

Online Appendix A “Small” Countries Included in Analysis

	Africa ¹	The Americas ¹	Asia ¹	Europe ¹	The Pacific ¹
	Algeria	Antigua and Barbado	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 d'Ivoire	Greenland	Israel	Hungary	Vanuatu
	Democratic Republic of the Congo	Grenada	Jordan	Iceland	
	Djibouti	Guadeloupe	Kazakhstan	Ireland	
	Egypt	Guatemala	Kuwait	Italy	
	Equatorial 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	Suriname	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 include 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 population 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	<i>Freedom in the World</i> , Freedom House (1999 - 2005 editions)
Literacy Rate	Literacy rate of population aged 15 and above, years 2000 to 2005.	Once	<i>The State of the World's Children 2008</i> , United Nations Children's 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)
Government Tax Receipts	Tax revenue as a fraction of GDP.	Annual/1998 - 2004	World Development Indicators
Corporate Tax Rate	Highest marginal corporate tax rate.	Annual/1998 - 2004	World Development Indicators
Telephone Employees	Number of telephone employees per fixed telephone line.	Annual/1998 - 2004	World Development Indicators
Gateway Capacity	Gateway capacity in gigabits per second.	Annual/1998 - 2004	OECD (2009), International Cable Protection Committee (http://www.iscpc.org/), and major subway cable consortium websites
IP Protection	Intellectual Property Rights component of Intellectual Property Rights Index	Once	Horst (2006)

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.

Online Appendix D Distinguishing Language-Specific and Non-Language Specific Content

Our model will accommodate non-language specific content if: 1) in each country, language-specific content is produced proportional to the fraction of the country's population using each language,³ and 2) in each country, language-specific and non-language specific content have the same marginal effect on adoption. Total content in country k at time t includes language-specific content across all languages and non-language specific content (Assumption 2 ensures that we can add these without weights):

$$(A1) \quad Content_{kt} = \sum_{j=1}^J LContent_{kjt} + NLContent_{kt}.$$

Assumption 1 implies:

$$(A2) \quad LContent_{kjt} = \frac{Users_{kj} LContent_{kt}}{Population_k} \text{ and } LContent_{kt} = \sum_{j=1}^J \frac{Users_{kj} LContent_{kt}}{Population_k}.$$

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.

³ This is the same assumption required when we allocate hosts to languages within countries.

Substituting Equation (A2) into (A1) and using the fact that $Population_k = \sum_{j=1}^J Users_{kj}$ we get:

$$(A3) \quad Content_{kt} = \sum_{j=1}^J \left(\frac{Users_{kj} [LContent_{kt} + NLContent_{kt}]}{Population_k} \right).$$

The term in brackets is total content for country k at time t so:

$$(A4) \quad Content_{kt} = \sum_{j=1}^J \frac{Users_{kj} Content_{kt}}{Population_k}.$$

This implies that we can measure country k 's content produced at time t in language j by allocating total content according to the fraction of the country's population using each language:

$$(A5) \quad Content_{kjt} = \frac{Users_{kj} Content_{kt}}{Population_k},$$

consistent with our procedure described on page 5 of the main text.

If Assumption 2 does not hold, the direction of bias due to the presence of non-language specific content will depend on the relative sensitivities. If adoption is more responsive to language-specific content then we will understate content's effect. If the opposite is true we will overstate it.