Communication Effort and the Cost of Language: Evidence from Stack Overflow

Jacopo Bregolin[*]
University of Liverpool

April 2022

Abstract

The transmission of information is crucial for productivity and growth, but language differences may limit its effectiveness. In this paper, I empirically investigate how the exogenous cost of language affects communication effort, and the trade-off faced by knowledge platforms in implementing their website in multiple languages. I exploit the introduction of websites for languages different from English on a question-and-answering platform and compare the behavior of non-English speaking users before and after the introduction. Results show that the quality of communication improves by more than 24% when writers use their first language, rather than English, and answers are 7% more likely to solve the questioner's problem, a 20% increase from the baseline. In addition, the size of the effect increases when the sender is more incentivized and when the questioner's effort is higher. With the introduction of other languages, the community size increases, but the quality of the contribution of the new joiners is lower. Finally, information is more dispersed. These results show that the platform should adopt multiple languages to maximise the quality and quantity of the information collected unless the size of the communities using those languages does not justify the cost.

JEL Codes: D82, D83, L17, L86, M21, M54, Z13
Keywords: Cost of language, information transmission, knowledge platforms

[*] email: jacopo.bregolin@liverpool.ac.uk. This paper extends the second chapter of my PhD dissertation. I am extremely indebted to my advisor, Milo Bianchi, co-advisor, Daniel L. Chen, and committee members Jacques Crémer and Daniel Ershov. Their support and guidance have been invaluable. I also thank April Franco, Astrid Hopfensitz, Jakub Lonsky, and Marta Troya-Martinez for helpful discussions and feedback. I thank participants at the University of Liverpool Management School internal seminar and Digital Economy Network workshop (Norwich) for useful comments.

1 Introduction

Complex languages and the transmission of information have been identified as crucial factors for human evolution, potentially being the main source of our differentiation from the animal world (Diamond 1991). The sharing of knowledge allows people to take advantage of each others' human capital investments, speeding up learning and productivity. Nevertheless, information transmission may be limited in several ways. On the one hand, communication may be affected by the incentives of the information holder. On the other hand, exogenous cognitive boundaries may constrain our ability to share information.[1]

In this paper, I study to what extent the use of a foreign language affects effort choices in communication, and whether this depends on incentives and reciprocal effort. I then compare the advantages and disadvantages of reducing the cost of language by decentralizing the language used. This is a major concern in all contexts where individuals do not share the same language, but organizations or institutions still want to maximize information sharing (Chen, Geluykens, and Choi 2006, Crémer et al. 2007, Ginsburgh and Weber 2011, Tenzer, Pudelko, and Harzing 2014). In the digital era, a leading example is provided by knowledge platforms that aim to aggregate information, like Wikipedia or Stack Overflow. These platforms aim to be global and face the challenging choice of using one or multiple languages.

To study this trade-off, I use data from Stack Overflow, a question-and-answer website on topics related to computer programming, and exploit the staggered introduction of versions of the website that use languages different from English. This natural experiment allows me to measure the effort choices of non-English speaking users before and after the introduction of the new site, that is before and after they were able to use their native language in addition to English.

The paper shows that users increase their communication effort by 24% when speaking in their native language, and are 7% more likely to provide satisfactory information. Incentive alignment and reciprocal effort increase the effect. In addition, the availability of multiple languages increases the number of contributors, but new contributors provide, on average, lower quality contributions. Finally, information gets more dispersed and potentially inefficiently duplicated.

The study of communication effort choices is particularly relevant in the context of Stack Overflow. Question-and-answer websites' success is strictly based on the quantity

[1] The economic literature has theoretically investigated both of these constraints but generally focuses on one or the other. On one side, the literature has looked at incentives and strategic information transmission, either without costs of effort in communication, i.e. the *Cheap Talk* literature (Crawford and Sobel 1982, Austen-Smith and Banks 2000, Asher and Lascarides 2013, Sobel 2013), or with strategic choice of effort (Dewatripont and Tirole 2005) as in the signalling literature (Spence (1973), Gambetta (2011)). On the other sides, the team-theory literature (Marschak and Radner (1972)) has focused on environments where incentives are perfectly aligned, but exogenous constraints affect the ability to communicate, for instance bounded cognitive abilities or costs in information processing (Arrow 1974, Bolton and Dewatripont 1994, Crémer, Garicano, and Prat 2007, Blume and Board 2013, Blume 2018, Dilmé 2018). In this paper, I put together the two strands and look at the interaction between incentives and exogenous costs.

and quality of the information provided: the platform has then all incentives to reduce the barriers to participation and to the provision of effort in communication.[2]

Should the platform have a unique website in English, or should it implement several websites in different languages? The optimal strategy is not trivial.[3] By allowing users to communicate in languages different from English, the platform reduces communication costs but segregates communities.

To illustrate the trade-off, I present a simple theoretical framework of communication between two agents, Bob and Alice, where Bob needs some information to achieve some task, and Alice may provide it. Alice decides her communication effort based on how much she internalizes Bob's utility, Bob's effort, her cost of language, and her expertise. The framework underlines two aspects. First, Alice's effort and participation decision depend on her cost of using the language available, potentially high if it is not her native language. Second, under high cost of language, she would participate only if she has high expertise. It follows that, for certain levels of expertise and cost, the availability of her native language would 1) increase her effort if she has high expertise, or 2) make her participate if she has low expertise, as she would not have participated otherwise. Overall, the framework suggests that introducing multiple languages would increase the quality of contributions and the community size, but new contributors would provide lower-quality content.

I investigate empirically this trade-off by observing users' communication efforts before and after their native language became available. Stack Overflow was created in 2008 in English, but, with time, the platform implemented additional websites in Russian, Portuguese, Japanese, and Spanish with the same purpose and function as the initial website. A unique Id for each user allows to track users native of those languages across the websites and compare their choices when English was their only option versus when their native language became available.

To proxy for communication effort, I look at two measures of communication quality. One is based on users' actual communication content and uses message characteristics, while the second is a measure of communication outcomes. More precisely, the former corresponds to the number of separate snippets of code included in the answer. Since questions relate to computer programming, a more developed and informative answer would include a step-by-step procedure that alternates text and code. More pieces of code would then signal higher quality.[4] The second measure instead exploits the fact that authors of the questions can *accept* one of the answers they receive if they consider it enough satisfactory. This choice is not mandatory, so it can reliably inform whether the questioner could solve his problem with the information received.

I then measure the degree of incentive alignment exploiting another feature of the

[2] StackOverflow is probably the main source of help for computer programmers. Solutions provided in the platform affect the code of programmers all over the world, and mistakes or bad information may have a large impact. It happened for instance that the most copied code snippet from StackOverflow had an error: https://programming.guide/worlds-most-copied-so-snippet.html.

[3] Indeed different platforms adopted different strategies. For example, Wikipedia is available in 326 languages and Quora in 24 languages.

[4] Note that I am not measuring the length of the code.

website, called *bounties*. Stack Overflow users can auction reputation points (i.e. virtual rewards) on given questions. In other words, they can commit to providing a reward to the author of an answer considered enough satisfactory. The size of the number of points at stake provides then a measure of how much the author of the answer is incentivized.

The method used for the analysis is staggered difference-in-difference. I execute a regression analysis at the answer level with communication quality as the dependent variable, the availability of the user's native-language website as a treatment dummy, and time and user fixed effects. I use the estimation technique developed by Borusyak, Jaravel, and Spiess (2021) and compare the results with the more standard Two-Way Fixed-Effects approach.[5] I then proceed with a heterogeneous analysis by interacting the treatment dummy with different levels of 1) questioner's effort, 2) incentive alignment, and 3) the degree to which users started contributing to their native-language website. Finally, I evaluate for externalities on the English website by limiting the analysis to only English answers.

Other dimensions of the trade-off are analyzed more descriptively by comparing the non-native-English users who were contributing in English before their native language became available with those who were not. This comparison allows us to assess both differences in community size and quality of contributions. Finally, I measure the dispersion of information using *tags*, i.e. labels attached to questions to categorize their content. If the same tag appears across different languages, then some information may have been provided multiple times, meaning that some efficiency was lost. At the same time, if some topics are treated in some languages but not English, then the multiplicity of languages suggests that the platform lost some ability to aggregate information.

Overall the paper shows that the trade-off is confirmed in the data. The introduction of native-language websites increases communication effort by 24%. This effect jumps to 110% if the user is highly incentivised, and to 34% if the questioner puts a lot of effort. Answers are 7% more likely to be *accepted*, a 20% increase from the baseline. At the same time, the overall quality of the English website is not affected significantly, and users who remain active in English slightly increase their effort even on the English site. In addition, at least 42% of non-native English speaking users are likely to not have joined the platform in absence of their native-language website, implying a substantial increase in community size.[6] Nevertheless, new contributors appear to provide significantly lower quality answers in their native language, compared to users already active in English before the native-language website became available. In addition, there is a substantial overlap of topics across languages: out of 247 topics, 83 appear in at least 2 languages. At the same time, 12 topics out of the 247 are not treated in English at all, meaning that the platform did not optimally aggregate information. These findings are consistent with evidence from Wikipedia (Bao, Hecht, Carton, Quaderi, Horn, and Gergle 2012)

[5] The approach by Borusyak et al. (2021) solves for econometric issues identified by the literature (Callaway and Sant'Anna 2020, de Chaisemartin and D'Haultfœuille 2020, Sun and Abraham 2020, Borusyak et al. 2021).

[6] It has been shown that larger community size is beneficial not only because they constitute a larger base of contributors, but because it creates additional incentives for contributing users to provide content (Zhang and Zhu 2011)

and suggest that there is potential for efficiency gains by imposing a single language.

From these results, we can infer that a knowledge platform highly benefits from multiple languages, as they both increase the community size and the quality of information collected. Nevertheless, these benefits shade away if the non-native-English users are very few, or they have a very low cost of using English. In this case, from an efficiency standpoint, a single language (or a limited number of languages) is preferable.[7]

While this trade-off is particularly relevant for knowledge platforms, it is generalizable to any economic environment or institution where language diversity imposes the critical choice between centralization or decentralization of language. A typical example is national states (Ginsburgh and Weber 2011), where we saw the homogenization of language, like in France (Blanc and Kubo 2021), or the maintaining of language diversity, like in Spain. Another example is the firms choosing between common or specialized languages (Crémer et al. 2007). Finally, the trade-off is relevant in international trade, where a common language is necessary to find agreements, but language costs may prevent efficient interactions (Melitz 2008, Lohmann 2011).

To my knowledge, this is the first paper to empirically quantify the role of the cost of language on both communication effort and communication outcomes. Some experimental literature has tested communication games, with or without communication frictions (Lafky and Wilson 2020 and Blume, DeJong, Kim, and Sprinkle 2001 respectively). The works by McManus (1985), Tainer (1988), Guillouët, Khandelwal, Macchiavello, and Teachout (2021), and Battiston, Blanes I Vidal, and Kirchmaier (2021) also study exogenous communication costs and their effect on outcomes. McManus (1985), Tainer (1988), and Guillouët et al. (2021) look, as in this paper, at the cost of using English and study its impact on wages and productivity. Battiston et al. (2021) focuses instead on communication frictions arising from not being able to talk face-to-face, rather than the language itself. These papers do not observe the actual communication, but only communication outcomes, so they do not quantify changes in communication quality. In addition, these papers are silent on the role of the other party's effort, and the incentives of the information holder.

The rest of the paper is organised as follows: section 2 discusses communication in Q&A websites and the case of Stack Overflow, section 3 presents a simple theoretical framework, section 4 presents the data, section 5 presents the analysis for the effect of the cost of language on communication effort, section 6 discusses the trade-off faced by the platform, and section 7 concludes.

[7] A single language may be preferable also if the platform values only high-quality information and does not value the community size, as larger community size comes at the cost of noisier information. Nevertheless, generally, platforms that do not remunerate contributors are not sustainable if they do not have a large community base. At the same time, noisy information can be reduced with other approaches, like moderation.

2 Communication in Q&A platforms

Question and Answers websites are online platforms that allow users to ask new questions or answer existing ones. Examples of such websites are *Stack Overflow*, *Yahoo! Answers*, or *Quora*.[8]

The content of these websites is particularly useful for the analysis of communication strategies for several reasons. First of all, they provide detailed data on information transmission, including the information requested and the information provided, both generally not observed. This richness allows measuring effort in information transmission on both sides of the communication. In addition, as communities tend to be large, interpersonal relationships may be weaker. As a consequence, communication strategies are less likely to be affected by unobserved factors, like friendship or long-term norms, very common within firms. Finally, question and answers websites allow the researcher to observe a very large number of communication interactions, allowing more flexible statistical analysis.

2.1 Stack Overflow

For the empirical investigation, the paper relies on data from Stack Overflow, a question and answers website that focuses on topics related to computer programming. Questions may concern, for instance, how to use programming languages for data analysis or software development, or how to solve coding bugs. The website has the objective to be the main resource of information for all possible problems that programmers may encounter.[9] Key features of the platform are that it is crowd-based and free of charge. In other words, any internet user who registers (for free) on the website can ask questions and/or provide answers to other questions. Contributors are not remunerated.[10]

Stack Overflow stands out from other sites because of the size of its welfare impact: many programmers are self-learned and Stack Overflow provides a large community willing to help. As of June 2021 indeed, Stack Overflow receives more than one hundred million monthly visits.[11] In addition, Stack Overflow provides to information seekers content easily accessible and searchable via browsers' search engines. The literature has identified these features as particularly important for productivity gains (Boudreau, Brady, Ganguli, Gaule, Guinan, Hollenberg, and Lakhani 2017, Goldfarb and Tucker (2019), Sandvik, Saouma, Seegert, and Stanton 2020).

2.2 Language used in Stack Overflow

As of today, there are five different websites of Stack Overflow, each using a different language, namely English, Russian, Japanese, Portuguese, and Spanish. Note that, apart from the language, their function is identical. Each website anyway became public

[8] *Yahoo! Answer* has shut down in April 2021

[9] https://www.joelonsoftware.com/2008/09/15/stack-overflow-launches/

[10] Nevertheless, the platform has implemented several incentive systems, including virtual rewards.

[11] https://stackoverflow.com/company

at different times. Stack Overflow was first launched in English in September 2008. The platform was implemented in English as the founders are Americans and the use of English is the norm in the programming community. Nevertheless, they realized that a significant part of the programming community would not be able or may have problems accessing English content. After some discussion, they decided to allow the opening of Stack Overflow in other languages than English.[12]

The platform designers chose those 4 additional languages on the ground that large communities of programmers speak them and, at the same time, they may not speak English. The introduction of each website followed some *beta* periods before the rollout of the final version.[13]

Figure 1 shows the timeline of the introduction of the different websites.

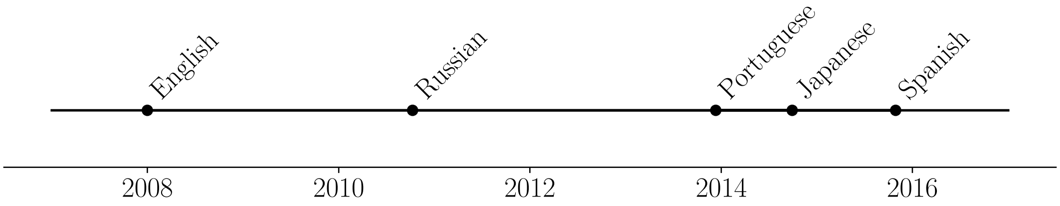


Figure 1: Timeline of the introduction of Stack Overflow websites.

The case of Stack Overflow in Russian
The introduction of Stack Overflow in Russian followed a slightly different process. In 2010, some Russian programmers decided to create a clone of English Stack Overflow in Russian. They created a website called HashCode which was replicating Stack Overflow features and purpose. Once the company behind Stack Overflow decided to open a version in Russian, they acquired HashCode, and on March 31st 2015 all posts from HashCode were imported into the Russian version of Stack Overflow. Formally then, the Russian version of Stack Overflow appeared in 2015. Figure 1 reports the 2010 date as the data includes all the HashCode content.

3 Theoretical Illustration

In this section, I present a simple framework to provide an overview of what and how different factors affect communication effort decisions. It aims to guide the empirical analysis and the interpretation of the results for both the intensive and extensive margin of communication choices.

The framework models unilateral information transmission in the Stack Overflow platform, that is, between who asks and who answers the questions[14]. In this simple environment, I abstract from strategic behavior and I assume that pairwise communication

[12] https://stackoverflow.blog/2014/02/13/cant-we-all-be-reasonable-and-speak-english/

[13] Appendix A.1 provides more details on the introduction of the websites

[14] The modelling approach and the functional forms are inspired by Calvó-Armengol, de Martí, and Prat (2015) and the communication literature in Organizational Economics

is independent of other communicating pairs.[15] In addition, I abstract from the incentive system implemented by the platform, and I rely on user and time fixed-effects at the stage of the analysis to control for heterogeneous sensitivity to the incentive system.

Let Bob be a programmer that needs to understand how to implement some features in his software. After multiple attempts, he decides to ask about his problem to the Stack Overflow community as, otherwise, he is not able to proceed with his project. Alice instead is a community member that sometimes answers questions on the platform.

Bob and Alice independently decide their communication strategies: it could be thought as they just keep the same strategies every time they participate on the website. This assumption is reasonable because the community is very large and both Bob and Alice cannot anticipate who will ask a question or provide the answer.

Alice then understands the solution to Bob's problem, decides whether to provide the answer and, in case, publishes her solution. Bob then implements the features thanks to Alice's help.

Note that the game is static as strategies are decided ex-ante.

More formally, let the information that Bob needs be θ, of which he only knows the ex-ante distribution:

$$\theta \sim \mathscr{N}\left(0, \frac{1}{s}\right).$$

In addition, let Bob's and Alice's efforts be defined, respectively, as E_Q and E_A, where E_Q captures the clarity and informativeness of the question, and E_A the clarity and informativeness of the answer. The cost of effort, C, depends on the cost of using a given language (λ) and the experience or general knowledge in the subject (k), and it is defined as:

$$C_Q = \frac{\lambda_Q}{k_Q}; \quad C_A = \frac{\lambda_A}{k_A}$$

for Bob and Alice's costs of effort respectively. The interpretation of the knowledge parameter is that the more the user is experienced, the more she can get to the point exactly, providing an accurate description of the question/solution. The crucial point that I want to capture is that the cost of language affects the intelligibility of the message, so that if a message is badly written it cannot be understood, while the experience affects the probability of misunderstanding, for example via misleading content.

Once Bob has published his questions, Alice provides the answer with the message m such that:

$$m = \theta + \varepsilon + \eta$$

where ε and η are noise terms that shrink with the agents' efforts. More precisely, $\varepsilon \sim \mathscr{N}\left(0, \frac{1}{E_Q}\right)$ and $\eta \sim \mathscr{N}\left(0, \frac{1}{E_A}\right)$.

[15] This assumption is justified by the fact that Stack Overflow is a very large community, and it is hard for users to have accurate beliefs over other users' decisions

The answer displayed in the platform is then a realization of the message distribution, which Bob can observe. Finally, let the action $a \in (-\infty, \infty)$ be what Bob will do to solve his problem.

The utility functions of Bob and Alice are, respectively,

$$\begin{aligned} U_Q &= -\left((a-\theta)^2 + C_Q^2 E_Q\right) \\ U_A &= -\left(\gamma(a-\theta)^2 + C_A^2 E_A\right). \end{aligned}$$

Bob wants to minimize any error in implementing the features in his software and will use the observed message to update his beliefs about the true value of θ. Alice will internalize Bob's utility to a certain degree $\gamma \in [0,1]$. In addition, since she knows Bob's prior and the message realization, she can anticipate Bob's action, given the message.

3.1 Optimal effort strategies

For a given question, how much effort Alice will decide to make?[16] Proceeding backward, Bob selects the action a^* such that:

$$a^* \equiv \arg\max_a \mathbb{E}[-\left((a-\theta)^2 + C_Q^2 E_Q\right)|m].$$

By bayesian updating, the optimal action is then given by:

$$a^* = \mathbb{E}[\theta|m] = \beta m \quad \text{with} \quad \beta \equiv \frac{E_Q E_A}{E_Q E_A + E_Q s + E_A s}.$$

In words, Bob weights the message by the expected informativeness.

To find her optimal effort level, Alice solves:

$$\max_{E_A \geq 0} \mathbb{E}[-\left(\gamma(a-\theta)^2 + C_A^2 E_A\right)]$$

and her best response is then given by:

$$R(E_Q) = \frac{E_Q(\sqrt{\gamma}k_A - s\lambda_A)}{\lambda_A(E_Q + s)}.$$

3.2 Implications of the model after a variation in the cost of language

How effort decisions are affected by variations in the exogenous cost of language? Let λ'_A be the initial level of exogenous communication cost and λ''_A be the new level. On the extensive margin, Alice provides an answer only if the cost of language is low enough, relative to her experience. More precisely, she participates if $\sqrt{\gamma}k_A > s\lambda''_A$. Her participation decision would change only if the condition is satisfied with a cost level λ''_A,

[16] Details on the steps are provided in appendix B

but not with a cost level λ'_A, or vice versa. On the intensive margin instead, her best response effort level would change by:

$$\Delta R(E_Q) = \frac{E_Q\left(\sqrt{\gamma}k_A - s\lambda''_A\right)}{\lambda''_A(E_Q + s)} - \frac{E_Q\left(\sqrt{\gamma}k_A - s\lambda'_A\right)}{\lambda'_A(E_Q + s)} \tag{1}$$

$$= \frac{E_Q\sqrt{\gamma}k_A(\lambda'_A - \lambda''_A)}{\lambda''_A\lambda'_A(E_Q + s)} = -\frac{E_Q\sqrt{\gamma}k_A\Delta\lambda_A}{\lambda''_A\lambda'_A(E_Q + s)}, \tag{2}$$

where $\Delta\lambda_A \equiv \lambda''_A - \lambda'_A$ is the size of the variation in the exogenous cost λ_A.

Equation 2 shows that, after a drop in the cost of language (i.e. $\Delta\lambda_A < 0$):

1. the effort choice of the answerer increases:

$$\Delta R(E_Q) > 0, \tag{3}$$

2. the change in the effort choice depends on the size of the change in the cost of language:

$$\frac{\partial \Delta R(E_Q)}{\partial \Delta\lambda_A} = -\frac{E_Q\sqrt{\gamma}k_A}{\lambda''_A\lambda'_A(E_Q + s)} > 0 \quad \text{if} \quad \Delta\lambda_A < 0 \tag{4}$$

3. the change in the effort is positive on the effort made by the questioner:

$$\frac{\partial \Delta R(E_Q)}{\partial E_Q} = -\frac{\sqrt{\gamma}k_A\lambda''_A\lambda'_A\Delta\lambda_A s}{\left[\lambda''_A\lambda'_A(E_Q + s)\right]^2} > 0 \quad \text{if} \quad \Delta\lambda_A < 0 \tag{5}$$

4. the change in the effort is positive on the degree of incentive alignment:

$$\frac{\partial \Delta R(E_Q)}{\partial \gamma} = -\frac{E_Q k_A \Delta\lambda_A}{2\sqrt{\gamma}\lambda''_A\lambda'_A(E_Q + s)} > 0 \quad \text{if} \quad \Delta\lambda_A < 0 \tag{6}$$

4 Data

In this section, I present the data used for the main analysis.

I retrieve the answers published in StackOverflow by two groups of users, the *Treatment* group and the *Control* group. The *Treatment* group is composed of users who face a shock in the cost of language. In other words, this group includes users for whom English is not the native language, and who may incur a cost reduction once the website in their native language becomes available. I assume that users who published posts in a language different from English are native to that language.[17] This assumption implies

[17] To justify this assumption, note that English is the most common language used in the community of programmers, suggesting that if a person is fluent in English would just use the English website. This is confirmed by the fact that 99.8% of English-platform users contributing as well in other languages contribute in only one other language

that when a non-English website was released, users speaking that language were able to publish in their native language, facing a drop in the cost of communication. The date at which the website in the native language of a given user became available is defined as the treatment date for that user.

The selected sample includes all users who posted at least one answer in English before treatment and at least one answer in another language, i.e. Russian, Japanese, Portuguese, and Spanish.[18] Note that users could keep writing answers in English after the treatment, but this is not a condition to be in the sample.

The *Control* group instead is composed of a random sample of users who did not participate in any of the non-English platforms of Stack Overflow. I assume that these users were not affected by the introduction of StackOverflow platforms in languages different from English.

Table 1 reports the total number of answers and the number of users who wrote them contained in the sample used for the analysis. To identify the platforms, I use *SO* for the Stack Overflow in English, while I add the first letter of the language to *SO* for the other platform: *SOJ*, *SOP*, *SOR*, and *SOS* are, respectively, Stack Overflow in Japanese, Portuguese, Russian, and Spanish. The *Treatment status* indicates whether the authors published the answers before or after being treated. Figure 2 shows instead the sample size across time, with each platform's sample stacked vertically.

The data are right-censored at the end of August 2017.

			#answers	#authors	Earliest	Latest
Group	Post in:	Status				
Control	SO		6976	536	2008-09-16	2017-08-27
Treatment	SO	Not yet Treated	128984	2680	2008-08-12	2015-10-29
		Treated	100610	2089	2010-10-10	2017-08-28
	SOJ	Treated	3435	204	2014-10-10	2017-08-25
	SOP	Treated	30273	1183	2013-12-12	2017-08-27
	SOR	Treated	8448	137	2010-12-20	2017-08-28
	SOS	Treated	15139	1156	2015-10-30	2017-08-28

Table 1: Total number of answers in the sample, unique authors that wrote them, and dates of the earliest and latest answer in the group. Values are grouped by 1) Treatment group of the author (Treatment or Control), 2) platform, and 3) whether the author was treated at the time he/she wrote the answer.

[18] As of 2021, only those languages are available.

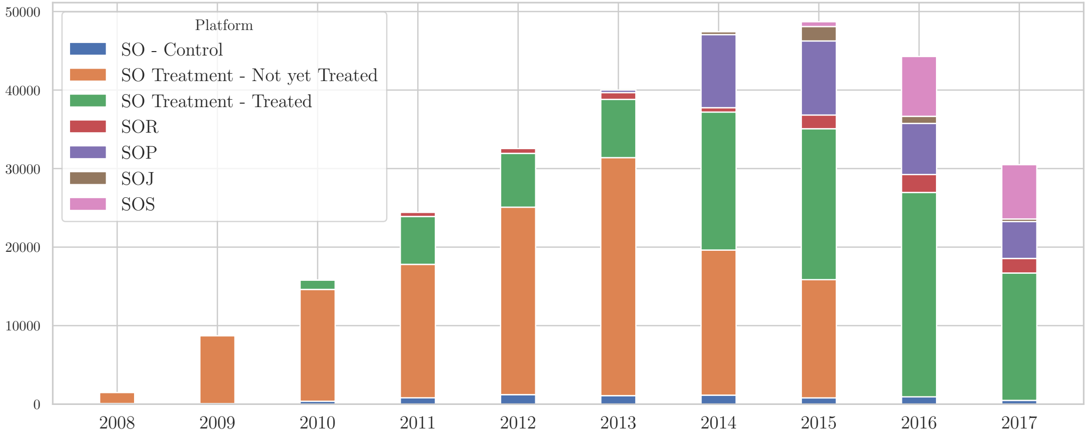


Figure 2: Number of answers in the sample for each year. The different colors identify the different platforms where those answers were published. SO, which corresponds to Stack Overflow in English, is split by answers published by users never treated (Control), users not yet treated, and users already treated.

4.1 Users' "adoption" of the native-language website

Once the platform implemented the non-English websites, treated users could participate using both their native language or English. Figure 3 shows the extent to which treated users adopted the non-English websites. It reports the distribution of the number of answers published by each treated user in the sample before and after being treated. Graphs separate users based on their native language.[19] It is possible to notice that, on average, users kept writing in English even after their native language became available.

There is anyway substantial heterogeneity in behavior before and after the native-language website became available. Figure 4 shows how many treated users published a certain quantity of answers on the native-language website, conditional on how much they published on the English website before the native-language became available. The figure reports separate plots for each website. The figure shows that users cluster on the extreme. On one side, as a general pattern across websites, some users who produce a lot in the English website before treatment also contribute a lot in the non-English one, and vice-versa. On the other side, some users contribute a lot in their native language after contributing little in English and vice-versa. The latter pattern is more clear on the Spanish and Portuguese websites. It suggests that some users have a low cost of using English, and the native-language platform did not bring many benefits to them, while other users have higher costs of using English, and, as a consequence, higher benefits from the introduction of the new websites.

To quantify the extent to which users switch to their native language platform, I

[19] For this statistic, it is relevant to split the sample by native language because users treated earlier had more time to publish answers after treatment. Indeed, the Russian sample is on a different scale because it was treated much earlier.

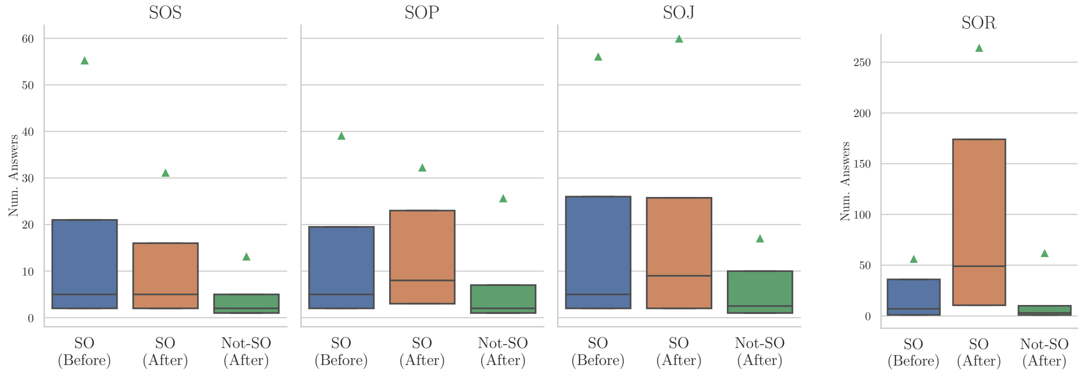


Figure 3: Statistics on the number of answers that each author published in English (Before and After the platform in her native language became available) and in her native language. Each plot reports the distribution conditioning on the native language of the author: from left to right, Spanish, Portuguese, Japanese, and Russian. The box reports the 25th, 50th, and 75th quantiles, while the triangle reports the mean.

compute users' rate of contribution to the native language website, relative to the total amount of contribution after treatment. For this specific purpose, and differently from previous statistics, I consider *contributions* both questions and answers. The measure is the result of the total number of questions and answers published by the user on the native-language website over the total number of questions and answers published after treatment (both in English and in the native language). Figure 5 shows the distribution of the measure, which confirms the above discussion: generally, after treatment, users either contribute only to the native-language website, or they mainly contribute to the English one. (Note that, by construction of the data, all users in the sample contribute to the native-language website, which explains the absence of a high spike at 0).

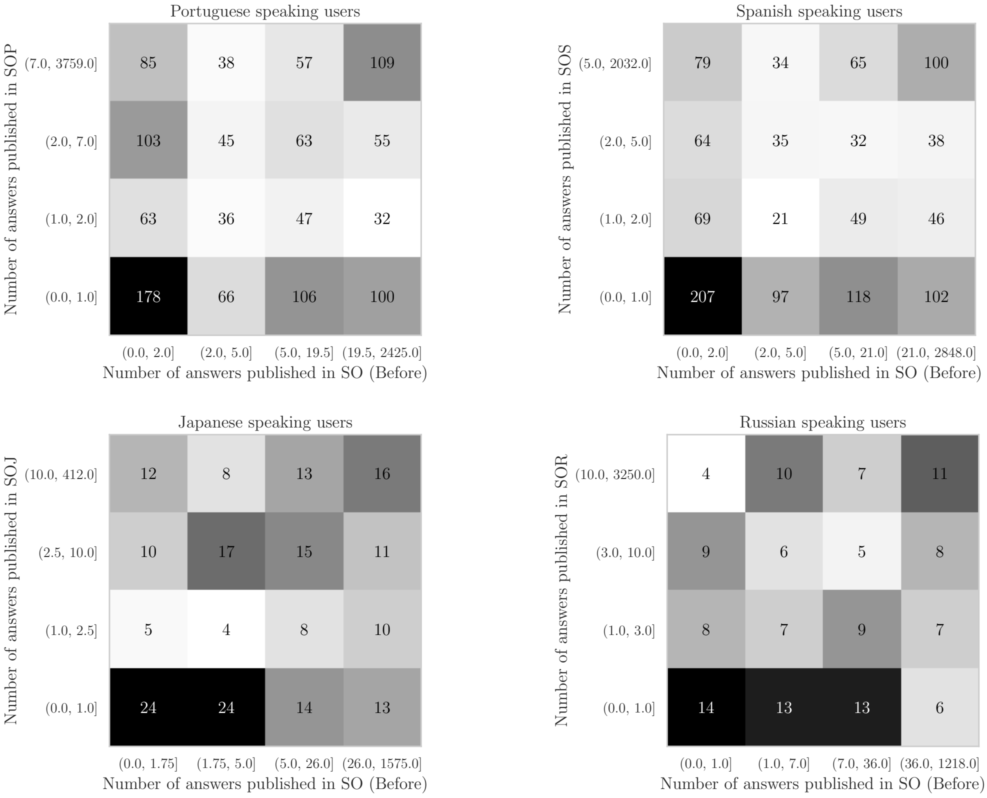


Figure 4: Distribution of users based on participation in English before the native-language platform became available and in the native language once it was available. Numbers in the plot correspond to the number of users in the sample who published a positive number of answers according to the respective intervals. Intervals are based on the 0.25, 0.5, and 0.75 quantiles of the respective distributions.

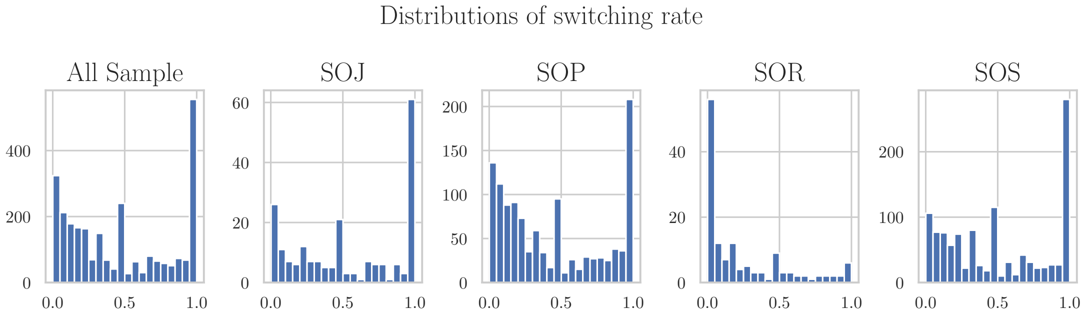


Figure 5: Distribution of switching rate measured as the total number of question and answers published by the user in the native-language website over the total number of questions and answers published after treatment.

4.2 Measure of answer quality

A standard and simple measure of textual informativeness is the text length, measured, for instance, as the number of words used. This proxy for quality, as well as all alternatives that use text measures, is language-specific and not comparable across languages.

To overcome this issue, the paper proxies for quality using the number of separated pieces of code contained in the answer. More precisely, each answer is an *html* script. Once users include code snippets in the answer, they add *code* sections (i.e. <*code*>...< /*code*>) such that the code will appear in a separate box with a different color background. The box mimics a programming/statistical software's console and makes the code more readable. The proxy of quality is then defined as the number of *code* sections in the answer. In the appendix, figure 15 shows an example of an answer with two snippets of code. The intuition behind this measure is that a typical answer about programming would include some textual explanation and some code snippet to illustrate the solution. The presence of multiple snippets may indicate that either the author is providing several pieces of information, or that she is explaining one piece of information more clearly, with a step-by-step procedure. In both cases, more snippets suggest higher informativeness of the answer. Table 2 reports the distribution of the number of code snippets across answers.

	mean	std	min	25%	50%	75%	max
Distribution Num. Codes							
Full sample	3.37	4.76	0.0	1.0	2.0	4.0	284.0
Before Treatment - SO	2.71	3.99	0.0	1.0	2.0	3.0	284.0
After Treatment - SO	3.55	4.65	0.0	1.0	2.0	4.0	153.0
After Treatment - SOJ	3.70	4.65	0.0	1.0	2.0	5.0	52.0
After Treatment - SOP	4.60	6.38	0.0	1.0	3.0	6.0	186.0
After Treatment - SOR	4.49	6.18	0.0	1.0	3.0	6.0	120.0
After Treatment - SOS	5.00	5.93	0.0	1.0	3.0	6.0	129.0
Never treated	2.26	3.34	0.0	0.0	1.0	3.0	59.0

Table 2: Distribution of the number of pieces of Codes across all answers of the sample

Another possible way to proxy for answers' quality is to measure the degree of appreciation from the community, that is, based on communication outcomes. The data provides two possible indicators of such measures: answers acceptances and up-votes. The user who asked the question can *accept* one of the answers as *best answer*. This indicates that the *accepted* answer was the one to solve his problem. At the same time, every registered user can up-vote (or down-vote) answers, similarly to how users allocate *likes* in other platforms.[20]. These measures anyway depend on time: users could accept or up-vote answers later than when the answer is published. This means that

[20] There are some exceptions on who can vote content. For details, see https://stackoverflow.com/help/privileges/vote-up

more recent answers, according to this measure, may be of lower quality mechanically.[21] Nevertheless, both measures correlate with the number of pieces of code. Figure 6 shows that answers with more pieces of code on average obtain a higher score, where the score is the number of the up-votes net of downvotes that the answer received. It also shows that *accepted* answers include, on average, a higher number of pieces of code. The pattern is consistent across websites, as it is possible to see in figure 7, but in general, does not apply to answers with zero pieces of code. This suggests that there may be two different types of questions, one that requires some code in the answer and one that does not. I will address this issue in the analysis via robustness checks.

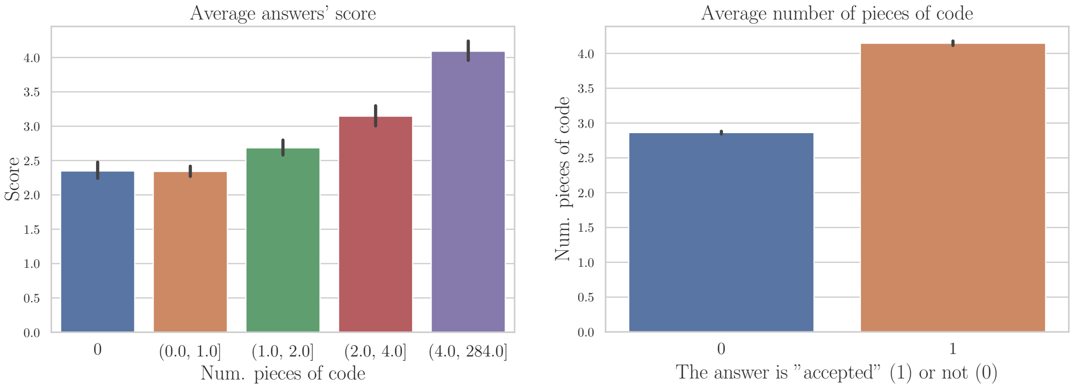


Figure 6: [**Left**]Average score obtained by answers, conditional on the answers having a certain number of pieces of code. Intervals of number of code snippets are based on the 0.25, 0.5 and 0.75 quantiles of the distribution across all answers in the sample. [**Right**] Average number of pieces of code across non-accepted (0) or accepted (1) answers. Vertical bars are 95% confidence intervals computed via bootstrapping.

[21] While this is a concern, it is not a major issue. In Stack Overflow, most up-votes and acceptances occur soon after publication.

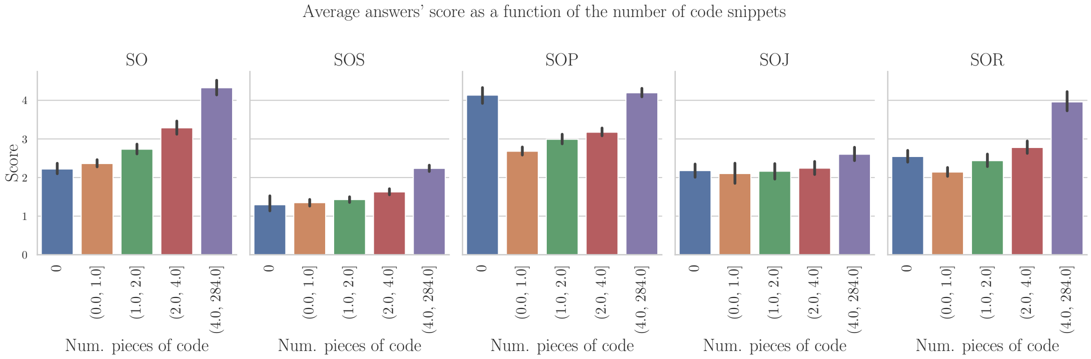


Figure 7: Average number of points obtained by answers, conditional on the answers having a certain number of pieces of code. Intervals of number of code snippets are based on the 0.25, 0.5 and 0.75 quantiles of the distribution across all answers in the sample. Vertical bars are 95% confidence intervals computed via bootstrapping.

4.3 Additional variables

The quality of the question

I proxy for the questioner's effort in communication with the number of pieces of code included in the question, that is, the same measure of quality used for the answers.

Incentive alignment

The data does not provide information on the degree to which the user internalizes the questioner's utility. Nevertheless, users can be incentivized by the questioner (or other participants) with reputation points via *bounties*. The platform allows the questioner to auction a certain amount of reputation points on a given question, and to promise to allocate these points to the user who would provide a satisfactory answer. The auctioned points are allocated at the discretion of the questioner (even though some automatic allocation rules may apply in certain cases) and the questioner loses them even if the points are not allocated.[22] This feature allows for variation in virtual remuneration, which can proxy for the degree of incentive alignment between the communicating parties. Figure 8 (right graph) reports the frequency distribution of bounty amounts.

Empathy

To capture the degree of empathy, I use 1) whether the two communicating parties share the same language, 2) the type of profile picture displayed by the questioner, and 3) whether the questioner displays a full name (i.e. name and surname). All this information would allow the user answering to know whether the questioner shares the same nationality and group identity.

The variable that captures the commonality of language between the user and the questioner is a dummy equal to 1 if the questioner displays his location, and the language spoken in that location corresponds to the native language of the author of the answer.[23]

[22] More details are available here: https://ell.stackexchange.com/help/privileges/set-bounties

[23] Since I do not have reliable ways to identify the nationality of users in the control group, this variable is missing for those users.

Note that this variable is based on the information available to the author of the answer, and it is not necessarily correct in reality. Nevertheless, to capture the degree of empathy, it is indeed relevant to rely only on the information available to the user. Note also that the "same language" variable takes always a value equal to 1 if the answer is published on a non-English website.

The variable for the full name of the questioner is a dummy equal to 1 if the displayed name of the questioner matches the pattern of two words separated by a space and with capital letters.

Finally, the type of questioner's picture corresponds to a categorical variable based on whether the questioner's profile displays the default avatar, a personalized picture, or none of them. Figure 8 (left graph) reports the frequency distribution.

Competition in answering

In Stack Overflow, more users can answer the same question and there is no ex-ante agreement on who will answer a given question. To capture this form of "competition", I adopt two proxies: the total number of answers the question has received, and the total number of views (i.e. impressions) of the question.

Table 3 reports descriptive statistics across answers of the sample for, from left to right, the questions' quality, the bounty amount, the dummy variables equal to 1 if the communicating party share the language and if the questioner displays a full name, the number of answers to the question, and the number of views to the question.

	Quality Q.	Bounty	Same Lang.	Q. Full Name	# Answers	# Views
mean	2.45	1.12	0.21	0.25	2.77	8239.55
std	2.94	14.41	0.41	0.43	4.42	71869.20
min	0.00	0.00	0.00	0.00	0.00	5.00
25%	1.00	0.00	0.00	0.00	1.00	140.00
50%	2.00	0.00	0.00	0.00	2.00	584.00
75%	3.00	0.00	0.00	1.00	3.00	2290.75
max	111.00	1000.00	1.00	1.00	518.00	8671208.00
# Sample	292919.00	293777.00	286801.00	287231.00	292926.00	292926.00

Table 3: Descriptive statistics of variables affecting effort provision. Respectively, columns correspond to 1) the number of pieces of code (i.e. quality) of the questions being answered by the answers of the sample; 2) bounty amount at stake on the questions that the answers are addressing; 3) a dummy equal to one if the author of the answer share the same native language as the questioner (note that this variable is always one in the platforms using a language different from English); 4) a dummy equal to 1 if the questioner displays both name and surname; 5) the number of answers received by the question that the answer is answering to; 6) the number of views received by the question that the answer is answering to.

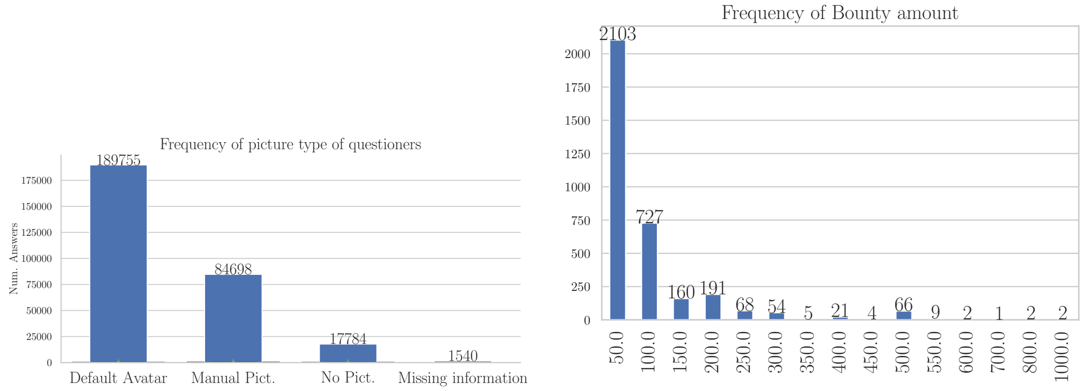


Figure 8: [**left**]Frequency distribution across answers of the type of profile picture used by the author of the question which the answer is answering. [**right**] Frequency distribution across answers of the amount of bounty at stake on the question that the answer is answering.

5 The effect of the cost of language on communication effort

Before moving to the analysis, a representation of the raw data may already provide suggestive evidence of users' behavior.

Figure 9 reports the average number of pieces of code, i.e. the measure of effort, made across answers before and after the non-English platforms became available. On the x-axis are reported 7-days periods before ad after. Note that they do not correspond to calendar weeks since the treatments are staggered. It is possible to see that while the average effort remains substantially similar on the English website, it is substantially higher on the non-English platforms. This shows that users on average include more pieces of code when they reply in their native language platform.

Figure 10 reports the same scatterplot, but separately for each non-English language. In this case, observations are answers written by authors native of one of the non-English languages. It follows that, in a given graph, all authors are treated at the same time, and the 7-days periods correspond to calendar dates. Compared to figure 9, these graphs include also the contributions from users of the Control (i.e. never-treated) group.

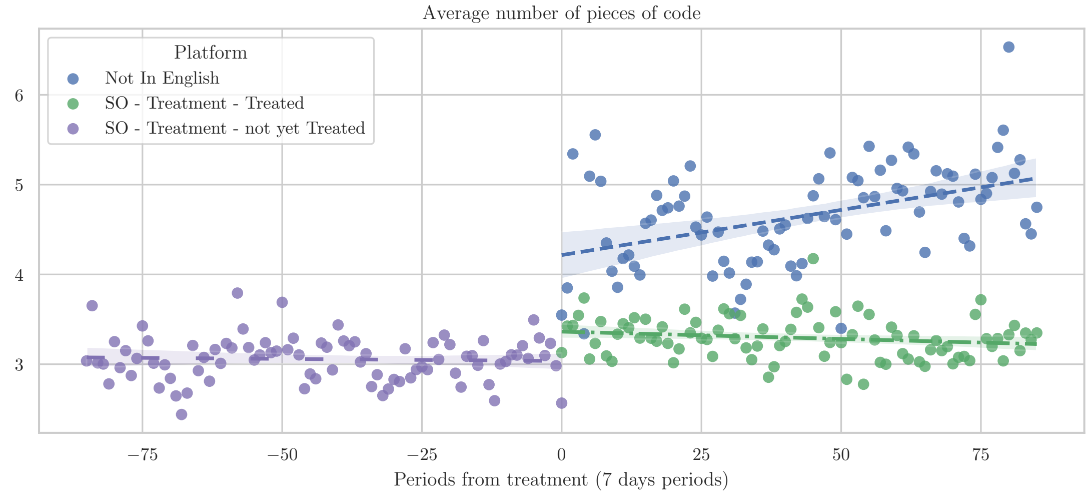


Figure 9: Average number of pieces of code in the 7-days periods before and after treatment, i.e. before and after the introduction of the native-language websites. In the after-period, separate colors discriminate between answers published in English and answers published in other languages.

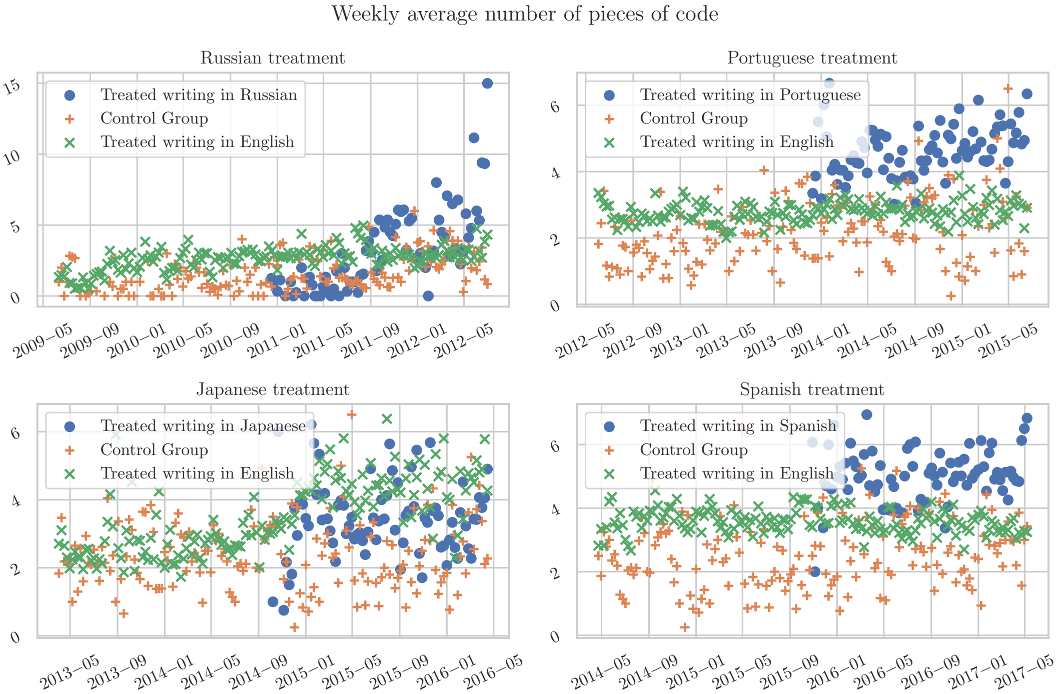


Figure 10: Average number of pieces of code across time. Graphs report data from the Control group of users (which is the same across graphs for corresponding dates) and data of the treated users, based on their native language.

5.1 Threats to identification

The identification of the effect requires the assumption that, at the time of treatment, there were no other variables that have simultaneously changed and affected communication decisions of treated users only. A possible concern in the setting of this analysis is that the native-language websites differ from the English site due to other factors than the language alone. In particular, there are two main threats to identification. The first relates to the inclination that users may have to put effort when using their native language compared to English. If Alice feels more empathetic toward other community members who speak her native language, she may be willing to put extra effort into her native-language website independently of the language cost (Lyons 2017, BenYishay and Mobarak 2019, Ginsburgh and Weber 2020). The second relates to the community size since indeed the communities of the new websites are substantially smaller than the English one. One possible implication is that if Alice cares to receive up-votes on her answer, she may need less effort to do so because of less competition.

To address these concerns, I will include in the analysis control variables that aim to capture both Empathy and Competition, as described in section 4.3.

5.2 Estimation

Let a user be *treated* if she is part of the *Treatment Group* and the website using her native language is already available.[24] The estimand of interest is then the average difference between the quality of answers published by treated users and the quality that those answers would have in a potential scenario where the authors could only use English. More precisely, let i index answers, j index users, so that $j(i)$ is the author of answer (i), and t index time periods. The estimation target is:

$$\tau = \sum_i w_i \tau_i$$

$$\text{with} \quad \tau_i = Y_i - Y_i(0)$$

$$\text{s.t.} \quad j(i) \text{ treated at time } t(i)$$

Where w_i are non-stochastic weights, Y_i is the outcome variable of answer i, and $Y_i(0)$ is the potential outcome of the answer i if $j(i)$ would not be treated.

To identify this effect, I exploit the staggered implementations of the non-English websites, which allows me to compare 1) the treated units with units not yet treated, and 2) the treated units with units that will never be treated. To account for the individual and time fixed effects, the literature has traditionally adopted the so-called Two-Way Fixed Effect estimation method (TWFE) which consists of a linear regression of the outcome variable on individual fixed effects, time fixed effects, and a dummy equal to 1 when the unit is treated. The regression is estimated via OLS. Nevertheless, in the context of the data used for the analysis, this approach is likely to provide biased

[24] For what follows, for simplicity I will identify a *treated* answer as an answer written by a *treated* user.

estimates.[25] To overcome this issue, I use the estimation strategy proposed by Borusyak et al. (2021), which is based on the prediction of the unobserved potential outcome using a model *trained* on the non-treated and control-group data. More precisely, the estimation strategy proceeds in three steps. First, it estimates via OLS the user and time fixed effects, using only non-treated answers, i.e. answers of both not-yet treated and never treated users. It then predicts the potential counterfactual outcome $\tilde{Y}_i(0)$ for the treated observations exploiting the estimates made in step one. This allows to compute the estimate of the treatment effect $\hat{\tau}_i = Y_i - \tilde{Y}_i(0)$ for each observation. Finally, the third step averages the difference between observed and predicted outcomes across all observations.[26]

This estimation strategy relies on the parallel trend assumption, homoskedastic errors, and no anticipation of the treatment. Note that in the context of this paper, even if the treatment is anticipated users' cost of language is unchanged until they are treated, making the no-anticipation assumption naturally satisfied.

As a matter of comparability with traditional approaches, the analysis will provide estimation results for both the Two-Way Fixed Effects method (TWFE hereafter) and the method proposed by Borusyak et al. (2021) (BJS hereafter).

Let i index answers, j index users, and t index time (weeks). In addition, let *numCodes* identify the quality measure based on the number of snippets of code of the answer, and L index the non-English languages, i.e. either Russian, Portuguese, Japanese, or Spanish.

TWFE

The Two-Way Fixed Effect estimation approach would then estimate the treatment effect by estimation via Ordinary Least Squares (OLS) of the following regression:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \beta D_{L(j(i),t(i))} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i,$$

[25] As discussed by several papers (Callaway and Sant'Anna 2020, de Chaisemartin and D'Haultfœuille 2020, Sun and Abraham 2020, Goodman-Bacon 2021, Borusyak et al. 2021) the two-way fixed effect estimation procedure estimates the treatment effect as a weighted average of all possible treatment effects for each user × period cell. The weights sum to one, but could be negative. This fact may be an issue in the context of this paper. I cannot rule out that the treatment effect is heterogeneous across time and users. Users may potentially take time to adjust to the new environment, and users with a higher cost of using English may be more impacted. This may cause biased estimates using OLS, potentially even of the opposite sign if the treatment effect increases over time. In addition, the method proposed by Borusyak et al. 2021 allows more flexibility in addressing the fact that the data constitute an unbalanced panel, as discussed in appendix D.2. Finally, because the TWFE estimation gives more weight to the treatment effect of the units treated for the longest period, in the context of this paper the treatment effect of the Russian sample is overweighted. As discussed in section 2.2, the Russian sample followed a relatively different history and its treatment effect could be, because of this reason, significantly different from the others.

[26] The literature has proposed other solutions, e.g. de Chaisemartin and D'Haultfœuille (2020) and Callaway and Sant'Anna (2020) suggest alternatives that rely only on the data just before and after the treatment of each cohort (i.e. the set of individuals treated at the same time). In the context of this paper, those solutions are less preferable because I observe an unbalanced panel (not all users participate every week). The selection of data may cause the creation of biased comparison groups.

where $D_{L(j(i),t(i))}$ is a dummy equal to 1 if author $j(i)$ at time $t(i)$ is able to use the website in her native language L, different from English. β is the coefficient of interest, capturing the treatment effect.[27] $\boldsymbol{W}_{i(jt)}$ is a vector of answer-specific control variables.

BJS

The alternative method proposed by Borusyak et al. (2021) instead estimates the treatment effect via a three-step procedure. First, it estimates via OLS a linear model on the non-treated sample:

$$\text{[Step 1]} \quad numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i \quad \text{if } j(i) \text{ not treated at time } t(i),$$

then, it predicts, using the estimated model, the potential outcome of treated units if were untreated, and compute the observation-specific treatment effect:

$$\begin{aligned} \text{[Step 2]} \quad \widehat{numCodes}_i &= \hat{\alpha}_{j(i)} + \hat{\alpha}_{t(i)} + \boldsymbol{W}_i'\hat{\boldsymbol{\gamma}} \quad \text{if } j(i) \text{ treated at time } t(i), \\ \hat{\tau}_i &= numCodes_i - \widehat{numCodes}_i \quad \text{if } j(i) \text{ treated at time } t(i). \end{aligned}$$

Finally, it averages all treatment effects, to obtain the average treatment effect:

$$\text{[Step 3]} \quad \hat{\tau} = \frac{1}{N_{post}} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i.$$

Where N_{post} is the number of answers published by treated users. Note that this is not the only possible way to compute the final treatment effect. As suggested by Baker, Larcker, and Wang (2022), in presence of an unbalanced panel like in this setting, the average treatment effect could be computed by first obtaining the average effect for each user, and then taking the average effect across users. Table 18 in the appendix presents results using that approach.

Table 4 reports the estimated treatment effect (i.e. *after*), corresponding to $\hat{\beta}$ in the Two-Way Fixed Effects specification and to $\hat{\tau}$ in the BJS specification. It shows that when users can write answers in their native language, on average they include significantly more pieces of code. This result confirms the theoretical implication of equation 3, stating that a reduction in the exogenous cost of language induces an increase in communication effort.[28]

5.2.1 The treatment effect depends on the questioner's effort

Equation 5 of the theoretical framework shows that the change in effort induced by the variation in exogenous cost increases with the effort made by the questioner. To test this

[27] Note that to have $D_{L(j(i))} = 1$ is not necessary that i is published in language L, as it could be published in English. To have $D_{L(j(i))} = 1$ is sufficient that language L is available to author $j(i)$.

[28] As discussed in the data section, the correlation between the number of pieces of code and up-votes is not satisfied when the answers have zero pieces of code. This may suggest that some answers do not need to include any code. Table 17 in the appendix reports the regression results after dropping all answers with zero pieces of code and selecting users that, given the remaining answers, were active both before and after treatment. Results are consistent.

	(1) TWFE	(2) TWFE 1	(3) TWFE 2	(4) TWFE 3	(5) BJS	(6) BJS 1	(7) BJS 2	(8) BJS 3
after	0.392^{*}	0.387^{*}	0.388^{*}	0.205^{*}	0.656^{***}	0.677^{***}	0.683^{***}	0.663^{***}
	(0.107)	(0.111)	(0.111)	(0.0551)	(0.0412)	(0.0397)	(0.0387)	(0.0751)
Observations	293777	292919	292919	280407	293777	292846	292846	199564
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QEffort	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes	No	No	Yes	Yes
Empathy	No	No	No	Yes	No	No	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Baseline Regressions' estimates where the dependent variable is the number of pieces of code. The estimate *after* corresponds to the average treatment effect, and corresponds to the parameters $\hat{\beta}$ or $\hat{\tau}$ if the specification adopted is the TWFE or the BJS respectively. Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

hypothesis, I separately estimate the treatment effect by different levels of effort made by the questioners.

As a proxy for the questioners' effort, I use the number of separated snippets of code that the questioner included in the question. Define this variable as $Qeffort$. I then bin this variable into four levels of effort, as described in table 5. The thresholds of each category correspond to the quartiles of variable $Qeffort$'s distribution.

The TWFE method's specification is then the following:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \sum_{\eta} \beta_\eta D_{L(j(i),t(i))} \mathbf{1}_{\eta(i)} + \boldsymbol{W}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where η identifies the level of questioner's effort, and $\mathbf{1}_{\eta(i)}$ is an indicator function equal to 1 if the question that the answer i is addressing contains a number of pieces of code corresponding to category η.

For what concern instead the BJS method, average treatment effects are taken within each category:

$$\hat{\tau}_\eta = \frac{1}{N_\eta} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i \mathbf{1}_{\eta(i)}$$

where N_η is the number of answers in the sample written by treated users whose question is of quality level η.

Table 6 reports the estimate results: column 1 and 2 contain the $\{\hat{\beta}_\eta\}_{\forall\eta}$, while column 3 and 4 contain the $\{\hat{\tau}_\eta\}_{\forall\eta}$. Results confirm that the treatment effect grows with higher level of questioner's effort.

5.2.2 The treatment effect depends on the incentive alignment

According to the model, as shown in equation 6, the effect of a decrease in the exogenous cost of effort is increasing with the degree of incentive alignment between questioner and

	average number of snippets of code in questions
Low	[0,1]
MediumLow	(1,2]
MediumHigh	(2,3]
High	(3,111]

Table 5: Categories for the effort level of the questioner

	(1)	(2)	(3)	(4)
	TWFE	TWFE 2	BJS	BJS 2
Low × after	0.143	-0.0522	0.374^{***}	0.388^{***}
	(0.129)	(0.0693)	(0.0638)	(0.0927)
MediumLow × after	0.581^{**}	0.401^{**}	0.868^{***}	0.869^{***}
	(0.100)	(0.0543)	(0.0788)	(0.107)
MediumHigh × after	0.578^{**}	0.400^{**}	0.884^{***}	0.912^{***}
	(0.103)	(0.0455)	(0.0708)	(0.0977)
High × after	0.592^{**}	0.413^{***}	0.977^{***}	0.927^{***}
	(0.0709)	(0.0236)	(0.0328)	(0.0596)
Observations	292919	280407	292846	199564
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QEffort	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: Estimates by level of questioner's effort. Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

answerer.

To measure the incentive alignment between the two parties, I use the values of the so-called *bounties*, as discussed in section 4.3. Bounties can be considered virtual payments and create a direct incentive to answer well the question. The bounty amount auctioned on the question and not yet assigned (i.e. *active*) provides a measure of incentive alignment.

I discretize the bounty amount into four categories: the low category is composed just of the zero amount, while the other three categories are based on the 33^{rd} and 66^{th} quantiles of the distribution of the positive amounts. The categories are reported in table 7.

Similarly to previous heterogeneity analysis, the TWFE method estimates the treatment effects with the following specification:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \sum_{\phi} \beta_\phi D_{L(j(i),t(i))} \mathbf{1}_{\phi(i)} + \boldsymbol{W}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where ϕ indexes the categories of the amount of active bounties open on the question addressed by the answer, and $\mathbf{1}_{\phi(i)}$ is an indicator function equal to 1 if the question that the answer is addressing has an amount of active bounty points of level ϕ.

For what concern instead the BJS method, average treatment effects are taken within each category:

$$\hat{\tau}_\phi = \frac{1}{N_\phi} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i \mathbf{1}_{\phi(i)}$$

where N_ϕ is the number of answers in the sample written by treated users whose questions have an amount of active bounty points of level ϕ.

Results are reported in table 8. They show that on average, the treatment effect is higher when authors are more incentive-aligned, as suggested by the theoretical framework.

	amount of bounties
Low	0
MediumLow	50
MediumHigh	100
High	[150,1000]

Table 7: Categories for the amount of bounties allocated to questions that a user answered in a given week

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.373^{*}	0.190^{*}	0.666^{***}	0.652^{***}
	(0.110)	(0.0534)	(0.0391)	(0.0758)
MediumLow × after	1.235^{*}	1.045^{*}	1.645^{***}	1.088^{***}
	(0.287)	(0.236)	(0.192)	(0.189)
MediumHigh × after	2.296	2.135	2.759^{***}	2.355^{***}
	(0.831)	(0.874)	(0.425)	(0.447)
High × after	3.008^{***}	2.651^{**}	3.477^{***}	2.976^{***}
	(0.268)	(0.209)	(0.388)	(0.408)
Observations	292919	280407	292846	199564
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QEffort	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Estimates by level of incentive alignment. Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

5.2.3 Who drives the effect?

The theoretical framework suggests that the size of the effect is larger the higher the drop in the cost of language.[29] I do not observe individuals' cost of using English, but, assuming some frictions in switching to the native language website, we would expect that users with a higher cost would be more likely to switch. I then categorize users by the share of posts (i.e. questions or answers) published on a non-English website relative to the total amount of posts published after the native-language website became available. This measure allows to characterize users by the degree they switch to the native-language platform.[30]

To estimate potential heterogeneity in this dimension, I categorize this proxy in four categories, as displayed in table 9. The boundaries of each category are based on the 25^{th}, 50^{th}, and 75^{th} quantiles of the distribution. I then estimate separate treatment effects for each category. More precisely, with c indexing the level of language cost, in the TWFE method the specification is the following:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \sum_c \beta_c D_{L(j(i),t(i))} \mathbf{1}_{c(j(i))} + \boldsymbol{W}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where $\mathbf{1}_{c(j(i))}$ is an indicator function taking value 1 if the user j belongs to the level category c.

For what concern instead the BJS method, the cost-based estimates will be obtained by averaging the treatment effects within each category of cost:

$$\hat{\tau}_c = \frac{1}{N_c} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i \mathbf{1}_{c(j(i))}$$

where N_c is the number of answers in the sample written by treated users with a cost of English within the category c.

Table 10 reports estimates for the four categories. Parameters corresponds to $\{\hat{\beta}_c\}_{\forall c}$ for the TWFE columns, and to $\{\hat{\tau}_c\}_{\forall c}$ for the BJS columns. It is possible to see that the effect is largely driven by users that switch more to the native-language website.

	share of answers not in English in the after-period
Low	[0,0.143]
MediumLow	(0.143,0.426]
MediumHigh	(0.4326,0.875]
High	(0.875,1]

Table 9: Categories for the exogenous cost of using English (boundaries rounded at 2 decimals)

[29] See equation 4.

[30] For more details, see section 4.1 and figure 5

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.0988	0.125	0.228^{***}	0.212^{*}
	(0.114)	(0.102)	(0.0571)	(0.101)
MediumLow × after	0.224	0.0889	0.472^{***}	0.217^{**}
	(0.122)	(0.106)	(0.0460)	(0.0795)
MediumHigh × after	0.660^{*}	0.232	0.562^{***}	0.644^{***}
	(0.198)	(0.125)	(0.0351)	(0.113)
High × after	1.475^{***}	0.838^{*}	1.883^{***}	2.214^{***}
	(0.142)	(0.174)	(0.0211)	(0.0825)
Observations	292919	280407	292846	199564
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QEffort	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Estimates of average treatment effect by level of exogenous cost of using English. Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

5.3 Alternative quality measure based on outcomes

The number of pieces of code in the answer is a proxy for communication effort based on the characteristics of the message. Another approach to proxy for quality is based on observable outcomes, such as, for instance, the questioner's appreciation of the answer.

StackOverflow allows authors of questions to "accept" an answer as the "best answer". This action is not mandatory and does not depend on the number of answers provided to the same question. The action simply allows questioners to mark that the given answer provided a satisfactory solution to the question they stated.

If authors of answers employ higher effort, the likelihood that their answer is accepted as the "best answer" should increase.

In this section, I estimate the treatment effect of a drop in the cost of language on the probability that the answer is accepted by the questioner as the best answer.

In a way similar to the previous analysis, I estimate the treatment effect using both the TWFE and the BJS methods.

TWFE

For the Two-Way Fixed Effects approach, the specification adopted is the following:

$$\mathbf{1}_{(\text{i is accepted})} = \alpha_{j(i)} + \alpha_{t(i)} + \beta^{BA} D_{L(j(i),t(i))} + \boldsymbol{W}_i' \boldsymbol{\gamma} + \varepsilon_i,$$

where $\mathbf{1}_{(\text{i is accepted})}$ is an indicator function that takes value equal to 1 if answer i is accepted as "best answer" and 0 otherwise.

BJS

For what concerns the BJS method, I follow again the three-step procedure:

$$
\begin{aligned}
&\text{[Step 1]} \quad \mathbf{1}_{(\text{i is accepted})} = \alpha_{j(i)} + \alpha_{t(i)} + \boldsymbol{W}_i'\gamma + \varepsilon_i \quad \text{if } j(i) \text{ not treated at time } t(i),\\
&\text{[Step 2]} \quad \mathbf{1}_{\widehat{(\text{i is accepted})}} = \hat{\alpha}_{j(i)} + \hat{\alpha}_{t(i)} + \boldsymbol{W}_i'\hat{\gamma} \quad \text{if } j(i) \text{ treated at time } t(i),\\
&\qquad\qquad \hat{\tau}_i^{BA} = \mathbf{1}_{(\text{i is accepted})} - \mathbf{1}_{(\text{i is } \widehat{\text{accepted}})} \quad \text{if } j(i) \text{ treated at time } t(i).\\
&\text{[Step 3]} \quad \hat{\tau}^{BA} = \frac{1}{N_{post}} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i^{BA}.
\end{aligned}
$$

Where N_{post} is the number of answers written by treated users.

Table 11 reports the estimates results. It shows that on average users are significantly more likely to have answers accepted once they can access the website in their native language.

	(1) TWFE	(2) TWFE 1	(3) TWFE 2	(4) TWFE 3	(5) BJS	(6) BJS 1	(7) BJS 2	(8) BJS 3
after	0.0211^{***}	0.0209^{***}	0.0203^{**}	0.00873	0.105^{***}	0.105^{***}	0.0931^{***}	0.0705^{***}
	(0.00245)	(0.00240)	(0.00244)	(0.00440)	(0.00425)	(0.00420)	(0.00340)	(0.00742)
Observations	293777	292919	292919	280407	293777	292846	292846	199564
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QEffort	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes	No	No	Yes	Yes
Empathy	No	No	No	Yes	No	No	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11: Treatment effects on the probability of having an answer accepted as "best answer". Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

6 Platform's trade-off

Knowledge platforms like Wikipedia and Stack Overflow, which aim to be global and maximize the quality of their content, have to decide whether they should allow the use of multiple languages. The introduction of multiple versions of the website in different languages has several implications and creates a nontrivial trade-off.

6.1 Benefit: increase in communication effort?

The main analysis discussed in this paper suggests that non-native English-speaking users benefit from a communication cost reduction if allowed to use their native language rather than English. This cost reduction significantly increases users' effort in information transmission. On average, using their native language rather than English, users include 0.66 additional pieces of code. Since the pre-treatment average is 2.71, the effect corresponds to a 24% increase in information quality.

Does this imply that information transmission becomes more effective? The analysis shows that, when a non-native English speaking user publishes an answer on her native-language website, she is 7% more likely that it gets *accepted.* This is a 20% increase compared to the pre-treatment average (35%). This shows that indeed native-language websites provide a substantial increase in social welfare for the Stack Overflow community, as the higher quality induces more positive outcomes.

6.2 Benefit: increase in community size?

For the platform, one of the advantages of introducing non-English websites is to potentially reach users who would not be participating otherwise. Table 12 shows that, out of nearly 93K users who published at least one answer/question on one of the non-English websites, 42.8% never registered on the English website. We could guess then that these users would not have joined the platform if their native language would not have been available.

In addition, 39.75% of users who registered on the English site before treatment (3.46% of the total) did not contribute before treatment. This suggests that also the contribution of these users may have been missing in absence of the non-English sites. Figure 11 shows the distribution of these users based on what type of participation they made after treatment. It is possible to see that the majority of these users participated only in their native-language site after treatment, and not in English.[31] This is consistent with the hypothesis that these users had a too high cost of using English to participate before treatment. At the same time, a quite substantial group of users started participating in English too, suggesting the presence of positive spillovers.

	After	Before	Not_registered	Tot
SOJ	1579	695	3588	5862
SOP	12178	3386	7800	23364
SOR	23661	279	23352	47292
SOS	7593	3720	5064	16377
Tot	45011	8080	39804	92895

Table 12: Number of active non-native English users who registered in the English website before treatment, after treatment, or did not register. Active means that they published at least an answer or question in the non-English websites of the corresponding row.

[31] Russian users do not follow the pattern. This anyway is due to the specificity of the history of the Russian website, as discussed in section 2.2

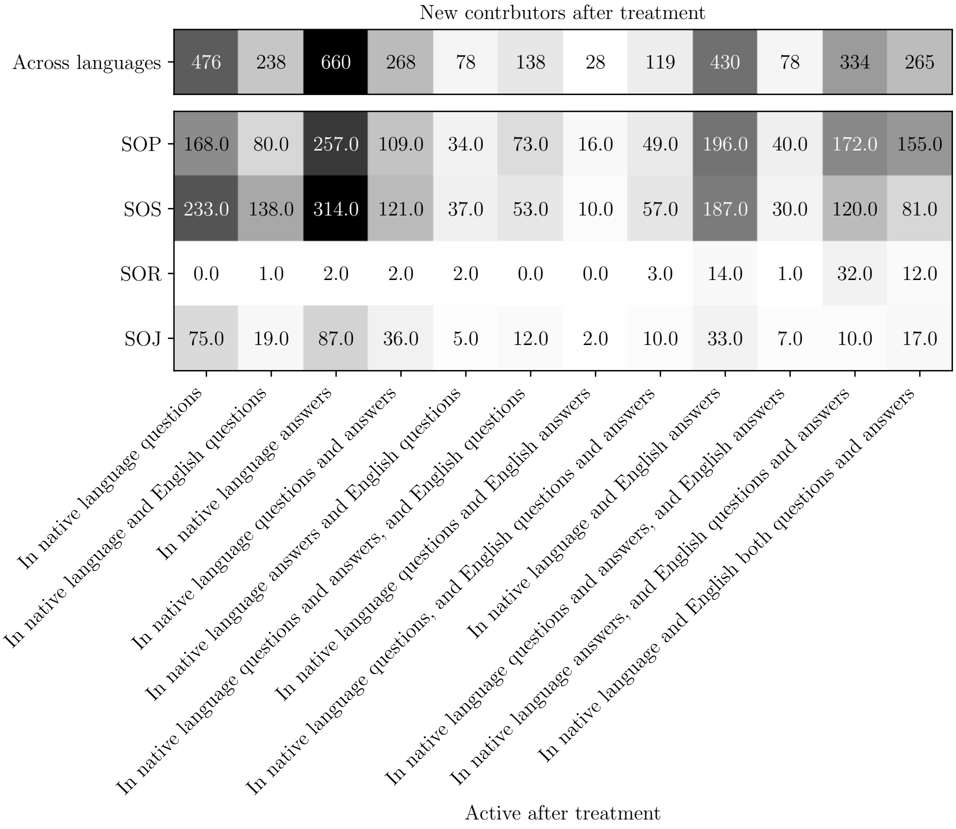


Figure 11: Sample of users who made at least one question and/or answer in a non-English language, who were registered in the English website before treatment, and who did not contribute any question/answer before treatment. Figure reports the number of such users based on what type of contribution they made after treatment.

6.3 Ambiguous: Externalities on the English platform?

It is reasonable to assume that users are time-constrained and cannot just increase participation with no boundaries. If that is the case, users with a cost of language high enough would switch to their native-language website, and substitute effort from the English to the native language website.

If these switching users are high expertise users, the absence of their contributions to the English website may reduce the overall welfare of the English platform. On the contrary, if switching users are low expertise users, the English website may see an increase in the average quality of its content.

To obtain a measure of the change in average quality produced by treated users, I estimate a simple OLS regression without user and time fixed effects. This estimation compares the average quality of answers produced by treated users in English with their English answers written before treatment, together with the ones of the control group.

For the estimation, I use the same sample described in section 4. By construction of the sample, the comparison groups may differ. Indeed, users may not participate in the English website after treatment.[32] The objective of this estimation is to measure the change in the average quality of answers produced by the treatment group, without controlling for a change in the composition of contributors. In this way, the estimate captures the change conditional on having, potentially, the best or worst contributors leaving the platform.

The estimating equation is the following:

$$numCodes_i = \beta D_{L(j(i),t(i))}\mathbf{1}_{(i \text{ in English})} + \theta D_{L(j(i),t(i))}\mathbf{1}_{(i \text{ NOT in English})} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i$$

Where $\mathbf{1}_{(i \text{ in English})}$ and $\mathbf{1}_{(i \text{ NOT in English})}$ identify, respectively, indicator functions that take value equal to 1 if the answer is written in the English site or not.

Table 13 reports the estimates for the β coefficient. Results show that on average answers' quality has increased in the English website after the introduction of non-English sites, but not significantly. This suggests that overall the platform has not faced significant externalities on the English website. The fact that the estimates are positive suggests that users leaving the English website in favour of their native-language one are, on average, less expert. This is also confirmed by figure 12, which shows that on average, users who left the English website after treatment were providing less quality content before treatment.

This evidence does not explain whether users who kept contributing to the English site also increased their effort on the English site, or they just kept the same effort. To investigate this question, I estimate the treatment effect on the English answers using both the TWFE and BJS estimation approaches. While I use the same approach as in section 5.2, the sample is different. To ensure that I observe users both before and after treatment, I select the sample of users that published at least 1) 1 answer on the English site before treatment, 2) one answer on a non-English site, and 3) 1 answer on

[32] The condition for a user to be part of the sample is to have published at least an answer in English before treatment, and an answer in a non-English site after treatment.

	(1) OLS	(2) OLS 1	(3) OLS 2	(4) OLS 3
after × InSo	0.841*	0.731*	0.713	0.681
	(0.263)	(0.260)	(0.258)	(0.261)
Observations	293777	292919	292919	280407
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QEffort	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes
Empathy	No	No	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13: Estimates of treatment effect on English answers' quality, without time and user fixed effects. Standard errors are clustered (*cse*) at the native-language level, i.e. at the treatment level.

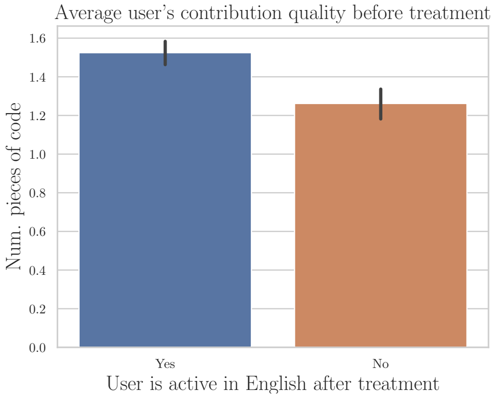


Figure 12: User-level average quality (number of pieces of code) of contributions before treatment, by participation in the English site after treatment.

the English site after treatment.[33] The estimating equation for the TWFE approach is the following:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \beta D_{L(j(i),t(i))}\mathbf{1}_{(i \text{ in English})} + \theta D_{L(j(i),t(i))}\mathbf{1}_{(i \text{ NOT in English})} + \boldsymbol{W}_i'\boldsymbol{\gamma} + \varepsilon_i$$

For what concern the BJS method, The steps are the same as described in section 5.2, with the exception of the last step, which is the following:

$$\hat{\tau} = \frac{1}{N_{eng}} \sum_{i|\text{j(i) treated at time t(i)}} \hat{\tau}_i \mathbf{1}_{(i \text{ in English})}$$

Where N_{eng} is the number of answers written on the English site by treated users.

[33] This last condition, i.e. that users were active in English after treatment, is what makes the sample different from the one used in the previous analysis. It induced a reduction of the sample size of around 10K observations (3%) compared to the data described in section 4.

Table 14 reports estimates results for β in the context of TWFE regressions, and τ in the context of BJS estimation. They suggest that the introduction of multiple languages had positive spillovers on the English website. The channel of this positive effect remains an open question.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	TWFE	TWFE 1	TWFE 2	TWFE 3	BJS	BJS 1	BJS 2	BJS 3
after × InSo	0.201^{**}	0.191^{**}	0.190^{**}	0.183^{*}	0.200^{***}	0.206^{***}	0.214^{***}	0.195^{*}
	(0.0238)	(0.0340)	(0.0336)	(0.0410)	(0.0585)	(0.0561)	(0.0554)	(0.0993)
Observations	284531	283710	283710	271349	231698	230814	230814	171863
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QEffort	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes	No	No	Yes	Yes
Empathy	No	No	No	Yes	No	No	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: Estimates of treatment effect on English answers' quality. Standard errors are clustered (cse) at the native-language level, i.e. at the treatment level.

These results show that, on the intensive margin, the platform is not suffering from negative externalities.

For what concerns externalities on the extensive margin instead, figure 13 shows how many users contributed a certain amount of answers in English after being treated, based on their contribution before treatment. It shows that a significant amount of users stopped contributing to the English website. In general, anyway, those are users who were already contributing little before, and users who were highly contributing before treatment kept contributing a lot in the post-treatment period. Finally, users who decreased the number of contributions were compensated by users who increased their participation. Note that the data is right-censored and for most users, the period before treatment was much longer than the period after. This means that these statistics may be downward-biased for the number of post-treatment contributions.

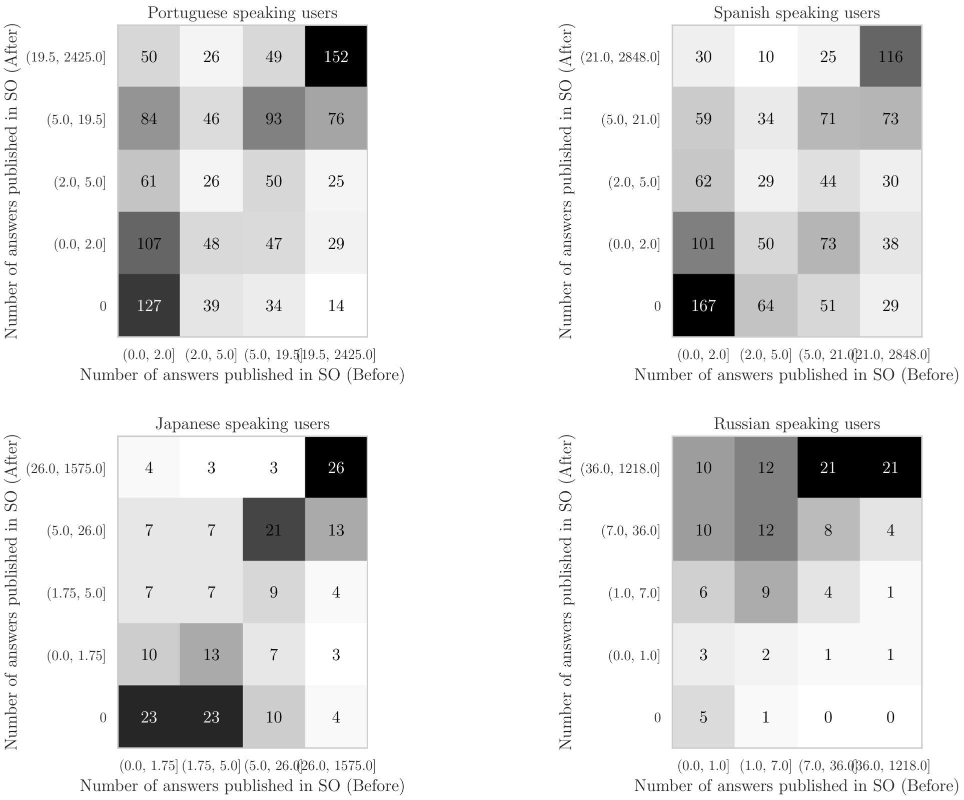


Figure 13: Distribution of users based on participation in English before and after treatment. Numbers in the plot correspond to the number of users in the sample who published 0 or more answers in English after treatment, based on their contribution before treatment. Intervals are based on the 0.25, 0.5, and 0.75 quantiles of the distributions of contributions before treatment.

6.4 Cost: increase in misleading answers?

If communication costs act as a barrier to participation for users not very expert on the topics of the questions, a reduction in communication costs can lead to more imprecise answers. To show this through the theoretical framework, let λ''_A be the new level of the cost of language, and λ'_A the initial level, where $\lambda''_A < \lambda'_A$.

Communication effort is positive, i.e. it results in a published answer, if the following condition is satisfied:

$$\sqrt{\gamma} k_A > s\lambda_A.$$

In other words, the user provides an answer if the cost of language is sufficiently lower than her expertise, k_A. From the platform perspective anyway, a good answer is an answer that is both well written and accurate. The platform would then like that all participating users would have at least a minimum level of expertise, say $\bar{k}_A$.

A reduction in the cost of language may induce an increase in the number of answers by users that do not satisfy a minimum level of k_A. In fact, let $\hat{k}_A < \bar{k}_A$. Then, if:

$$\lambda''_A < \frac{\sqrt{\gamma}\hat{k}_A}{s} < \lambda'_A,$$

a user with an insufficient level of expertise would not answer questions on the English website, but she would provide answers on her native language website.

This implies that, if the distribution of expertise across users is the same for different levels of cost of using English, then, on average, who do not contribute in English before treatment but do contribute in their native language when available, have lower expertise. We should then observe that those users contribute to the native-language website with lower quality answers compared to users who were active in English before treatment.

Figure 14 shows that indeed this is the case. On average, users who were registered but not active before treatment provide lower-quality contributions compared to users who were active before treatment.

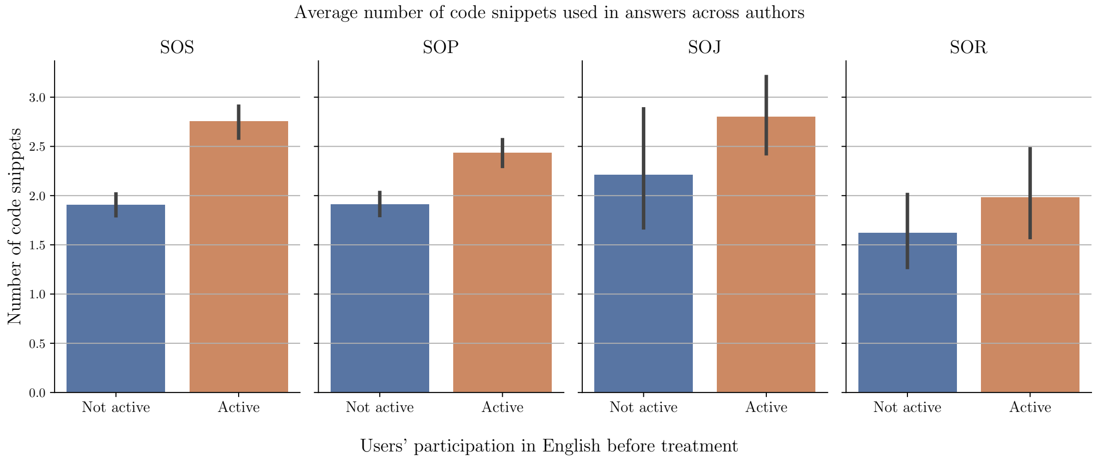


Figure 14: Sample of users who made at least an answer in a non-English language, and who were registered on the English website before treatment. The figure reports the average of the average message quality of each user's contributions, based on whether the user has published answers on the English website before the native-language website became available. Message quality is measured as the number of code snippets appearing in the answer. Vertical black lines are confidence intervals computed via bootstrapping.

6.5 Cost: reduction in knowledge aggregation?

From an economic perspective, to have information shared in a multiplicity of languages is inefficient. The use of the same language would allow to maximize the aggregation of information and minimize search costs. Nevertheless, as noted by the Stack Overflow team itself, imposing a language over the others would exclude people who cannot learn that language, and would mean deciding arbitrarily what language should be the only one.[34]. This trade-off between efficiency and ethics is not only relevant for Stack Overflow but in general on any discussion about centralization versus decentralization of languages (Ginsburgh and Weber 2011, Blanc and Kubo 2021, Blouin and Dyer 2022). For what concern knowledge platform and Wikipedia in particular, the literature has indeed found that the multiplicity of websites caused the dispersion of information (Bao et al. 2012).

To test if it is the case also for Stack Overflow, I first identify a list of all existing programming languages. I retrieve this list from Wikipedia, which lists 677 programming languages.[35] To avoid confusion with natural languages, let me call the programming languages PLs. I then check if, for each of these PLs, there exists a tag in the Stack Overflow websites. In Stack Overflow, tags are used to categorize the content of the questions. This implies that, if a tag exists, then at least one question has addressed that topic. If a tag exists in some languages, but not in others, it means that only the community of that language has addressed that topic.

Table 15 shows the number of PLs that appeared in 0, 1, 2, 3, or 4 languages, where the languages are Spanish, Portuguese, Russian, and Japanese, and whether they also appeared in English. It shows that out of the 677 PLs, only 28 of them appear in all 5 languages (including English). Out of the 247 PLs discussed in at least one language, 219 are discussed only in some of the languages, meaning that the information is not accessible for users not speaking those languages. In addition, 83 PLs are discussed in at least two languages, suggesting that there could be efficiency gains if everyone would speak the same language. Finally, 12 PLs are discussed only in languages different from English. This suggests the potential risk that the implementation of additional languages has reduced the variety of information in English. These results are consistent with the findings by Jia, Tumanian, and Li (2021).

Number of non-English languages with the tag	0.0	1.0	2.0	3.0	4.0
Whether tag is in English site					
0.0	430	8	3	1	
1.0	152	29	17	9	28

Table 15: Number of programming languages for which at least a question has been made in 1, 2, 3, or 4 languages. The 4 languages are Spanish, Portuguese, Russian, and Japanese. Rows split the sample based on whether the tag appears in the English website (1) or not (0)

7 Conclusion

This paper studies the trade-off faced by knowledge platforms when deciding to make their website available in either one or multiple languages.

It shows that the benefits of allowing contributions in multiple languages are substantial, mainly because it reduces communication costs for the users native to those

[34] https://stackoverflow.blog/2014/02/13/cant-we-all-be-reasonable-and-speak-english/

[35] List retrieved from https://en.wikipedia.org/wiki/List_of_programming_languages

languages. On one side, I show that at least 42% of those users were unlikely to participate if their native-language website was not available. This is relevant in showing that participation decisions may be limited by language barriers. On the other, users increase by 24% their communication effort after their native language becomes available. This effect is driven by users that, after their native language became available, have switched the most to it, reducing contributions in English. These users are likely to be the ones with the highest cost of using English, and who then faced the largest drop in communication costs by using their native language.

In addition, the increase in effort due to a reduction in the cost of language is positively correlated to the questioners' effort and incentives. When answering in their native language, users increase effort by up to 34% if the questions are in the top quartile by quality, and up to 110% if they are highly incentivized via virtual remuneration. This suggests that strategies and policies that aim to reduce the cost of language to favour information transmission may be under-effective if not paired with incentives on both sides of the communication flow.

The paper then shows that there are no clear negative externalities for the English website, but rather positive: even if, overall, quality is not significantly different, the users who keep contributing to the English website after treatment increase their effort in English. For what concerns the extensive margin instead, a substantial amount of users stop participating in English when their native language is available. These users anyway were not very active before treatment, and users that were contributing a lot in English kept a high level of participation.

According to this evidence, it seems advisable to introduce websites in multiple languages, as it increases the community size and the quality of the information collected. Nevertheless, the paper shows also some drawbacks. First of all, the new inflow of participation induced by the availability of additional languages is characterized by lower-quality contributions, which is reducing the overall improvement in information quality. This is justifiable by the fact that a high cost of language acts as a barrier to participation for inexpert users. Second, there is naturally a decrease in efficiency in information aggregation. If the same topic is addressed in more than one language it means that multiple users spent time and effort to potentially provide the same piece of information. At the same time, if some information is provided only in some languages but not others, then some users are not able to access it. Both issues would be solved by imposing a single language. It follows that, if the communities of non-English speakers are small and few people would benefit from multiple languages, a single language is preferable. This anyway would raise ethical concerns, as it would exclude minorities or constrain their access to the platform (Jeon, Jullien, and Klimenko 2021).

Overall is not clear what is the optimal strategy, which then depends on the long-term objective of the platform (e.g. how global it wants to be) and the size of the communities using given languages. It seems wise for Stack Overflow to have implemented additional websites for only some of the most common languages outside English.

While the analysis is specific to the context of Stack Overflow, the results may contribute to different environments. A large literature has addressed communication costs as a major constraint to efficient economic activities, but, to my knowledge, it has not quantified the problem. This is relevant anyway for a variety of decision-makers. To give a few examples, when firms need to form teams of employees of different nationalities, they need to assess the advantages of pairing co-workers of the same nationality (Lyons 2017, Corritore, Goldberg, and Srivastava 2020). In defining the hierarchy structure of the company, managers need to evaluate the advantage of hiring *translators* or imposing the same language across teams (Crémer et al. 2007). Finally, national states may want to understand the exact benefit of imposing a homogeneous language before taking initiative toward centralization.

This paper is silent on how organizations could compensate and alleviate part of the trade-off using external technologies, as shown for instance on eBay product titles by Brynjolfsson, Hui, and Liu (2019). Indeed, live translations and search engines that allow for searches across languages may solve the trade-off.[36] Many issues anyway may reduce the benefits of those technologies. For example, a lot of expressions and concepts require a complete rewriting to convey the same message in different languages, something that only human translators can achieve.[37] Future research should then be devoted to understanding to what extent existing or potential future technologies could be instrumental.

Finally, future work should be devoted to investigating the external validity of these findings in the context of face-to-face communications with personal interactions.

References

Arrow, K. J. (1974). *The Limits of Organization.* W. W. Norton & Company, Inc. 2

Asher, N. and A. Lascarides (2013). Strategic conversation. *Semantics & Pragmatics 6*(2), 1–62. 2

Austen-Smith, D. and J. S. Banks (2000). Cheap talk and burned money. *Journal of Economic Theory 91*(1), 1–16. 2

Baker, A. C., D. F. Larcker, and C. C. Y. Wang (2022, May). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics 144*(2), 370–395. 23

Bao, P., B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle (2012, May). Omnipedia: bridging the wikipedia language gap. *CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1075–1084. 4, 37

Battiston, D., J. Blanes I Vidal, and T. Kirchmaier (2021, March). Face-to-face communication in organizations. *Review of Economic Studies 88*(2), 574–609. 5

[36] Kitenge and Lahiri (2021) find that in international trade, the advantage of accessing the internet is lower if trading partners use a similar language. The availability of machine translations may explain this fact.

[37] Calefato, Lanubile, and Minervini (2010) find that, in a context with few speakers, a real-time translation that employs Google translate service would produce 62% of adequate translations.

BenYishay, A. and A. M. Mobarak (2019, May). Social Learning and Incentives for Experimentation and Communication. *The Review of Economic Studies 86*(3), 976–1009. 21

Blanc, G. and M. Kubo (2021). Schools, language, and nations: Evidence from a natural experiment in France. *Working paper*. 5, 37

Blouin, A. and J. Dyer (2022). How cultures converge: An empirical investigation of trade and linguistic exchange. *Working paper*. 37

Blume, A. (2018, May). Failure of common knowledge of language in common-interest communication games. *Games and Economic Behavior 109*, 132–155. 2

Blume, A. and O. Board (2013, March). Language barriers. *Econometrica 81*(2), 781–812. 2

Blume, A., D. V. DeJong, Y.-G. Kim, and G. B. Sprinkle (2001). Evolution of communication with partial common interest. *Games and Economic Behavior 37*, 79–120. 5

Bolton, P. and M. Dewatripont (1994, November). The firm as a communication network. *The Quarterly Journal of Economics 109*(4), 809–839. 2

Borusyak, K., X. Jaravel, and J. Spiess (2021). Revisiting event study designs: Robust and efficient estimation. *Working Paper*. 4, 22, 23, 47, 48

Boudreau, K. J., T. Brady, I. Ganguli, P. Gaule, E. Guinan, A. Hollenberg, and K. R. Lakhani (2017, October). A Field Experiment on Search Costs and the Formation of Scientific Collaborations. *The Review of Economics and Statistics 99*(4), 565–576. 6

Brynjolfsson, E., X. Hui, and M. Liu (2019, December). Does Machine Translation Affect International Trade? Evidence from a Large Digital Platform. *Management Science 65*(12), 5449–5460. 39

Calefato, F., F. Lanubile, and P. Minervini (2010, August). Can Real-Time Machine Translation Overcome Language Barriers in Distributed Requirements Engineering? In *2010 5th IEEE International Conference on Global Software Engineering*, pp. 257–264. ISSN: 2329-6313. 39

Callaway, B. and P. H. C. Sant'Anna (2020, December). Difference-in-Differences with multiple time periods. *Journal of Econometrics*. 4, 22

Calvó-Armengol, A., J. de Martí, and A. Prat (2015). Communication and influence. *Theoretical Economics 10*, 649–690. 7

Chen, S., R. Geluykens, and C. J. Choi (2006, December). The importance of language in global teams: A linguistic perspective. *Management International Review 46*(6), 679. 2

Corritore, M., A. Goldberg, and S. B. Srivastava (2020, June). Duality in Diversity: How Intrapersonal and Interpersonal Cultural Heterogeneity Relate to Firm Performance. *Administrative Science Quarterly 65*(2), 359–394. 39

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. *Econometrica 50*(6), 1431–1451. 2

Crémer, J., L. Garicano, and A. Prat (2007). Language and the theory of the firm. *The Quarterly Journal of Economics 122*(1), 373–407. 2, 5, 39

de Chaisemartin, C. and X. D'Haultfœuille (2020, September). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review 110*(9), 2964–2996. 4, 22

Dewatripont, M. and J. Tirole (2005). Modes of communication. *Journal of Political Economy 113*(6), 1217–1238. 2

Diamond, J. (1991). *The Third Chimpanzee, The Evolution and Future of the Human Animal*. Hutchinson Radius. 2

Dilmé, F. (2018, January). Optimal languages. *Working Paper*. 2

Gambetta, D. (2011). *Codes of the Underworld: How Criminals Communicate*. Princeton University Press. 2

Ginsburgh, V. and S. Weber (2011, April). *How Many Languages Do We Need?: The Economics of Linguistic Diversity*. Princeton University Press. 2, 5, 37

Ginsburgh, V. and S. Weber (2020, June). The Economics of Language. *Journal of Economic Literature 58*(2), 348–404. 21

Goldfarb, A. and C. Tucker (2019, March). Digital Economics. *Journal of Economic Literature 57*(1), 3–43. 6

Goodman-Bacon, A. (2021, December). Difference-in-differences with variation in treatment timing. *Journal of Econometrics 225*(2), 254–277. 22

Guillouët, L., A. K. Khandelwal, R. Macchiavello, and M. Teachout (2021). Language barriers in multinationals and knowledge transfers. *Working paper*. 5

Jeon, D.-S., B. Jullien, and M. Klimenko (2021, July). Language, internet and platform competition. *Journal of International Economics 131*, 103439. 38

Jia, J., V. Tumanian, and G. Li (2021, October). In favour of or against multi-lingual Q&A sites? Exploring the evidence from user and knowledge perspectives. *Behaviour & Information Technology 40*(13), 1390–1405. 37

Kitenge, E. and S. Lahiri (2021). Is the Internet bringing down language-based barriers to international trade? *Review of International Economics n/a*(n/a). 39

Lafky, J. and A. J. Wilson (2020, January). Experimenting with incentives for information transmission: Quantity versus quality. *Journal of Economic Behavior and Organization 169*, 314–331. 5

Lohmann, J. (2011, February). Do language barriers affect trade? *Economics Letters 110*(2), 159–162. 5

Lyons, E. (2017, July). Team Production in International Labor Markets: Experimental Evidence from the Field. *American Economic Journal: Applied Economics 9*(3), 70–104. 21, 39

Marschak, J. and R. Radner (1972). *Economic theory of teams.* Yale University Press. 2

McManus, W. S. (1985). Labor Market Costs of Language Disparity: An Interpretation of Hispanic Earnings Differences. *The American Economic Review 75*(4), 818–827. 5

Melitz, J. (2008, May). Language and foreign trade. *European Economic Review 52*(4), 667–699. 5

Sandvik, J. J., R. E. Saouma, N. T. Seegert, and C. T. Stanton (2020, 04). Workplace Knowledge Flows. *The Quarterly Journal of Economics 135*(3), 1635–1680. 6

Sobel, J. (2013). Giving and receiving advice. In D. Acemoglu, M. Arellano, and E. Dekel (Eds.), *Advances in Economics and Econometrics*, Econometric Society Monographs, Chapter 10. Cambridge University Press. 2

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics 87*(3). 2

Sun, L. and S. Abraham (2020, December). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*. 4, 22

Tainer, E. (1988). English Language Proficiency and the Determination of Earnings among Foreign-Born Men. *The Journal of Human Resources 23*(1), 108–122. 5

Tenzer, H., M. Pudelko, and A.-W. Harzing (2014, June). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies 45*(5), 508–535. 2

Zhang, X. M. and F. Zhu (2011, June). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *The American Economic Review 101*(4), 1601–1615. 4

Appendix A Additional details on Stack Overflow

A.1 The introduction of new websites

The creation of new websites of Stack Overflow follows a specific process. The main objective is to ensure, before the launch, a sufficiently active community base that will guarantee the growth and the sustainability of the website in the long run. First of all the website is proposed in an ad-hoc platform called *Area 51*, where users registered can support the proposal and start publishing questions and answers. If the website idea receives enough attention and contributions, then it proceeds to the *beta* period, it gets its URL, and it is accessible as an independent site. The *beta* period is split into two steps: first, in the so-called *private beta*, only users that were active in supporting it in the early stage can contribute. Then, when it becomes *public beta*, everyone can register and contribute. Once all features are implemented, the website is said to *graduate*, entering its final stage. At each stage, the incentive system may slightly vary. For example, some *privileges* are reachable with different with different amounts of points, generally being lower requirements in earlier stages.

Data is available starting from the *private beta* period. Table 16 reports the dates for the start of each stage for websites in different languages.

platform	**proposal**	**private beta**	**public beta**	**graduation**
SO		01/08/2008	-	15/09/2008
SO - R	01/06/2012	27/03/2015	27/03/2015	11/12/2015
SO - J	-	29/09/2014	16/12/2014	[not graduated]
SO - S	02/08/2012	01/12/2015	15/12/2015	17/5/2017
SO - P	05/11/2010	12/12/2013	29/01/2014	15/5/2015

Table 16: Dates in which the platforms passed the different development stages. SO correspond to Stack Overflow in English, while the initials R, J, S, P stand for Russian, Japanese, Spanish, and Portuguese, respectively.

Appendix B Details about the theoretical framework

B.1 Second stage

$$\begin{aligned}
& a^* \equiv \arg\max_a \mathbb{E}[-\left((a-\theta)^2 + C_Q^2 E_Q\right)|m] \\
\iff & a^* \equiv \arg\max_a -a^2 - \mathbb{E}[\theta^2|m] + 2a\mathbb{E}[\theta|m] + C_Q^2 E_Q \\
\iff & -2a^* + 2\mathbb{E}[\theta|m] = 0 \\
\iff & a^* = \mathbb{E}[\theta|m] = \beta m \quad \text{with} \quad \beta \equiv \frac{E_Q E_A}{E_Q E_A + E_Q s + E_A s}
\end{aligned}$$

where the last equality holds because of Bayes Normal updating.

B.2 First stage

$$\max_{E_A \geq 0} \mathbb{E}[-\left(\gamma(a-\theta)^2 + C_A^2 E_A\right)]$$

Given the action expected to be chosen by Bob, the problem rewrites as:

$$\begin{aligned}
&\max_{E_A \geq 0} -\gamma \mathbb{E}[(\beta m - \theta)^2] - C_A^2 E_A \\
\iff &\max_{E_A \geq 0} -\gamma \mathbb{E}[\beta m - \theta]^2 + -\gamma \mathbb{V}[\beta m - \theta] - C_A^2 E_A \\
\iff &\max_{E_A \geq 0} -\gamma \mathbb{V}[\beta m - \theta] - C_A^2 E_A \\
\iff &\max_{E_A \geq 0} -\gamma \left(\beta^2 \frac{1}{s} + \beta^2 \frac{1}{E_A} + \beta^2 \frac{1}{E_Q} + \frac{1}{s} - 2\beta \frac{1}{s} \right) - C_A^2 E_A \\
\iff &\max_{E_A \geq 0} -\gamma \left(\beta^2 \frac{1}{\beta s} + \frac{1}{s} - 2\beta \frac{1}{s} \right) - C_A^2 E_A \\
\iff &\max_{E_A \geq 0} -\gamma \left(\frac{1}{s}(1 - \beta) \right) - C_A^2 E_A
\end{aligned}$$

and

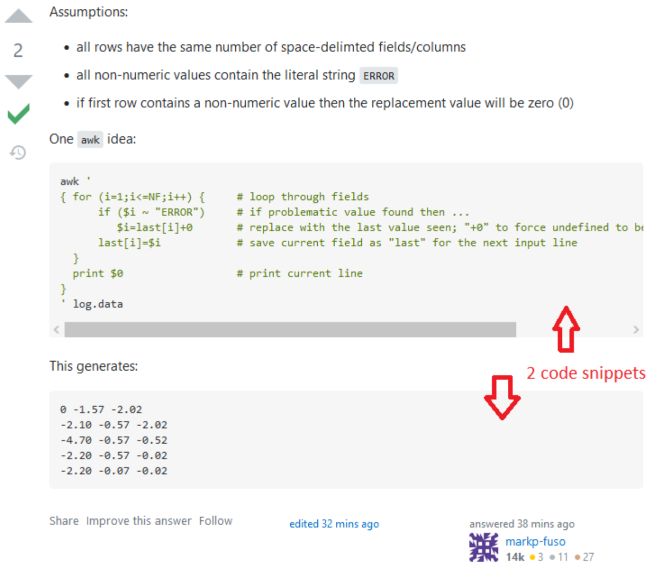
$$\begin{aligned}
\frac{\gamma E_Q^2}{(E_Q E_A + E_Q s + E_A s)^2} &= C_A^2 \\
(E_Q E_A + E_Q s + E_A s)^2 &= \frac{\gamma E_Q^2 k_A^2}{\lambda_A^2} \\
E_A(E_Q + s) &= \frac{E_Q(\sqrt{\gamma} k_A - s\lambda_A)}{\lambda_A}
\end{aligned}$$

The best response is then given by:

$$R(E_Q) = \frac{E_Q(\sqrt{\gamma} k_A - s\lambda_A)}{\lambda_A(E_Q + s)}$$

Appendix C Additional details about the data and the measures

C.1 Quality measure



Assumptions:

- all rows have the same number of space-delimted fields/columns
- all non-numeric values contain the literal string `ERROR`
- if first row contains a non-numeric value then the replacement value will be zero (0)

One `awk` idea:

```
awk '
{ for (i=1;i<=NF;i++) {     # loop through fields
      if ($i ~ "ERROR")     # if problematic value found then ...
         $i=last[i]+0       # replace with the last value seen; "+0" to force undefined to be
      last[i]=$i            # save current field as "last" for the next input line
  }
  print $0                  # print current line
}
' log.data
```

This generates:

```
0 -1.57 -2.02
-2.10 -0.57 -2.02
-4.70 -0.57 -0.52
-2.20 -0.57 -0.02
-2.20 -0.07 -0.02
```

Figure 15: Example of an answer in Stack Overflow where the number of snippets (i.e. the proxy for quality) is equal to 2.

C.2 Participation choices

Appendix D Robustness for DiD analysis

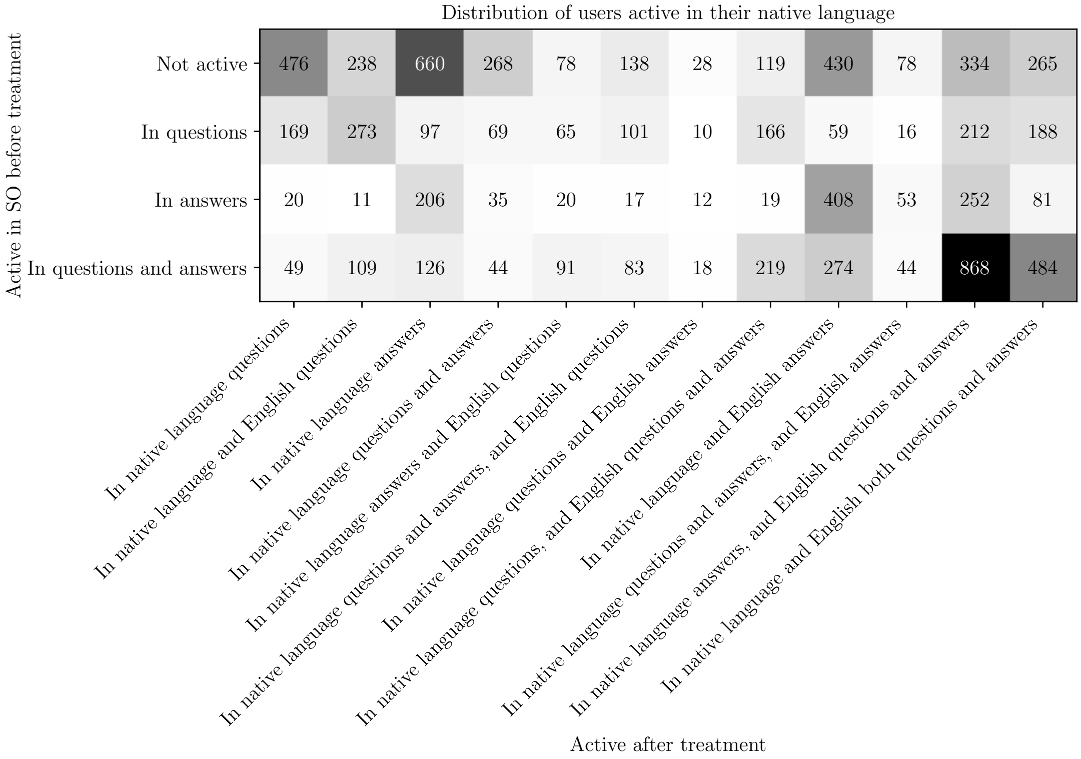


Figure 16: Distribution of users based on what type of contributions they have made. The sample conditions for 1) The user must have registered in Stack Overflow (English) before being treated, and 2) The user participated with at least one question/ answer in a non-English website. The y-axis identifies contributions made before being treated (i.e. in English only), while the x-axis the contributions made after treatment.

D.1 removing answers with 0 pieces of code

Table 17 reports regression results comparable to the estimation in table 4 after dropping all answers with zero pieces of code and selecting users that, given the remaining answers, where active both before and after treatment.

	(1) TWFE	(2) TWFE 1	(3) TWFE 2	(4) TWFE 3	(5) BJS	(6) BJS 1	(7) BJS 2	(8) BJS 3
after	0.442^{*}	0.434^{*}	0.434^{*}	0.209	0.838^{***}	0.864^{***}	0.866^{***}	0.910^{***}
	(0.141)	(0.147)	(0.147)	(0.0743)	(0.0366)	(0.0358)	(0.0344)	(0.0776)
Observations	228244	227818	227818	218908	228244	227812	227812	152843
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QEffort	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes	No	No	Yes	Yes
Empathy	No	No	No	Yes	No	No	No	Yes

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 17: Baseline Regressions' estimates where the dependent variable is the number of pieces of code, after dropping all answers with zero pieces of code. The estimate *after* corresponds to the average treatment effect, and corresponds to the parameters $\hat{\beta}$ or $\hat{\tau}$ if the specification adopted is the TWFE or the BJS respectively. *cse* represents the level at which the standard errors have been clustered: either at the users' native language (i.e. the level at which the treatment takes place) or at the user level.

D.2 Alternative way to compute the average treatment effect

Since the panel is unbalanced, the Borusyak et al. (2021) based treatment effect overweights the effort choice of users who are more active in the post-treatment period. This is simply because in the main analysis the treatment effect is computed as a equally-weighted average of each observation's treatment effect. Users with more answers (i.e. observations) in the post-period will have a stronger weight in the final average. This is not necessarily a bad thing: the platform may want to take into consideration the different level of contributions, as these communities show a lot of heterogeneity in user types.

An alternative option, to avoid the overweighting of the most active users, is to first average within each user, and then take the average across users. More specifically, the third step in the estimation process would become, with j indexing users and t indexing time:

$$\text{[Step 3]} \quad \hat{\tau} = \frac{1}{J}\sum_{j}\left(\frac{1}{N_k}\sum_{i|k(i)\,=\,j(i),k(i)\text{ treated at time t}}\hat{\tau}_i\right).$$

where N_k is the number of answers made by user k after she was treated, and J is

	(1)	(2)	(3)	(4)
Treatment effect	0.057	0.077	0.086	0.146
Controls				
Qeffort	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes
Empathy	No	No	No	Yes

Table 18: Point estimates for the treatment effect using an alternative way to compute the final Average Treatment Effect, still using nevertheless the approach by Borusyak et al. (2021). More specifically, once I compute the treatment effect on each observations, instead of taking an overall average across observations, I first average within each user, so to obtain the average treatment effect at the user level. I then compute the average across users. This approach weights equally the users, while the approach presented in the paper gives more weight to the users that are more active in the post-treatment period.

the total number of users. The estimates computed in this way are presented in table 18. They show a smaller but still positive effect.