adoption: Lessons from Electronic Payments," *RAND Journal of Economics*, 35, 260 – 276.

Hammarström, H. (2005). "Review of Raymond J. Gordon, Jr. (ed.) 2005 Ethnologue: Languages of the World, SIL International," working paper.

Hargittai, E. (1999). "Weaving the Western Web: Explaining Differences in Internet Connectivity Among OECD Countries," *Telecommunications Policy*, 23, 701 – 718.

Hausman, J. A. and W. E. Taylor (1981). "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377 – 1398.

International Telecommunications Union (ITU) (1999). "Challenges to the Network: Internet for Development."

_____ (2001). "IP Telephony."

_____ (2002). "Internet for a Mobile Generation."

_____ (2003). "Birth of Broadband."

_____ (2004). "Portable Internet."

_____ (2005). "Key Indicators of the Telecommunication/ITC Sector."

Kiiski, S. and M. Pohjola (2002). "Cross-Country Diffusion of the Internet," *Information Economics & Policy*, 14, 297 – 310.

Litan, R. E. and A. M. Rivlin (2001). "Projecting the Economic Impact of the Internet," *American Economic Review*, 91, 313 – 317.

Lyman, P. and H. R. Varian (2003). "How Much Information?, 2003," manuscript available at http://www2.sims.berkeley.edu/research/projects/how-muchinfo-2003/).

OECD (2009). *Internet Access for Development*, OECD Publishing, available at www.oecdbookshop.org.

Paolillo, J. C. and A. Das (2006). "Evaluating Language Statistics: The Ethnologue and Beyond," working paper.

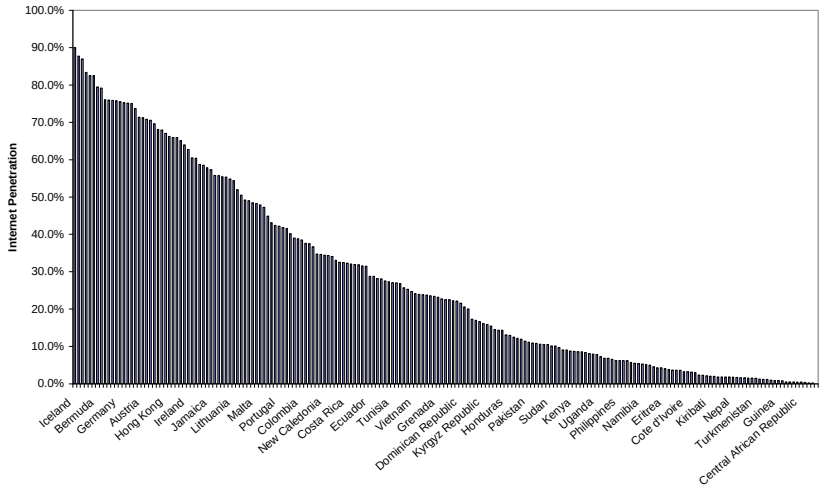
Rochet, J. and J. Tirole (2003). "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 1, 990 – 1029.

Scott Morton, F., F. Zettelmeyer, and J. Silva-Risso (2001). "Internet Car Retailing," *Journal of Industrial Economics*, 49, 501 – 519.

Sinai, T. and J. Waldfogel (2004). "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?" *Journal of Urban Economics*, 56, 1 – 24.

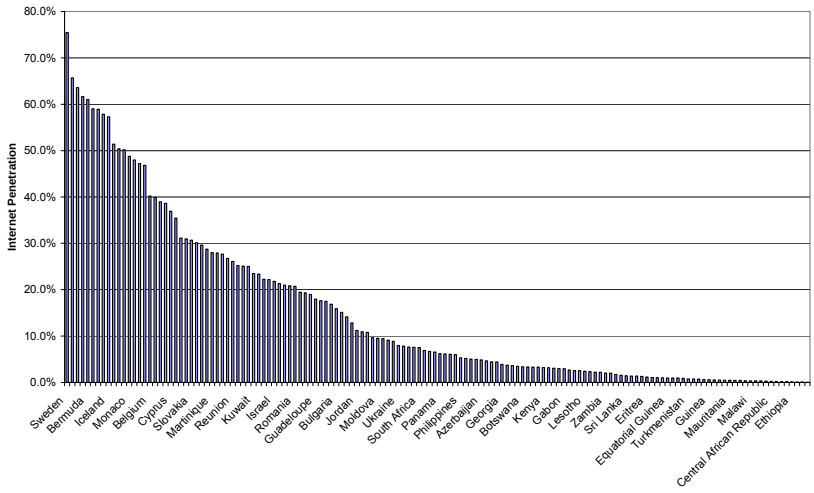
Staiger, D. and J. H. Stock (1997). "Instrumental Variables Regressions with Weak Instruments," *Econometrica*, 65, 557 – 586.
United Nations (1999 – 2005). *United Nations Commodity Trade Statistics Yearbook*, 1999 – 2005 editions, New York.
Wallsten, S. (2005). "Regulation and Internet Use in Developing Countries," *Economic Development and Cultural Change*, 53, 501 – 523.
Wallsten, S. (2006). "Broadband and Unbundling Regulations in OECD Countries," working paper.
Wunnava, P. and D. Leiter (2009). "Determinants of Inter-Country Internet Diffusion Rates," *American Journal of Economics and Sociology*, April 2009.

Figure 1 Internet Penetration across Countries in 2008



Source: International Telecommunications Union in *World Development Indicators*, World Bank. Internet penetration (fraction of population with Internet access) for 197 countries sorted from highest to lowest penetration. Not all country names displayed due to lack of space.

Figure 2 Internet Penetration in Sample "Small" Countries in 2004



Source: International Telecommunications Union in *World Development Indicators*, World Bank. Internet penetration (fraction of population with Internet access) for 164 sample "small" countries sorted from highest to lowest penetration. Not all country names displayed due to lack of space.

Table 1 Descriptive Statistics, 164 "Small" Countries, 1998 – 2004

Variable	N	Mean	Standard Deviation	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					
Language-Trade Interaction (US$ millions)	734	27.568	20.906	1.599	162.213
Log[Gateway Capacity (gigabits per second)]	1,169	0.331	1.154	0.000	8.144
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	922	12.000	20.789	0.068	175.385

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

Table 2 Profiles of “Large” and “Small” Countries for Included Languages

		Worldwide		"Large" Countries				Largest "Small" Country		
Ranking[1]	Language	Total # Users (millions)	Total Content (1000s hosts)[2]	Country	# of Users	Total # Users (millions)	% of Worldwide	Country	Total # Users (millions)	% 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%	Domincan Rep.	6.89	2.14%
				Columbia	34.00					
				Argentina	33.00					
				Spain	28.17					
				Venezuala	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 Kingdor	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-Bassau	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 [3]	138,648.22							

[1] Most-spoken languages by first-language speakers according to Gordon (2005). If blank not ranked.

[2] Average numer of hosts across six years of data.

[3] Based on year 2000 data from "World Population to 2300," United Nations, New York, 2004.

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	0.033	
	(0.000)	(0.366)	
"Large" Country Content	0.293	0.083	0.517
(Instrument)	(0.000)	(0.021)	(0.000)

Signficance levels are in parentheses.

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

	1998 Log Prices	2000 Log Prices	2001 Log Prices	Relevant Content	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

Standard errors in parentheses. * = 10% significance, ** = 5% significance, *** = 1% significance. Standard errors for relevant content regressions clustered at the country level. Dummy variables for missing values included for all variables in price regressions. Relevant content regressions also include the control variables in the second-stage regression.

Table 5 Effect of Content on Internet Adoption for All Sample Countries, 1998 – 2004, Second-Stage, Panel Data Estimates (N = 1,169)

	RE	FE	HT-GLS	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		
Wald χ^2-statistic	1,785.3		1,719.3	1,743.6
Specification Test	69.7		20.3	

Standard errors in parentheses. * = 10% significance, ** = 5% significance, *** = 1% significance. Year dummes and dummy variables for missing values included for all variables in all regressions. Prices and relevant content 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 Hausman-Taylor (HT) estimates use the covariance matrix specified in Hausman and Taylor (1981).

Table 6 Estimated Effects of Variables on Adoption by "Small" Countries

Variable	N[1]	Average Annual Change 1998 - 2004	Implied Increase in Adoption for "Small" Countries[2]	% of Annual Increase in Internet Usage by "Small" Countries[3]
"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% [4]

[1] Data are missing for some countries in some years.
[2] Marginal effect evaluated at the means of all other independent variables.
[3] Relative to the average annual increase in Internet users in "small" countries (0.022).
[4] Assumes all content is "relevant" as defined in the text.
"Large" countries are identified in Table 2 and "small" countries in Online Appendix A.

Table 7 **Effect of Content on Internet Adoption for All Sample Countries, 1998 – 2004, Second-Stage Estimates (N = 1, 169)**

	Hausman-Taylor				Language	
	Own Content		Year Effects		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	
Wald χ^2-statistic	1,757.5		1,747.1			

* = 10% significance, ** = 5% significance, *** = 1% significance. Prices and relevant content instrumented in all regressions. Dummy variables for missing values 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 through 4 also contain country and year fixed-effects while Columns 5 and 6 also include country, year, and language fixed-effects.

Table 8 **Alternative/Supplementary Explanations for Effect on Adoption for All Sample Countries, 1998 – 2004, Second-Stage, Hausman-Taylor Estimates (N = 1,169)**

	Trade		Trade Compreh.		Gateway Capacity	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Time-Varying Exogenous						
Per-Capita GDP	0.0103	0.0010 ***	0.0104	0.0010 ***	0.0106	0.0010 ***
Log Normalized Internet Price (1998)	0.0137	0.0413	0.0218	0.0415	0.0082	0.0415
Log Normalized Internet Price (2000)	-0.0392	0.0284	-0.0396	0.0284	-0.0399	0.0285
Log Normalized Internet Price (2001)	-0.0010	0.0060	-0.0011	0.0060	-0.0017	0.0061
Fraction School Enrollment	-0.0036	0.0220	-0.0038	0.0220	-0.0018	0.0221
Relevant Content	0.0014	0.0004 ***	0.0014	0.0004 ***	0.0016	0.0004 ***
Direct Network Effects	0.0006	0.0002 ***	0.0007	0.0002 ***		
Log[Gateway Capacity]					-0.0024	0.0018
Rel. Content*Log[Gateway Capacity]					0.0010	0.0005 **
Time-Varying Endogenous						
Telephone Infrastructure	-0.1721	0.0168 ***	-0.1699	0.0168 ***	-0.1741	0.0168 ***
Civil Liberties Index	-0.0001	0.0037	-0.0010	0.0037	-0.0007	0.0038
Time-Invariant Exogenous						
Gini Coefficient	-0.0260	0.0968	-0.0176	0.0997	-0.0093	0.1013
Fraction Urban Population	0.0971	0.0474 **	0.1040	0.0484 **	0.0903	0.0488 *
Age Below 20	0.6090	0.6070	0.8779	0.6226	0.8944	0.6310
Age 20 to 39	0.0744	0.4164	0.1508	0.4282	0.2308	0.4344
Age 40 to 64	1.2627	0.8943	1.6246	0.9142 *	1.5763	0.9256 *
Household Size	-0.0122	0.0096	-0.0084	0.0100	-0.0076	0.0102
Time-Invariant Endogenous						
Literacy Rate	0.0107	0.1170	0.0889	0.1252	0.1095	0.1272
σ_ε	0.044		0.044		0.045	
ρ	0.772		0.783		0.788	
Wald χ^2-statistic	1,783.7		1,770.7		1,737.7	

* = 10% significance, ** = 5% significance, *** = 1% significance. Relevant content and prices instrumented in all regressions. Year dummes and dummy variables for missing values included for all variables in all regressions. Standard errors calculated using the covariance matrix specified in Hausman and Taylor (1981).

Online Appendix A "Small" Countries Included in Analysis

	Africa[1]	The Americas[1]	Asia[1]	Europe[1]	The Pacific[1]
	Algeria	Antigua and Barbadu	Armenia	Albania	Fiji
	Angola	Aruba	Azerbaijan	Andorra	French Polynesia
	Benin	Bahamas	Bahrain	Belarus	Guam
	Botswana	Barbados	Bangladesh	Belgium	Kiribati
	Burkina Faso	Belize	Bhutan	Bosnia and Herzegovina	Marshall Islands
	Burundi	Bermuda	Brunei Darussalam	Bulgaria	Micronesia
	Cameroon	Bolivia	Cambodia	Croatia	New Caledonia
	Cape Verde Islands	Costa Rica	Cyprus	Czech Republic	New Zealand
	Central African Republic	Dominica	Georgia	Denmark	Papua New Guinea
	Chad	Dominican Republic	Indonesia	Estonia	Samoa
	Comoros	El Salvador	Iran	Finland	Solomon Islands
	Congo	French Guiana	Iraq	Greece	Tonga
	Cote dIvoire	Greenland	Israel	Hungary	Vanuatu
	Democratic Republic of the Congo	Grenada	Jordan	Iceland	
	Djibouti	Guadeloupe	Kazakhstan	Ireland	
	Egypt	Guatemala	Kuwait	Italy	
	Equitorial Guinea	Guyana	Kyrgyzstan	Latvia	
	Eritrea	Haiti	Laos	Lithuania	
	Ethiopia	Honduras	Lebanon	Luxembourg	
	Gabon	Jamaica	Malaysia	Macedonia	
	Ghana	Martinique	Maldives	Malta	
	Guinea-Bissau	Netherlands Antilles	Mongolia	Moldova	
	Kenya	Nicaragua	Nepal	Netherlands	
	Lesotho	Panama	Oman	Norway	
	Liberia	Paraguay	Pakistan	Poland	
	Libya	Puerto Rico	Palestinian West Bank and Gaza	Romania	
	Madagascar	Saint Kitts & Nevis	Philippines	Slovakia	
	Mali	Saint Lucia	Qatar	Slovenia	
	Mauritania	Saint Vincent & the Grenadines	Saudi Arabia	Sweden	
	Mauritius	Surimane	Singapore	Switzerland	
	Morocco	Trinidad & Tobago	South Korea	Ukraine	
	Mozambique	Uruguay	Sri Lanka		
	Namibia	U. S. Virgin Islands	Syria		
	Reunion		Tajikistan		
	Rwanda		Thailand		
	Sao Tome e Principe		Turkey		
	Seychelles		Turkmenistan		
	Sierra Leone		United Arab Emirates		
	Somalia		Uzbekistan		
	Sudan		Viet Nam		
	Swaziland		Yemen		
	Tanzania				
	Togo				
	Tunisia				
	Uganda				
	Zimbabwe				
# Countries	46	33	41	31	13
Ethnologue # Countries	57	51	50	45	25

[1] Classifications according to Gordon (2005). Regressions also inlcude the following countries and territories with missing language information: Afghanistan, Faroe Islands, Falkland Islands, Hong Kong, Liechtenstein, Macao, Mayotte, Monaco, Myanmar, San Marino, Serbia and Montenegro, and Tuvalu.

Online Appendix B Variable Descriptions and Data Sources

Variable	Description	Frequency/ Availability	Data Source
Internet Users	Fraction of population with some form of Internet access.	Annual/1998 - 2004	ITU (1999, 2001, 2002, 2003, 2004, 2005)
Per-Capita GDP	GDP per-capita in current U.S. dollars using purchasing power parity.	Annual/1998 - 2004	World Bank
Telephone Infrastructure	Fraction of the population with telephone main lines in use.	Annual/1998 - 2004	ITU (1999, 2001, 2002, 2003, 2004, 2005)
Normalized Internet Price	Internet monthly access price for 20 hours of off-peak use (1998 and 2000) as fraction of GDP per capita; Internet monthly access price for 30 hours of peak use (2001) as fraction of GDP per capita.	Annual/1998, 2000 - 2001	ITU (1999, 2001, 2002)
Fraction School Enrollment	Fraction of eligible populaion enrolled in primary education, years 1999 to 2004.	Annual/1999 - 2004	United Nations Statistics Division
Civil Liberties Index	Civil liberties measured on a one-to-seven scale, with one representing the lowest degree of freedom and seven the highest, years 1998 to 2004.	Annual/1998 - 2004	*Freedom in the World* , Freedom House (1999 - 2005 editions)
Literacy Rate	Literacy rate of population aged 15 and above, years 2000 to 2005.	Once	*The State of the World's Children 2008* , United Nations Childrens Fund
Gini Coefficient	Gini coefficient of inequality of income distribution, various years from 1995 to 2006.	Once	2006 United Nations Human Development Report, Table 15
Age	Fraction of population in year 2000 in four age brackets: 1) below age 19, 2) 20 to 39, 3) 40 to 64, and 4) 65 and above.	Once	United Nations Statistics Division
Fraction Urban Population	Fraction of population living in urban areas, year 2000.	Once	United Nations Statistics Division
Household Size	Average number of people per household.	Once	World Development Indicators
Relevant Content	Millions of hosts of "relevant" content. See text for detailed description.	Annual/1998 - 2004	Gordon (2005) (language) and Internet Systems Consortium (hosts)
Own Content	Millions of hosts. See text for detailed description.	Annual/1998 - 2004	Internet Systems Consortium
"Large" Country Content	Millions of hosts. See text for detailed description.	Annual/1998 - 2004	Gordon (2005) (language) and Internet Systems Consortium (hosts)

Online Appendix C Technical Details of Hosts Data Collection

The technical details of ISC's data collection are complex due to the sheer size of the Internet but ISC essentially counts the number of Internet Protocol (IP) addresses that have been assigned a Uniform Resource Locator (URL), which is the website address that users enter into a browser to locate content. An IP address is associated with a single host which is how ISC finds the host names. A request is sent to each active IP address requesting the unique host name. A host may have more than one IP address associated with it so ISC resolves these duplicates. Each computer on the Internet is assigned an IP address between 1 and 2^{32} but only those that have been assigned a URL are in use. To determine which have been assigned a URL, ISC must send a query to that address. Since it would take too long for ISC to query every possible address in use, it uses a sophisticated sampling algorithm to reduce the time.[1]

In its survey ISC gathers the URL of each host computer. This address contains a two-digit country code (e.g., .za for New Zealand, .uk for United Kingdom, and .ca for Canada) called a country-code Top Level Domain (ccTLD). ISC assigns each domain to a country based on the ccTLD.[2] The ccTLD does not necessarily imply that the computer is physically located within the country. Instead, assigning a ccTLD requires a local presence such as citizenship, resident address, or local administrative contact.

The relationship between hosts and addresses (URLs) is complicated. All web pages have a unique URL and are part of a sub-domain which is in turn part of a domain. A domain name such as "google.com" can have many sub-domains such as "www.google.com," "video.google.com," "appengine.google.com," and "investor.google.com". In the early days of the Internet a host commonly had a single sub-domain name. However, sub-domains now commonly map to multiple IP addresses and therefore multiple hosts. The domain naming system is not critical to ISC's host counting since the hosts are uniquely named and have a unique IP address. ISC identifies the sub-domain associated with each host for purposes of allocating hosts to countries.

[1] More details can be read at http://www.isc.org/index.pl?/ops/ds/.
[2] ISC also adjusts for "generic" ccTLD's, such as .com, .edu., and .org, that do not always have a country suffix.