Making the 'Next Billion' Demand Access*

The Local-Content Effect of `google.co.za` in Setswana

Bastiaan Quast†

Abstract

Recent attempts to connect the current 'next billion' to the Internet in places such as sub-Saharan Africa have not met expectations. In places where Internet infrastructure has come online and prices have gone down, the expected consequent increase in uptake was not observed. Internet adoption in a certain language is a two-sided market with positive cross-side network effects. As a result of this, it is difficult to isolate the causal effect of one on the other. The exogenous introduction of the Setswana language interface on the South African Google Search website was a spillover of the development of that interface for the Botswanan Google website. This exogenous improvement in the accessibility of Setswana-language content has resulted in a substantial increase in the number of native Setswana speakers coming online and owning personal computers. This is turn has also led to increased usage of the Setswana language online. This adoption appears to also lead to improvements in employment.

1 Introduction

Internet uptake is a two-sided market, with users on one side and content creators on the other side. Positive cross-side network effects mean that increases in content leads to increases in user adoption and visa versa. This market exists separately for each language but for many indigenous languages this virtuous circle fails to start properly, keeping usage and content levels low.

With this study I seek to answer the question whether an increase in local language content, does indeed lead to an increase in uptake of Internet usage among native speakers of this language.

Because of the cross-side network effects in a two-sided market, any observed changes are inherently endogenous. I remedy this problem by using an exogenous shock in accessibility of Setswana language content in South Africa, namely the introduction of the Google Search interface in Setswana. I find that this leads to a strong increase in both the proportion of households reporting to have spent on Internet access in the last 30 days, as well as individuals owning a computer. This in turn has led to a large increase in the usage of the Setswana

*http://qua.st/internet-access

†http://qua.st | bquast@gmail.com | bastiaan.quast@graduateinstitute.ch | Maison de la paix, Geneva, Switzerland

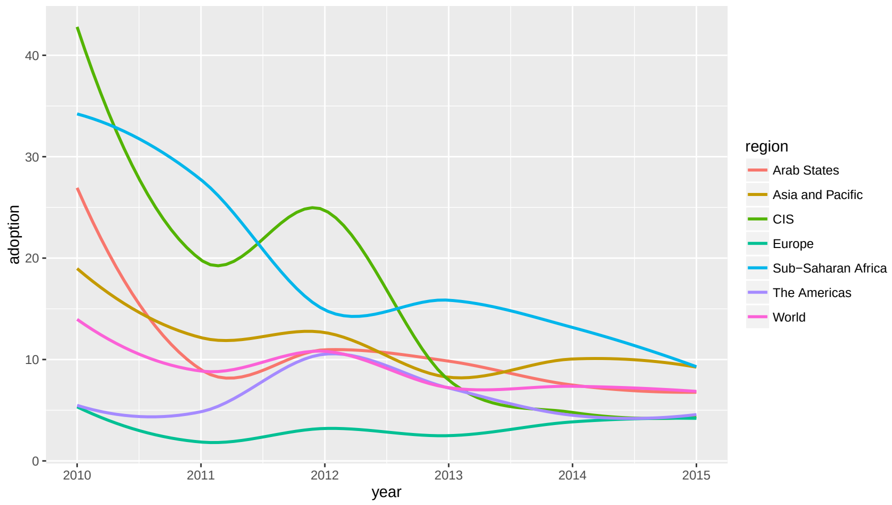
language online (in Google search queries). There also appears to be a strong improvement in employment status among individuals who spend on Internet or own a computer after the introduction of the interface. Suggesting that the expanded demographic of Setswana Internet users is benefiting from increased Internet adoption in terms of employability.

The term ‘Connecting the Next Billion’ was introduced in The Economist’s 2006 ‘End of Year Report’ (Standage, 2006), discussing the infrastructural requirements for connecting the second billion individuals to the Internet. Since then, close to 2 billion people are estimated to have been connected to the Internet, up from the just over one billion at the time of writing (Sanou, 2015). However, it seems increasingly unlikely that the current ‘Next Billion’ will be connected as easily as the previous ones.

In the period 2010-2014 the average annual growth of Internet bandwidth in sub-Saharan Africa was over fifty percent. This increased bandwidth also causes downward pressure on the cost of Internet access, which brought the sub-Saharan average cost of a 500MB prepaid Internet bundle down to around $10, increasingly putting it within range of the emerging middle classes. Yet, despite increased range and improved affordability, sub-Saharan Africa is showing stagnation in the growth of Internet connected individuals.

As is shown in Figure 1, growth of Internet usage in Sub-Saharan Africa is rapidly decreasing. Unlike in other regions in the figure, this observed stagnation is not a consequence of near market saturation, as adoption levels are still relatively low.

Figure 1: Internet Adoption



A crucial factor in Internet adoption by native speakers of a language is the interplay with content creators using that language. This dynamic is know as a two-sided market, which is characterised as having two different sides, which exhibit positive cross-side network effects (Parker and Van Alstyne, 2000a,b, 2005).

In the case of Internet adoption in a certain language, these sides are on the one hand the content creators, such as news websites and on the other hand the content consumers, or Internet users. Ideally, adoption should follow a virtuous circle, whereby the content offering encourages more users to come online, which in turn incentivises more content creation and so on (Rochet and Tirole, 2003, 2006). Unfortunately this virtuous circle sometimes fails to properly start for certain languages, this is especially important as it typically concerns communities who's linguistic characterisation can already hamper economic growth (Arcand and Grin, 2013). Herein also lies the difficulty with finding empirical evidence supporting these dynamics, since the process of adoption by users and content creators is inherently endogenous. With this paper I seek to empirically address the question if increased accessibility of content does indeed lead to an increase in Internet adoption.

Much research has been done on Internet and language with respect to the preservation of smaller languages, in particular indigenous languages. The focus here is often the preservation of the language, such as the below comment by Wikipedia founder Jimmy Wales (Forbes, 2010).

> The Web is a powerful tool for preserving languages that would otherwise be lost. We see this in a lot of the smaller European languages that have very active Wikipedia projects. For example, the Welsh Wikipedia is quite an active community and they have 27,000 articles and this is true even though virtually everyone who speaks Welsh also speaks English.

Unfortunately, this is less true for many indigenous languages outside of Europe. Increasingly language availability is also being considered as a method for improving demand for connectivity (Gandal, 2006; Pena-Lopez, 1999) as well as more generally in development economics as a whole (Arcand, 1996). In a recent paper,Viard and Economides (2014) use macro level connectivity data and a model whereby countries that share languages are used to isolate the effect of content on demand for connectivity, which they find to be positive and significant.

There are good reasons for striving for increased connectivity of remote populations. Jensen and Oster (2009) find that the introduction of cable television in Indian states has a pronounced positive effect on attitudes towards the oppression of women, violence against women, son preference, and as well as decreased fertility. Internet access can provide an in some ways similar window on the outside world, we might expect some of these things to also follow more widespread Internet adoption. Sinai and Waldfogel (2004) show that expanded Internet usage in cities can lead to less racism, an issue that can be of particular relevance to South Africa. The overcoming of ethnic polarisation can in turn be conducive to economic growth (Arcand, Guillaumont et al., 2000).

As mentioned above, Internet connectivity is a two-sided market, which makes it difficult to empirically isolate a causal effect of content availability on demand for Internet connectivity. This paper exploits an exogenous increase in the accessibility of content in the Setswana language in South Africa, in order to isolate the increase of Internet usage among native speakers. In 2010 Google collaborated with a team of Botswanan linguists (Otlogetswe, 2010) to make its Botswanan website (`google.co.bw`) available in the local language: 'Setswana'. In addition to being spoken in Botswana, there is also a sizeable population of Setswana speakers across the border in South Africa, where it is also one of the official state languages. This led to the Setswana-language interface also being introduced on the South-African Google website (google.co.za), as a spillover of the translation work originally performed for Google's Botswanan website. This introduction led to a large increase in the number of native Setswana speakers reporting to have spent some amount of money in the past 30 days on Internet access as well as increased computer ownership.

The Google Search interface represents a very small number of words on the Internet and it is not required to use a certain interface language in order to search for content in this language. Yet, the search page is in many cases the first website viewed by users and thereby has a substantial impact on the decision to further engage or not and if they chose to do so, in which language this will be. Asides from being able to understand the interface of the website, having this interface be in a local language also encourages usage of this local language, which in turn reveals more local language content.

In short, we can identify two main channels through which this promotes increased online engagement, which together constitute the theory of change. Firstly, the ability to read and understand the words of the interface increases the chance that a user continues to use the website and the Internet at large. Secondly, the visibility of local language content increases the likelihood of the user entering search queries in the local language and thereby finding more content in the local language.

The data used for this study comes from the South African National Income Dynamics Survey, provided by Southern Africa Labour and Development Research Unit (2008, 2012, 2013), the data is further discussed in section 2. After which section 3 discusses the methods employed in this study, specifically, the discussion of the identification strategy can be found is subsection 3.1 and the use of the Difference-in-Differences estimator in subsection 3.2. Further, I present the results of the estimation in section 4. Finally, I conclude in section 5.

2 Data

The National Income Dynamics Study (NIDS) collects data on a representative set of around ten thousand South African households across several time periods. The first wave was gathered in 2008, the second wave in 2010, and the third wave was gathered in 2012 (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013).

In addition to this, in may 2016 a fourth wave of data has also been partially published, which I used to relate the expanded demographic of Setswana speaking Internet users to improvements in employment levels.

The dataset contains an extensive household questionnaire, which contains detailed information on income and expenditure. In particular, it breaks household expenditure down into many forms of food and non-food expenditure, one of which is household expenditure on Internet access in the last 30 days. In addition to this, the household income is calculated and imputed with other income such as home ownership. The individual (adult) questionnaires also contain information on linguistic skills in both English and in the interviewees native language, as well as a series of variables relating to communication technology ownership and utilisation, in particular computer ownership. In table Table 4 an overview of the dependent variables, broken down by wave and native language is presented.

Table 1: Dependent Variable Descriptive Statistics

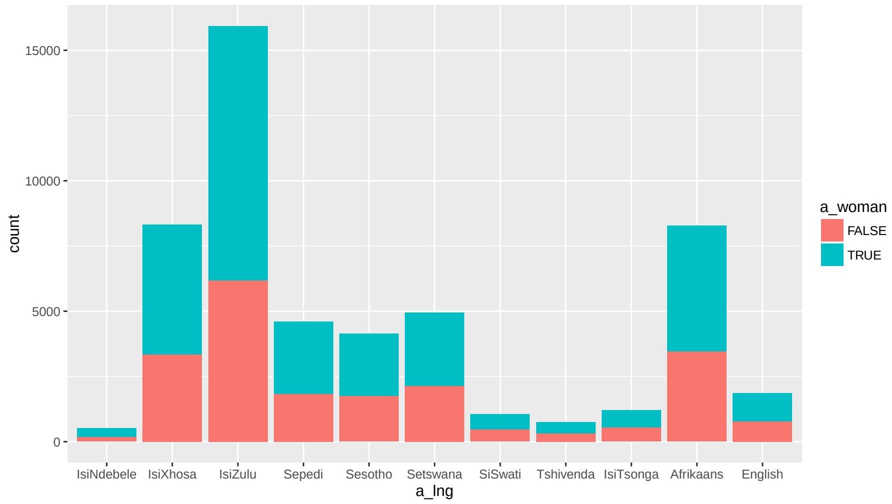
	Wave	Setswana	Other
Computer Ownership	1	0.0351	0.0567
	2	0.0359	0.0417
	3	0.0711	0.0632
Household Internet Expenditure	1	0.0068	0.0162
	2	0.0037	0.0224
	3	0.0106	0.0140

I use both household expenditure on Internet access and computer ownership as dependent variables. Household expenditure includes a variety of way in which this expense can be made, including paying for a fixed-line subscription, a mobile Internet subscription, as well using Internet in an Internet cafe. More than 99% of South Africa is covered by mobile Internet networks, this figure did not change during the time periods used in this study (Union, 2015). Computer ownership is recorded in the individual adult questionnaire, which provides me with more observations, unlike expenditure in the last 30 days, this shows a more long-term investment.

In addition to the variables of interest, I include a number of relevant covariates, such as household income and education levels. Furthermore, I include information on English and native language reading and writing skills. Around 45% of the individuals report being able to read and write English fluently, whereas around 55% report being able to do so in their native language.

In Figure 2 we can see the number of native speakers for each language in the dataset, coloured by gender. The dataset contains a total of 51,612 observations (adult individuals), of which 2,806 are female native Setswana speakers and 2,140 are male native Setswana speakers.

Figure 2: Native Language and Gender



3 Empirical Methodology

With this paper I aim to answer the question whether increased content or accessibility of content leads to an increase in demand. This section begins with a discussion of the identification strategy employed, followed by a explanation of the estimator used to operationalise this.

3.1 Identification Strategy

This paper exploits the introduction of the Setswana interface language to Google Search in South Africa as a spillover of the development of that interface for the Botswanan Google Search website. By comparing the number of native Setswana speakers in South Africa being Internet users, with the number of South Africans with a different native language around the same time, I isolate the effect of this introduction.

The Setswana language interface was first developed for the Botswanan Google Search website (`google.co.bw`). As such, the introduction of Setswana to the South African Google Search (`google.co.za`) was a spillover effect of that development. This allows me to rule out any possible endogeneity issues that might otherwise arise in contexts such as these. For instance, the Afrikaans language is almost solely spoken in South Africa. When we observe that the introduction of the Afrikaans Google Search interface occurs around the same time as a growth in the number of native Afrikaans Internet users, it will be hard to isolate the effect from the introduction from its cause (since an increase in native Afrikaans Internet users would be a good reason to introduce it as an interface language).

Substantial numbers of Setswana speakers exist in Botswana, South Africa, Zimbabwe, and to some extend Namibia. However, the language is most important in Botswana, where it is spoken by approximately 80% of all people, and where it is the only official language other than English. As such, it is also the place where most linguistic work on the Setswana language takes place. The Setswana Google Search interface was also developed at the University of Botswana by prof. Otlogetswe.

It is worth noting that it is very common to not personally own a computer, therefore 'paying for Internet access' also includes a lot of people who use the Internet in other locations such as Internet cafe's.

In addition to using the propensity to spend on Internet (in the last thirty days), I also use the propensity to own a computer as a dependent variable.

3.2 Regression Specification

As mentioned in the above section, I compare the change in the level of Internet users among native Setswana speakers in South Africa, with that of native speakers of other language in South Africa around the introduction of the Setswana interface to the South-African Google Search, after the second wave of the NIDS. For this I use a Difference-in-Differences estimator (Abadie, 2005; Imbens and Jeffrey M. Wooldridge, 2009) using a native-Setswana speaker dummy variable (`setswana`), interacted with an event dummy variable (`event`). The former is `TRUE` when the native language of the individual (`a_lng`) is `Setswana` and FALSE otherwise. The latter is FALSE for data collected prior to the introduction of the Setswana interface language (late 2010, here wave 1 and 2) and TRUE after this introduction (here wave 3). The model than takes the form as described in equation (1).

$$y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it} \tag{1}$$

Where α_i represents the individual fixed effects, λ_t represent the time fixed effects, and X_{it} are the time varying covariates. The ϵ_{it} is the error term. Finally the term of interest is D_{it} which represents the treatment effect.

The h_nfnet variable is recorded at a household level, as such a standard error correction needs to be applied (White, 1980), the model with standard error corrections is reported in the appendix.

Lastly, the dependent variables are both logical or binary variables, as such, normally a model such as logit should be used. However, since I am using Difference-in-Differences, this model would be undefined (Jeffrey M Wooldridge, 2010), I therefore use a standard linear model.

4 Results

In the base model, I use an interaction of the event dummy and setswana dummy in order to isolate the effect on the explanandum, a dummy variable describing household expenditure on Internet in the last thirty days or not (Internet_expenditure, household non-food Internet). The results of this estimation are presented in Table 6.

I find that the interaction term of the event dummy (event) and the native Setswana speaker dummy (setswana) is positive and highly significant, with a p-value around 0.0018. Both the individual dummy variables (event and setswana) yield significant negative parameter estimates.

In addition to this, the covariates included in the estimation are also highly significant. The highest education level of the individual (best_edu) and the household income (hhincome) are both positive and significant. The parameter estimate of woman here is negative but not at all significant, this is unsurprising as I use Internet expenditure at a household level. Most women live in a household which includes men and visa versa, suggesting that this effect cannot be isolated in this estimation. I further investigate this issue in a separate estimation discussed below. The variables describing linguistic skills in reading and writing in both English and the native language do yield many significant results, though lower levels of English writing skill seems to be correlated with a lower propensity to use the Internet (a_edlitwrten for levels 2 and 3, but not the very lowest: 4).

In an alternative formulation, I include the native language variable as a categorical variable (language), interacted with the event dummy. In this estimation I only find significantly positive results for Setswana and Venda (as small language from the region bordering Zimbabwe), and a significantly negative effect for the language Afrikaans.

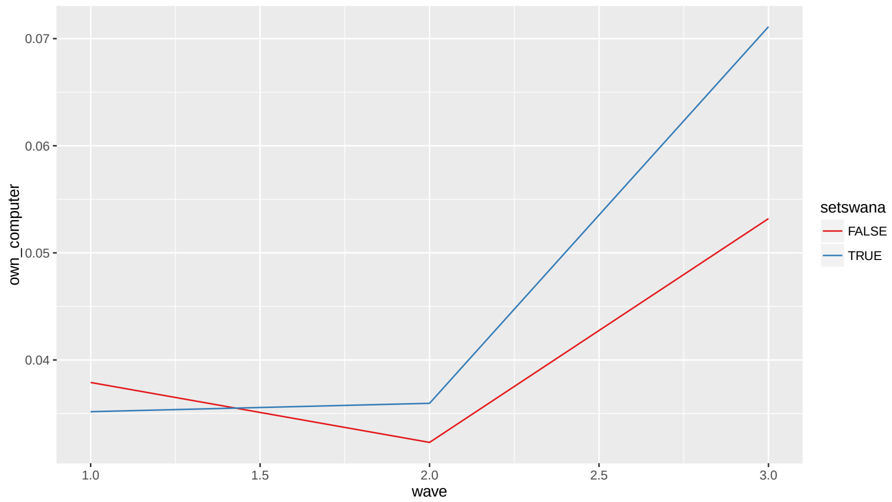
Furthermore, I also use a variant of the base model, in which the propensity of adults (own_computer) to own a computer is used as an explanandum. This is of particular relevance, as the explanandum here (own_computer) differs from

Table 2: Internet Access and Computer Ownership

	Internet	(P > |t|)	Computer	(P > |t|)
event * setswana	**0.012**	0.00	**0.024**	0.00
event	**-0.012**	0.00	-0.054	0.01
setswana	**-0.014**	0.00	**-0.015**	0.00
income	**0.000**	0.00	**0.000**	0.00
woman	-0.001	0.23	**-0.023**	0.00
education	**0.001**	0.00	**0.006**	0.00
Intercept	-0.000	0.64	**0.008**	0.02
Observations	47665		46464	

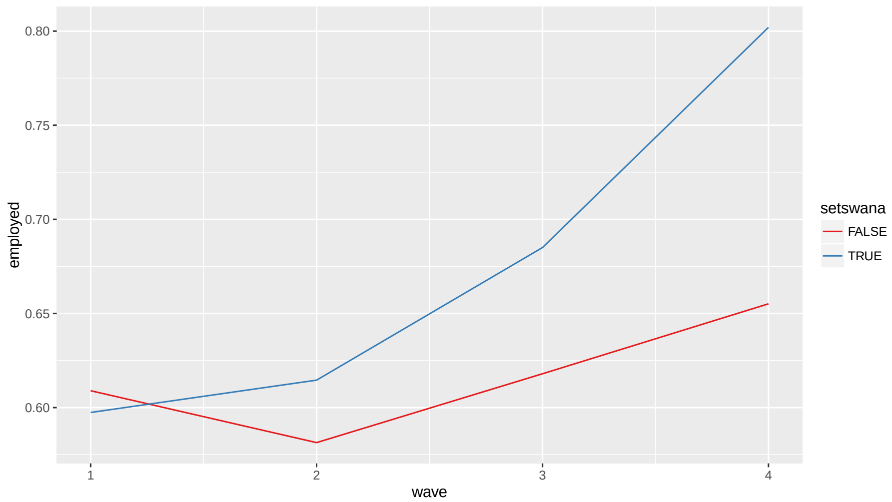
the base model's explanandum in two ways. Firstly, it does not include expenditure on Internet in ways such as Internet cafes, but focusses on actual ownership, signalling a more long-term investment and interest. Secondly, the `Internet_expenditure` variable is at a household level, whereas the `own_computer` variable is at the level of an individual adult. The results from this estimation are included in Table 5. This form of the estimation yields similar results to those estimated in the base model. Firstly I find that the variable of interest, the interaction term between the event and the language dummy (`eventTRUE:setswanaTRUE`) is positive and highly significant, with a p-value smaller than `0.001`. The individual dummy variables (`event` and `setswana`) again are significant and negative with the former's p-value smaller than `0.01` and the latter's smaller than `0.001`. In terms of the linguistic skill, I find that the lower levels of English reading as well as English writing are correlated with lower propensities of computer ownership. Similar to Internet expenditure model, household income (`hhincome`) and highest level of education (`best_edu`) are both positive and highly significant (p-value: `~0`). However, unlike in the household Internet expenditure model, the sex of the individual here is highly significant, specifically, parameter estimate of `woman` is negative and highly significant (p-value: `~0`). As mentioned above, this variable is difficult to interpret when using a household-level variable as an explanandum, however, here, the computer ownership variable is at an individual level, which makes the coefficient more interpretable.

Figure 3: Computer Ownership Setswana



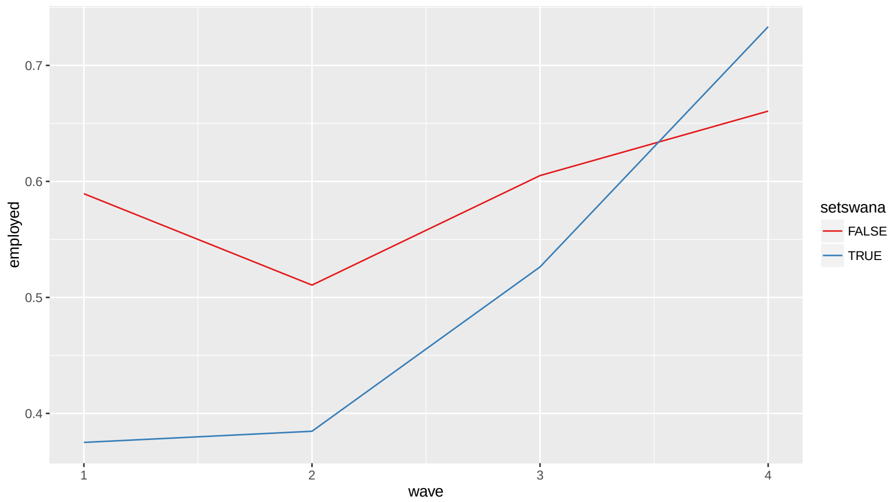
In the below figure, I plot the employment status of individuals that own a computer in the first wave after the introduction of the Setswana-language interface (wave 3). The figure shows that there is a sharp uptick in the proportion of employed individuals among Setswana speakers.

Figure 4: Employment for Individuals who Own a Computer in Wave 3



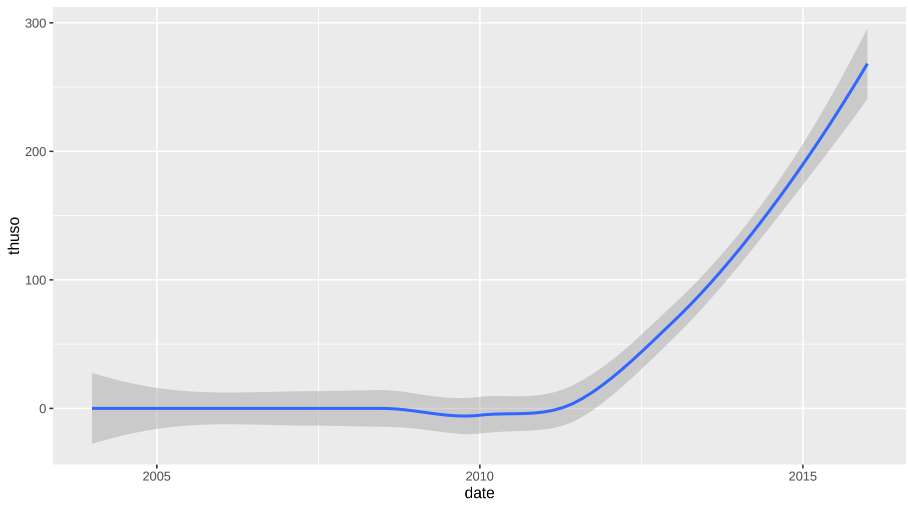
Similarly, the individuals that lived in a household that spent on Internet access in the last 30 days also saw a strong increase in the proportion of employed individuals, overtaking the proportion for the rest of the population.

Figure 5: Employment for Individuals with Internet Expenditure in Wave 3



This last effect is also visible in Figure 6, which illustrates how usage of the Setswana word 'thuso', meaning 'help' in search queries, was zero prior to the introduction of the Setswana search interface and became widespread thereafter.

Figure 6: Usage of Setswana Words on Google.co.za



5 Conclusions and Limitations

In conclusion, despite recent advances in the reach, speed, and affordability of Internet connectivity in sub-Saharan Africa, actual uptake has been stagnant. Internet adoption is a two-sided market, as a result, one aspect of this is that users benefit from cross-side network effects from content. Unfortunately, in certain languages, a sufficient amount of content is not always created, as a result, Internet adoption among native speakers of these languages can lag behind.

This paper demonstrates that this failure can in part be attributed to the dynamics of a two-sided market, whereby the vicious circle of few users and little content perpetuates a situation of low levels of adoption. Due to the endogenous nature of two-sided markets, there are few methods of isolating a causal effect.

I exploit the introduction of the Setswana-language interface on `google.co.za`, as a spillover of the development of this interface for `google.co.bw`, I find that it leads to a substantial increase in Internet usage and computer ownership among native Setswana speakers.

Furthermore, when comparing the Setswana speakers that own a computer or spend on Internet access after the event, with non Setswana speakers, we see a marked increase in the proportion of employed individuals over that among the rest of the population. This increase is even stronger in wave 4 of the data, suggesting that it takes some time for the effect to fully materialise and that it is persistent.

This increase of Internet usage among the Setswana speaking population as a result of the newly introduced interface language on `google.co.za`, suggests that there is a serious lack in the availability of local content in many indigenous African languages, which serves as an impediment to further Internet adoption.

This suggests that the effect is unlikely to be ephemeral in nature, since computer ownership constitutes a more long-term investment in Internet access.

As I discussed in the results section, the increase in computer ownership for Setswana speakers between wave 2 and wave 3, was 115%, compared to 70% for the rest of the population, or a 50% greater increase. The effect on Internet expenditure in the household, which includes expenditure in Internet cafes etc. is even greater. The proportion of Setswana households that spent on Internet in the last 30 days increased by 217%, whereas for the rest of the population it fell be 22%.

References

Abadie, Alberto

2005 "Semiparametric difference-in-differences estimators", *The Review of Economic Studies*, 72, 1, pp. 1–19.

Arcand, Jean-Louis

1996 "Development economics and language: the earnest search for a mirage?", *International Journal of the Sociology of Language*, 121, 1, 243Ã¢ÂÂ266.

Arcand, Jean-Louis and Francois Grin

2013 "Language in Economic Development: Is English Special and is Linguistic Fragmentation Bad?", in *English and development: Policy, pedagogy and globalization*, ed. by Elizabeth J. Erling, Multilingual Matters, chap. 11, pp. 243–266.

Arcand, Jean-Louis, Patrick Guillaumont, Sylviane Guillaumont Jeanneney et al.

2000 *Ethnicity, communication and growth*, University of Oxford, Department of Economics, Centre for the Study of African Economies.

Croissant, Yves

2013 "pglm: Panel Generalized Linear Model", *R package version 0.1-2*, `http://CRAN.%20R-project.org/package=%20pglm`.

Croissant, Yves, Giovanni Millo et al.

2008 "Panel data econometrics in R: The plm package", *Journal of Statistical Software*, 27, 2, pp. 1–43.

Forbes

2010 "Jimmy Wales: The Wiki World", *Forbes*, `http://www.forbes.com/2010/06/15/forbes-india-jimmy-wales-the-wiki-world-opinions-ideas-10-wales.html`.

Gandal, Neil

2006 "Native language and Internet usage", *International journal of the sociology of language*, 2006, 182, pp. 25–40.

Git Team

2016 *Git: Software Code Manager*, 137 Montague ST STE 380, Brooklyn, NY 11201-3548, `http://www.git-scm.org/`.

Hoekwater, Taco, Hartmut Henkel and Hans Hagen

2016 *LuaTeX*, `http://www.luatex.org/`.

Imbens, Guido W. and Jeffrey M. Wooldridge

2009 "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47, 1 (Mar. 2009), pp. 5–86, DOI: `10.1257/jel.47.1.5`, `http://www.aeaweb.org/articles/?doi=10.1257/jel.47.1.5`.

Jensen, Robert and Emily Oster

2009 "The power of TV: Cable television and women's status in India", *The Quarterly Journal of Economics*, pp. 1057–1094.

Knuth, Donald Ervin

1984 "Literate programming", *The Computer Journal*, 27, 2, pp. 97–111.

Lamport, Leslie

1985 *I1 (\ LaTeX)—A Document*, pub-AW, vol. 410.

Leisch, Friedrich

2002 "Sweave: Dynamic generation of statistical reports using literate data analysis", in *Compstat*, Springer, pp. 575–580.

LyX Team

2016 *LyX*, Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA, `http://www.lyx.org/`.

Otlogetswe, Thapelo J.

2010 "Setswana Google is here!", *T.J. Otlogetswe Blog*, `http://otlogetswe.com/2010/08/13/setswana-google-here/`.

Parker, Geoffrey G and Marshall W Van Alstyne

2000a "Information complements, substitutes, and strategic product design", in *Proceedings of the twenty first international conference on Information systems*, Association for Information Systems, pp. 13–15.

2000b "Internetwork externalities and free information goods", in *Proceedings of the 2nd ACM Conference on Electronic Commerce*, ACM, pp. 107–116.

2005 "Two-sided network effects: A theory of information product design", *Management science*, 51, 10, pp. 1494–1504.

Pena-Lopez, Ismael

1999 "Challenges to the Network: Internet for Development".

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Rochet, Jean-Charles and Jean Tirole

2003 "Platform competition in two-sided markets", *Journal of the European Economic Association*, 1, 4, pp. 990–1029.

2006 "Two-sided markets: a progress report", *The RAND journal of economics*, 37, 3, pp. 645–667.

Sanou, Brahima

2015 "The World in 2015: ICT facts and figures", *International Telecommunications Union.*

Sinai, Todd and Joel Waldfogel

2004 "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?", *Journal of Urban Economics*, 56, 1, pp. 1–24.

Southern Africa Labour and Development Research Unit

2008 *National Income Dynamics Study, Wave 1*, version 5.3, http://www.nids.uct.ac.za/home/.

2012 *National Income Dynamics Study, Wave 2*, version 2.3, http://www.nids.uct.ac.za/home/.

2013 *National Income Dynamics Study, Wave 3*, version 1.3, http://www.nids.uct.ac.za/home/.

Standage, Tom

2006 "Connecting the next billion", *The Economist-The World in 2006*, p. 117, http://www.economist.com/node/5134746.

Union, International Telecommunication

2015 *The World in 2011: ICT Facts and Figures*, ITU.

Venables, William N and Brain D. Ripley

2013 *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

Viard, V Brian and Nicholas Economides

2014 "The Effect of Content on Global Internet Adoption and the Global Ã¢ÂÂDigital DivideÃ¢ÂÂ", *Management Science*, 61, 3, pp. 665–687.

White, Halbert

1980 "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Wooldridge, Jeffrey M

2010 *Econometric analysis of cross section and panel data*, MIT press.

Xie, Yihui

2015 *Dynamic Documents with R and knitr*, Chapman and Hall/CRC, vol. 29, ISBN: 978-1498716963, http://yihui.name/knitr/.

Zeileis, Achim

2004 "Econometric Computing with HC and HAC Covariance Matrix Estimators", *Journal of Statistical Software*, 11, 10, pp. 1–17, http://www.jstatsoft.org/v11/i10/.

2006 "Object-Oriented Computation of Sandwich Estimators", *Journal of Statistical Software*, 16, 9, pp. 1–16, http://www.jstatsoft.org/v16/i09/.

Zeileis, Achim and Torsten Hothorn

2002 "Diagnostic Checking in Regression Relationships", *R News*, 2, 3, pp. 7–10, http://CRAN.R-project.org/doc/Rnews/.

6 Clustering

Table 3: Clustering

```r
library(plm)     # panel linear model estimation
library(lmtest)  # Standard Error corrections
library(broom)   # output formatting using tidy()

# specify panel model
plm4_3 <- formula(as.numeric(h_nfnet)  ~ interface_intro*setswana  +
                                         factor(a_edlitrden)   +
                                         factor(a_edlitwrten)  +
                                         factor(a_edlitrdhm)   +
                                         factor(a_edlitwrthm)  +
                                         a_woman               +
                                         hhincome              +
                                         best_edu               )
# estimate
plm4_3e <- plm(plm4_3, data=pNIDS, model='within')

# correct errors
tidy( coeftest(plm4_3e, vcov=vcovHC(plm4_3e,
                                    type="HC0",
                                    cluster="group")) )
```

term	estimate	std.error	statistic	p.value
interface_introTRUE	-0.0019402	0.0012392	-1.5657830	0.1174144
setswanaTRUE	0.0027958	0.0137038	0.2040204	0.8383396
factor(a_edlitrden)2	0.0061294	0.0039180	1.5643993	0.1177388
factor(a_edlitrden)3	0.0028083	0.0042618	0.6589578	0.5099301
factor(a_edlitrden)4	0.0001637	0.0044802	0.0365363	0.9708551
factor(a_edlitwrten)2	-0.0088583	0.0039162	-2.2619681	0.0237095
factor(a_edlitwrten)3	-0.0080226	0.0042172	-1.9023800	0.0571351
factor(a_edlitwrten)4	-0.0060115	0.0046454	-1.2940715	0.1956549
factor(a_edlitrdhm)2	-0.0029617	0.0043786	-0.6764141	0.4987852
factor(a_edlitrdhm)3	-0.0035411	0.0061579	-0.5750552	0.5652601
factor(a_edlitrdhm)4	-0.0056185	0.0068075	-0.8253295	0.4091939
factor(a_edlitwrthm)2	0.0020764	0.0043717	0.4749531	0.6348253
factor(a_edlitwrthm)3	0.0045419	0.0063446	0.7158756	0.4740761
factor(a_edlitwrthm)4	0.0077089	0.0074611	1.0332094	0.3015178
a_womanTRUE	-0.0012994	0.0012422	-1.0461001	0.2955268
hhincome	0.0000007	0.0000003	2.4693445	0.0135439
best_edu	0.0001411	0.0003131	0.4507013	0.6522095
interface_introTRUE:setswanaTRUE	0.0070425	0.0034066	2.0673056	0.0387175

7 Dependent Variable Breakdown

Table 4: Descriptive statistics on Ownership and Expenditure

a_lng	wave	a_owncom	h_nfnet
IsiNdebele	1	0.0331126	0.0000000
IsiNdebele	2	0.0270270	0.0284091
IsiNdebele	3	0.0333333	0.0000000
IsiXhosa	1	0.0112269	0.0012043
IsiXhosa	2	0.0186335	0.0054517
IsiXhosa	3	0.0345508	0.0019756
IsiZulu	1	0.0132693	0.0013730
IsiZulu	2	0.0127744	0.0086366
IsiZulu	3	0.0252420	0.0044928
Sepedi	1	0.0265273	0.0048309
Sepedi	2	0.0226818	0.0021994
Sepedi	3	0.0718697	0.0050477
Sesotho	1	0.0366044	0.0062598
Sesotho	2	0.0457010	0.0213640
Sesotho	3	0.0949535	0.0113032
Setswana	1	0.0351724	0.0068681
Setswana	2	0.0359537	0.0037783
Setswana	3	0.0711086	0.0106264
SiSwati	1	0.0441640	0.0000000
SiSwati	2	0.0612813	0.0091185
SiSwati	3	0.0458221	0.0134771
Tshivenda	1	0.0334928	0.0000000
Tshivenda	2	0.0000000	0.5441176
Tshivenda	3	0.0225080	0.0000000
IsiTsonga	1	0.0235294	0.0000000
IsiTsonga	2	0.0118203	0.0929368
IsiTsonga	3	0.0397196	0.0023529
Afrikaans	1	0.1345441	0.0465116
Afrikaans	2	0.0904605	0.0424710
Afrikaans	3	0.1106225	0.0399458
English	1	0.2969374	0.1016043
English	2	0.3234127	0.1070707
English	3	0.3156934	0.1023766

8 Covariate Descriptive Statistics

Figure 7: Household Income

```
ggplot(adulthh, aes(x=hhincome, fill=a_lng )) +
    stat_bin(bins=50)
```

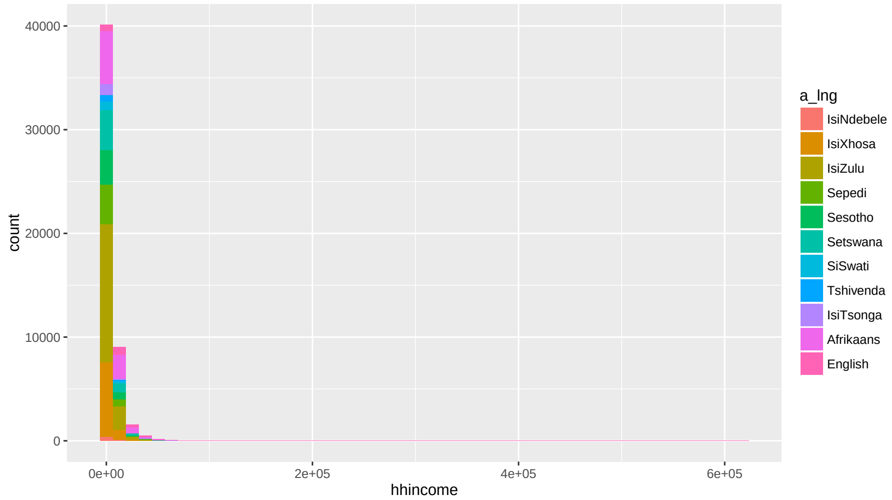
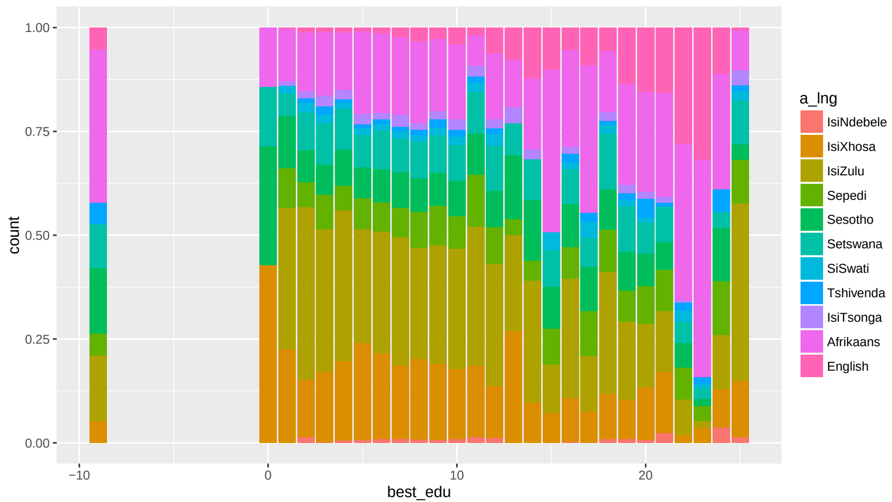


Figure 8: Years of Education

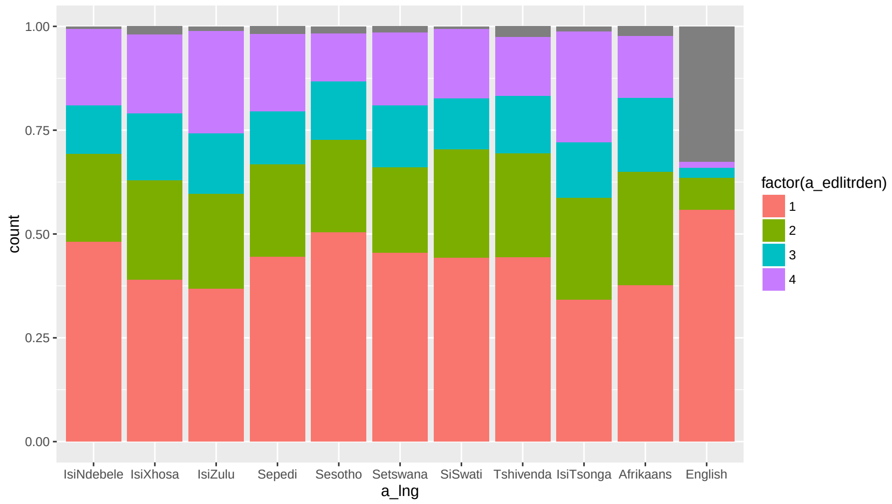
```
ggplot(adulthh, aes(x=best_edu, fill=a_lng )) +
    geom_bar(position = 'fill')
```



The Figure 9 describes the skill of individuals in reading the English language, where 1 the best and 4 is the worst, grey values are `NA`.

Figure 9: English Language Reading Skills

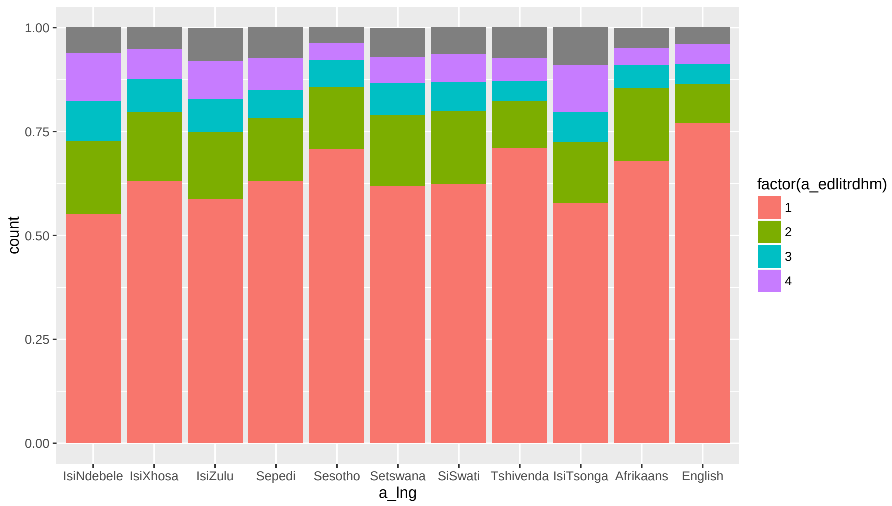
```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrden)) ) +
        geom_bar(position = 'fill')
```



The Figure 10 does the same but with regards to the native language.

Figure 10: Native Language Reading Skills

```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrdhm)) ) +
        geom_bar(position = 'fill')
```



9 Original Estimates

Table 5: Computer Ownership

```
lm(a_owncom ~ interface_intro*setswana_logical +
              factor(a_edlitrden)               +
              factor(a_edlitwrten)              +
              factor(a_edlitrdhm)               +
              factor(a_edlitwrthm)              +
              a_woman                           +
              hhincome                          +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0075393	0.0031352	2.4047103	0.0161891
interface_introTRUE	-0.0053820	0.0020917	-2.5729824	0.0100856
setswana_logicalTRUE	-0.0147502	0.0041653	-3.5411916	0.0003987
factor(a_edlitrden)2	-0.0306749	0.0069077	-4.4406931	0.0000090
factor(a_edlitrden)3	-0.0310090	0.0090330	-3.4328617	0.0005978
factor(a_edlitrden)4	-0.0389174	0.0116674	-3.3355679	0.0008519
factor(a_edlitwrten)2	-0.0173238	0.0068925	-2.5134122	0.0119602
factor(a_edlitwrten)3	-0.0210640	0.0089384	-2.3565895	0.0184476
factor(a_edlitwrten)4	-0.0187291	0.0114540	-1.6351572	0.1020227
factor(a_edlitrdhm)2	-0.0017925	0.0063219	-0.2835349	0.7767681
factor(a_edlitrdhm)3	-0.0041655	0.0088001	-0.4733526	0.6359638
factor(a_edlitrdhm)4	-0.0267964	0.0118797	-2.2556443	0.0240974
factor(a_edlitwrthm)2	0.0010896	0.0063817	0.1707311	0.8644360
factor(a_edlitwrthm)3	-0.0020020	0.0088176	-0.2270503	0.8203856
factor(a_edlitwrthm)4	-0.0357886	0.0118921	-3.0094517	0.0026186
a_womanTRUE	-0.0230335	0.0019598	-11.7530143	0.0000000
hhincome	0.0000058	0.0000001	56.8794124	0.0000000
best_edu	0.0058419	0.0002002	29.1761771	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0238052	0.0066799	3.5637169	0.0003660

Table 6: Living in Household that Spent on Internet (last 30 days)

```r
lm(h_nfnet ~  interface_intro*setswana_logical +
              factor(a_edlitrden)                +
              factor(a_edlitwrten)               +
              factor(a_edlitrdhm)                +
              factor(a_edlitwrthm)               +
              a_woman                            +
              hhincome                           +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0008565	0.0018351	-0.4667340	0.6406924
interface_introTRUE	-0.0117376	0.0012170	-9.6449716	0.0000000
setswana_logicalTRUE	-0.0135286	0.0024255	-5.5777454	0.0000000
factor(a_edlitrden)2	0.0015879	0.0040272	0.3942904	0.6933684
factor(a_edlitrden)3	0.0002543	0.0052755	0.0481963	0.9615600
factor(a_edlitrden)4	-0.0036574	0.0068160	-0.5365861	0.5915561
factor(a_edlitwrten)2	-0.0101933	0.0040178	-2.5369998	0.0111839
factor(a_edlitwrten)3	-0.0107676	0.0052214	-2.0621916	0.0391951
factor(a_edlitwrten)4	-0.0071685	0.0066937	-1.0709421	0.2842010
factor(a_edlitrdhm)2	-0.0025418	0.0036848	-0.6898084	0.4903181
factor(a_edlitrdhm)3	-0.0028899	0.0051291	-0.5634293	0.5731453
factor(a_edlitrdhm)4	-0.0079926	0.0069070	-1.1571655	0.2472107
factor(a_edlitwrthm)2	0.0008117	0.0037229	0.2180373	0.8274010
factor(a_edlitwrthm)3	0.0013993	0.0051372	0.2723915	0.7853222
factor(a_edlitwrthm)4	-0.0069174	0.0069137	-1.0005354	0.3170567
a_womanTRUE	-0.0013881	0.0011452	-1.2121278	0.2254696
hhincome	0.0000027	0.0000001	46.2987087	0.0000000
best_edu	0.0013582	0.0001164	11.6710813	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0119717	0.0038666	3.0961443	0.0019617

10 Software

The estimation is primarily performed using R (R Core Team, 2016), specifically using `lm()` and `glm()` functions included in the `stats` package (Venables and Ripley, 2013). Additionally, I make the of the `plm()` and `pglm()` functions which are available in packages by the same names (Croissant, 2013; Croissant and Millo, 2008). Standard error corrections are computed using the `lmtest` and `sandwich` packages Zeileis (2004, 2006); Zeileis and Hothorn (2002).

In order to make the result as easily reproducible as possible, this research and writing in the article has been done exclusively using open-source software such as R (R Core Team, 2016). This document is written and LyX (LyX Team, 2016) in the LaTeX(Lamport, 1985) language and compiled using the LuaTeX implementation(Hoekwater et al., 2016). The integration of R code and output in the document is performed using a process call literate programming Knuth (1984) using the knitr implementation Xie (2015) of the Sweave framework (Leisch, 2002).

All changes are logged using the version control system Git (Git Team, 2016) and publicly available on GitHub at https://github.com/bquast/Making-Next-Billion-Demand-Access/[1]

[1]The repository can by cloned to a local computer by entering in following command in a terminal (with Git installed):
`git clone https://github.com/bquast/Making-Next-Billion-Demand-Access.git`