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

Jacopo Bregolin*
University of Liverpool

March 2024

Abstract

The transmission of information is crucial for productivity and growth. However, language differences may limit its effectiveness. This is particularly relevant for knowledge platforms that aim to be global, given the cultural diversity of the pool of users. In this paper, I empirically investigate how the exogenous cost of language affects communication quality and the trade-offs faced by knowledge platforms in implementing their website in multiple languages. I exploit the staggered introduction of websites for languages other than English on a question-and-answer platform to demonstrate two main facts. First, non-native English speakers who contribute in English increase their answers' quality once able to use their native language, and their answers are more likely to solve the questioner's problem. The native-language answers drive the effect, which is larger when the question quality is higher and the incentives are stronger. Second, users who participate in their native language but not in English contribute lower-quality answers compared to those who contribute in English. This suggests that lower language barriers allow users with less expertise in the subject to participate. These results show that the platform should adopt multiple languages to maximise the quality of the information collected, although doing so may also result in an inflow of low-quality content from new users.

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

*email: jacopo.bregolin@liverpool.ac.uk. I am extremely grateful to three anonymous referees, the associate editor, and the editor for comments that substantially improved the paper. I would like to thank for helpful discussions and feedback Patrick Bennett, Milo Bianchi, Gordon Burtch, Daniel L. Chen, Jacques Crémer, Daniel Ershov, April Franco, Guido Friebel, Astrid Hopfensitz, Jakub Lonsky, Nicolas de Roos, Marta Troya-Martinez, Shalini Mitra, and Bobby Zhou. I thank for useful comments participants at the University of Liverpool Management School internal seminar, Digital Economy Network workshop, Munich Summer Institute, SIOE, Platform Strategy Research Symposium, and ESEM.

1 Introduction

Communication as a way to transmit information is fundamental for the development of our society. The sharing of knowledge allows people to take advantage of each others' human capital investments, speeding up learning and productivity. Knowledge platforms that aggregate information, like Wikipedia and Stack Overflow, play a crucial role in this respect as they can reach large audiences and leverage a global pool of knowledge. However, both online and offline, the quality of the shared information can be limited in several ways. On the one hand, communication can be affected by the too little incentives of the information holder. On the other hand, exogenous cognitive boundaries, such as the lack of proficiency in the language used, can constrain our ability to share information.[1] These issues are detrimental to knowledge platforms as content quality determines their success.[2] Some platforms use incentive systems such as virtual rewards to address the first problem and allow their users to communicate in several languages to address the second problem. While a large literature investigates the effectiveness of incentives, research on the identification of platforms' trade-offs when implementing their site in multiple languages is sparser.

This paper contributes to the literature by investigating the quality trade-offs that knowledge platforms face when they allow their users to use multiple languages rather than just English. Intuitively, lower language barriers may let existing users provide higher-quality contributions and let new users join the platform. Quality trade-offs arise if new users have less expertise and provide lower-quality content. This paper tests these hypotheses studying if and to what extent existing users provide higher-quality information when they are allowed to use their native language rather than English, and whether new users provide lower quality contributions compared to existing users.

The analysis leverages data from Stack Overflow, an English-based question-and-answer website focused on topics related to computer programming. The platform introduced non-English versions of the website in a staggered fashion, allowing the researcher to observe contributions from two sets of users: those who were active in English and started to contribute in their native language once it became available, and those that

[1] The economic theory literature has investigated both constraints but generally focuses on one or the other. On one side, the literature has looked at incentives and strategic information transmission, either with costless 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 side, the team theory literature (Marschak and Radner 1972) has focused on environments where incentives are perfect but exogenous constraints affect the ability to communicate, e.g. 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 combine these two strands and examine the interaction between incentives and exogenous costs.

[2] Stack Overflow is probably the primary source of help for computer programmers. Solutions provided in the platform influence programmers' code worldwide, so mistakes or inaccurate information can have a significant impact. For instance, the most copied code snippet from Stack Overflow contained an error: https://programming.guide/worlds-most-copied-so-snippet.html. While advanced generative AI tools are potentially taking over as the main source of programmers' support, Stack Overflow content is most likely fed into AI algorithms.

registered in the English site but did not contribute, and only participated in their native language when allowed. I refer to the first set of users as *old joiners*, and to the second set of users as *new joiners*. In particular, the researcher can observe *old joiners*'s contributions before and after their native language became available, and compare *old joiners*' contributions with those of *new joiners* within the native-language sites.

The paper shows two main facts. First, once *old joiners* are allowed to use their native language, both their answers' quality and the probability that their answers solve the questioner's problem increase. These effects are driven by answers written in the native language, and increase with answerer incentives and the question quality. Second, *new joiners*'s contributions in non-English languages are of lower quality compared to *old joiners*' contributions, possibly because of *new joiners*'s lower expertise. These results suggest that by implementing multiple languages, the platform reduces users' communication costs. This encourages users to produce quality content but also allows for the participation of inexpert contributors.

I illustrate the trade-off with a simple theoretical framework of communication between two users, Bob and Alice, where Alice may provide the information that Bob needs to complete a task. Alice decides her communication quality based on how much she internalises Bob's utility, the question's quality, her expertise, and the cost of language.[3] The framework highlights two aspects. First, the quality of Alice's answer depends on her cost of using the language available, which is potentially high if it is not her native language. Second, her participation is subject to having sufficiently high expertise relative to the language cost. It follows that, for certain levels of expertise and cost, the availability of her native language would 1) increase her answer's quality if she has high expertise or 2) allow her to participate if she has low expertise. Overall, the framework suggests that introducing multiple languages would increase the quality of contributions from existing users. However, it would also allow for the participation of new contributors who may provide lower-quality content.

To test these predictions, I use the staggered introduction of new language sites by Stack Overflow. The platform was created in 2008 in English; over time, it sequentially implemented additional websites in Russian, Portuguese, Japanese, and Spanish with the same purpose and function as the original website. A unique identifier for each user allows the researcher to observe users' contributions across sites.

To proxy for communication quality, I look at two empirical measures. One measure is based on the users' communication content and text characteristics; the second one measures 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 suggest higher quality.[4] The second measure exploits the fact that authors of the questions can

[3] The cost of language represents the degree of difficulty in using a particular language for communication. In the context of this paper, it mostly corresponds to the inverse of the degree of proficiency in that language. I use this terminology throughout the paper to be consistent with its role as a cost in the theoretical problem.

[4] Note that the measure is not based on the length of the code. Generally, a shorter code is more

accept one of the answers they receive if they consider it satisfactory. This action is not mandatory, so it can reliably indicate whether the questioner was able to solve his problem with the information received. I then measure how much the answerer was incentivised by exploiting the *bounties* website feature. Stack Overflow users can auction reputation points (i.e. virtual rewards) on given questions. In other words, they can reward the author of an acceptable answer. A higher number of points at stake implies more substantial incentives.

The analysis develops two main empirical strategies. The first strategy aims to estimate how lower language barriers affect communication quality, while the second strategy aims to test whether new users provide lower-quality content compared to existing users. More precisely, the first empirical analysis studies how the availability of *old joiners*' native languages impacts their contributions' quality. It relies on a staggered difference-in-difference approach, where a user becomes treated when the platform implements the site in the user's native language. 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 (2022) and compare the results with the more standard two-way fixed effects approach.[5] I study the complementarities and heterogeneity of this effect by interacting the treatment dummy with different levels of relevant variables. The second empirical strategy aims to compare the answers of *new joiners* with those of *old joiners* on the non-English sites. By exploiting the fact that questions can receive multiple answers, this strategy consists of a question fixed effects ordinary least squares regression, where answer quality is the dependent variable and different dummy variables identify whether the author is a *new joiner* or an *old joiner*. I also include order-of-publication fixed effects and control for the amount of time the user has been registered on the platform to capture the experience of the platform's functioning.

The paper shows that the trade-off is confirmed in the data. The introduction of native language websites increases *old joiners*' communication quality by 21.4% overall, with quality in native-language answers being 56.5% higher than in pre-treatment English answers. The overall effect increases to 118% if users are highly incentivised and to more than 30% when the question's quality is in the top quartile. Answers are also 17.2% more likely to be *accepted* by the questioner in this case, suggesting that higher quality leads to more effective communication. The estimates suggest substantial heterogeneity: users who are mainly active in English after treatment have close to a null effect, while users who are primarily active in their native language increase answer quality by up to 76.2%. By contrast, the effect is relatively homogeneous across different levels of pre-treatment participation and across non-English languages. These results suggest that the platform benefits from allowing communication in multiple languages.

efficient, so code length is not a good proxy for answer quality.

[5] In a nutshell, Borusyak et al. (2022) first estimate the treatment effect for every treated observation and then obtain the average treatment effect on the treated (ATT) by averaging across them. This approach solves econometric issues identified in the literature (Callaway and Sant'Anna 2020, de Chaisemartin and D'Haultfœuille 2020, Sun and Abraham 2020, Borusyak et al. 2022).

Nevertheless, for questions in the native language, answers by *new joiners* have 0.46 standard deviations lower quality compared to *old joiners* and 13.8 percentage points lower probability of being *accepted*. While other reasons may explain why *new joiners* do not participate on the English site, the substantial difference in answer quality suggests that the availability of sites in the users' native languages allows for the participation of users with lower expertise.

Overall, knowledge platforms that target quality highly benefit from multiple languages as non-native English users increase the quality and effectiveness of contributions. Nevertheless, these benefits are reduced when the non-native English users are very few, have a low cost of using English, or have low expertise. In this case, from an efficiency standpoint, a single language (or a limited number of languages) is preferable.[6]

While this trade-off is particularly important for knowledge platforms, it is also relevant to other economic and managerial environments. An example is national states (Ginsburgh and Weber 2011) that can impose a single language, like in France (Blanc and Kubo 2021), or maintain language diversity as in Spain. Firms provide other examples. For instance, many firms need to choose between common and specialised 'languages' across divisions (Crémer et al. 2007) and evaluate the trade-offs of multinational teams with language frictions (Chen, Geluykens, and Choi 2006, Tenzer, Pudelko, and Harzing 2014). Finally, this trade-off is relevant in international trade, where a common language is necessary to find agreements but language costs can prevent efficient interactions (Melitz 2008, Lohmann 2011).

To my knowledge, this is the first paper to empirically quantify the role of language costs in communication quality and its outcomes. Some experimental literature has tested communication games with and without communication frictions (Lafky and Wilson 2020 and Blume, DeJong, Kim, and Sprinkle 2001, respectively). The works of 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. As in this paper, McManus (1985), Tainer (1988), and Guillouët et al. (2021) investigate the cost of using English and study its impact on wages and productivity. By contrast, Battiston et al. (2021) focus on the communication frictions arising from being unable 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 the changes in communication quality.

The rest of this paper is organised as follows: section 2 describes the Stack Overflow platform; section 3 provides intuitions over the main results with empirical facts; section 4 presents a theoretical framework and predictions; section 5 describes empirical proxies; section 6 derives empirical hypothesis; section 7 provides the empirical analysis and estimates for the effect on *old joiners*' answer quality; section 8 provides the empirical analysis and estimates for the difference in answer quality between *old joiners* and *new joiners*; section 9 discusses managerial implications; and section 10 presents the conclusion of this study.

[6] Note that this analysis does not consider other possible implications such as trade-offs on the number of contributions and the community size. See section 9 for a more extensive discussion.

2 Communication in Q&A platforms

Question-and-answer websites are online platforms that allow users to ask questions or answer them. Examples of such websites include *Stack Overflow*, *Yahoo! Answers*, and *Quora*.[7] These platforms are particularly useful for analysing communication as, compared to offline environments, provide larger samples, richer data, and less personal interactions, suggesting less confounding factors.

2.1 Stack Overflow

For the empirical investigation, this paper relies on data from Stack Overflow, a question-and-answer 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 source of information for all possible problems that programmers may encounter.[8] The key features of the platform are that it is crowd-based and free of charge. In other words, any internet user can register on the website for free and ask questions, provide answers, or both. Contributors are not remunerated.[9]

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, Stack Overflow receives more than one hundred million monthly visits.[10] In addition, Stack Overflow offers information seekers content that is easily accessible and searchable via browser 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 Languages used in Stack Overflow

As of June 2021, there are five different Stack Overflow websites, each in a different language, namely English, Russian, Japanese, Portuguese, and Spanish.[11] Apart from the language, their function is identical. Each website became public at a different time. 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 realised that a significant part of the programming community may be unable to access English content. After some discussion, they decided to open Stack Overflow in languages other than English.[12]

[7] *Yahoo! Answers* was shut down in April 2021.

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

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

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

[11] URLs of the sites are as follows. English: `https://stackoverflow.com/`, Russian: `https://ru.stackoverflow.com/`, Japanese: `https://ja.stackoverflow.com/`, Portuguese: `https://pt.stackoverflow.com/`, Spanish: `https://es.stackoverflow.com/`.

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

The platform designers chose those four additional languages because large communities of programmers speak them and, at the same time, 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.

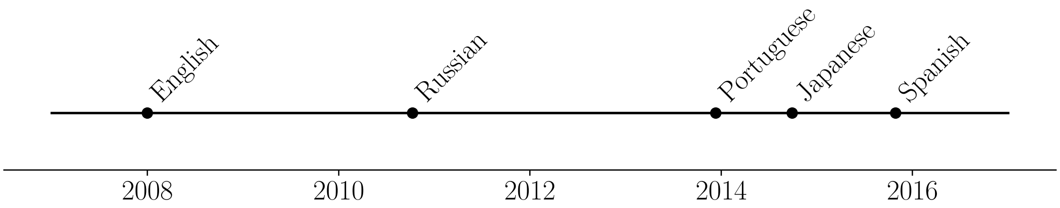


Figure 1: Timeline of the introduction of the Stack Overflow websites.

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 the English Stack Overflow in Russian. They created a website called HashCode, replicating Stack Overflow's features and purpose. Once the company behind Stack Overflow decided to open a version in Russian, it acquired HashCode: on 31 March 2015, all posts from HashCode were imported into the Russian version of Stack Overflow. Formally, the Russian version of Stack Overflow appeared in 2015. The data available include all the HashCode content, implying that the Russian data are available from 2010. Figure 1 reflects this fact and reports the 10 October 2010 as the date for the site launch.

3 Stylised Facts

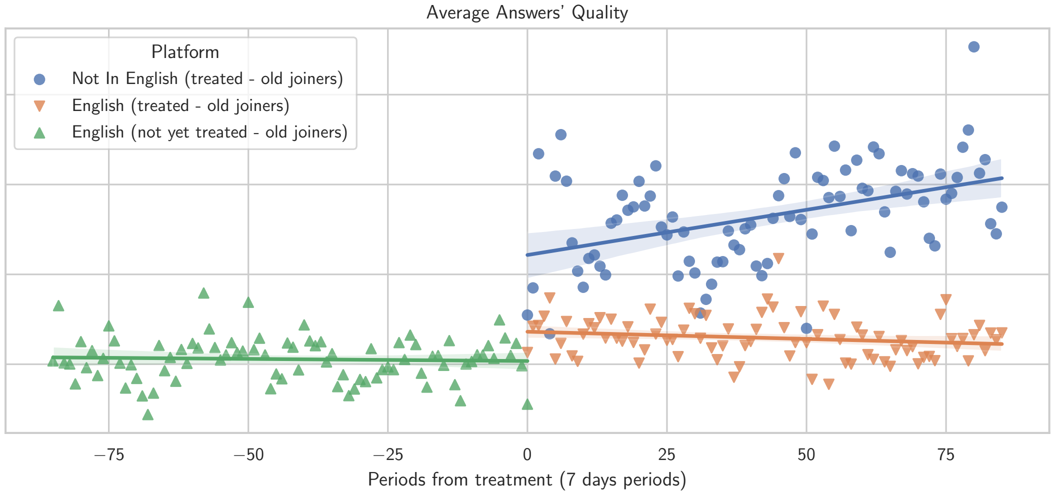
The introduction of new sites that use different languages mainly affects users who are native speakers of those languages. These users may have or have not contributed in English before their native language became available. I refer to the former as *old joiners*. In practice, these users participated on the English site before their native language became available and then participated in their native language once available. If the availability of the native language defines a treatment, *old joiners* are *not-yet-treated* when they can only use English, and *treated* afterwards. The researcher can observe their contributions before and after treatment. On the contrary, this is not possible with users who have not contributed to the English site before their native language was introduced. I refer to these users as *new joiners* on the condition that they had joined the platform before their native language became available.[14]

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

[14] Users joining when their native language was already available were not directly affected by the introduction of the new sites. I use the users' account registration date to identify when they joined.

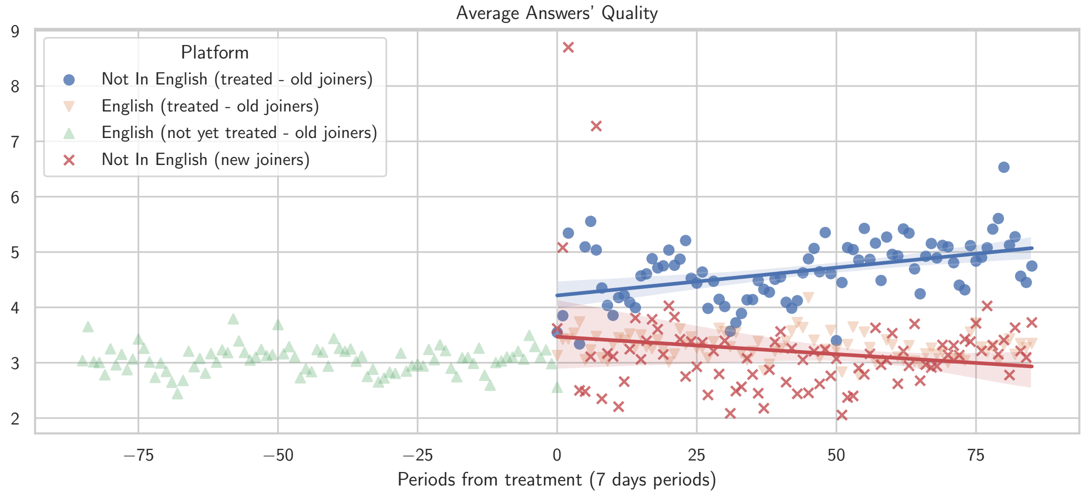
Intuitively, the platform, by allowing users to communicate in their native languages, decreases their cost of participation. On the one hand, this may allow *old joiners* to increase their contribution quality. On the other hand, it may allow *new joiners* to start contributing. The latter effect may be positive for the platform as it allows the community size to grow; however, it may have negative implications on content quality if *new joiners* are poor contributors.

This paper studies these possible effects on quality in detail. The raw data already provide suggestive evidence of the quality trade-off faced by the platform. Figure 2 plots the *old joiners*' average answer quality based on the language used before and after treatment. Quality is measured as the number of separate snippets of code appearing in the text of the answer. The figure shows that *old joiners*'s answers in the native language have a higher quality compared to those written in English, suggesting that the platform faces an increase in content quality. However, this beneficial effect may be compensated by lower-quality content from new participants. Figure 3 augments figure 2 by adding the *new joiners*'s average answer quality. It shows that, on average, *new joiners* contribute lower-quality answers than *old joiners*. A possible mechanism driving this effect is that language costs act as screening devices for knowledgeable users as the other users find it too costly to participate. By removing language barriers, the costs are reduced, and inexpert users inflow the platform with low-quality content. This paper provides some theoretical insights to explain these mechanisms, derives testable hypotheses, and investigates them empirically.



Notes. Each point represents the average number of pieces of code that *old joiners* have included in their answers in a 7-day period. The averages are based on the language used and whether the answer was posted before or after the introduction of the users' native language. Not all *old joiners* participated in each 7-day period.

Figure 2: Average answer quality before and after the introduction of the native language websites.



Notes. Each point represents the average number of pieces of code that *old joiners* and *new joiners* have included in their answers in a 7-day period. The averages are based on the language used and whether the answer was posted before or after the introduction of the users' native language. Not all *old joiners* and *new joiners* participated in each 7-day period.

Figure 3: Average answer quality of *old joiners* and *new joiners* before and after the introduction of the native language websites.

4 Theoretical Illustration

In this section, I present a simple framework to provide insights into how the cost of language and other factors affect communication quality between a questioner and an answerer on Stack Overflow. This section aims to guide the empirical analysis and the interpretation of the results in showing how answer quality depends on the question's quality, the incentives between sender and receiver, and the answerers' expertise and cost of language, where the latter is the critical variable of interest.

In the context of this paper, communication quality is a joint measure of the clarity and informativeness of the information transmitted and directly impacts the degree to which the receiver understands the information. Communication quality depends on both the question's and answer's quality, which are the outcomes of the transformation of inputs (e.g. effort, time) in the texts of the question and the answer. The framework does not model these processes. Instead, it sets question quality and answer quality as the choice variables, implicitly assuming that the users can manipulate those inputs to achieve a desired level of quality.[15]

The framework simplifies and adapts the model proposed by Calvó-Armengol, de Martí,

[15] This definition of quality is compatible with the one used by Dessein and Santos (2006) (section V.A) and Dessein, Galeotti, and Santos (2016), where communication quality is defined as the probability of a successful communication and is endogenous even though, in their setting, it is exogenous from the communicating parties. By contrast, Dewatripont and Tirole (2005) refer to this probability as the *communication technology*.

and Prat (2015) to pairwise communication and unilateral information transmission.[16] Simplicity comes at the cost of several limitations. In particular, in the model, pairwise communication is independent of other communicating pairs and the answerers' communication choices do not depend on other answerers. In addition, the framework focuses on modelling a single interaction and does not provide insights into how communication costs affect the quantity of answers provided.

These assumptions can be partly justified by the fact that the framework focuses on explaining quality rather than participation intensity and the realisation of quality is conditional on participation. Nevertheless, under these assumptions, the model abstracts from factors that may affect quality choices. These factors include the competition arising from other possible answerers and the potential non-random distribution of the questioner's characteristics and the question's topic. Section B.5 in the appendix discusses the issue in more detail, while the empirical analyses address these potentially confounding factors.

4.1 Model setup

Let Bob be a programmer who needs to understand how to implement some features in his software. He decides to ask about his problem to the Stack Overflow community as, otherwise, he cannot proceed with his project. Alice is a community member who sometimes answers questions on the platform. Alice defines her strategy of communication quality before observing the information that Bob needs.[17]

Let the information that solves Bob's problem be θ. Both Bob and Alice know its ex-ante distribution:

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

Here, s represents the precision of the prior: if s is high, Bob cannot be too wrong even if he picks a solution at random; if s is low, he may make major mistakes.[18]

[16] An alternative model with similar assumptions is the one presented by Dewatripont and Tirole (2005).

[17] In this context, I use the term *strategy* as used in game theory, where it refers to a function that maps possible values of contingent states (e.g. question quality, topic, language) to a level of answer quality. The assumption of committing to a strategy before observing the information implies that Alice forms her communication strategy while taking expectations over the information that Bob may need. This assumption is realistic if we believe that Alice adopts a long-term strategy that she maintains every time she answers a question. Note that the strategy may still depend on question-specific features and Alice's choice will be different for every answer she writes. The alternative assumption would be that Alice forms her strategy only after observing the information the questioner needs. The latter assumption substantially reduces the tractability of the model, and I do not have specific reasons to believe that this assumption is more accurate than the former.

[18] The precision of the prior can be interpreted as how diverse the possible solutions can be. For instance, since Stack Overflow hosts only questions about programming languages and software, s is a bit higher than what it would be in a platform that hosts any type of question as the range of possible solutions is more limited.

First, Bob writes the question by choosing its quality Φ_Q. He makes the quality choice by minimising the loss function

$$U_Q = \mathbb{E}\left[-\left((a^* - \theta)^2 + C_Q^2 \Phi_Q\right)\right],$$

where $a^* \in (-\infty, \infty)$ is the solution that he will pick in the last stage to solve his problem. Thus, the first argument represents his disutility from how different the solution he picks is from the correct solution. C_Q is his cost of communication.[19]

Second, Alice reads Bob's question: based on her communication strategy, she replies by choosing the answer's quality Φ_A. Alice cares about Bob choosing the right solution to a degree $\gamma \in [0, 1]$. γ determines how much she is inclined to help Bob, which may depend on platform-level incentives (e.g. virtual rewards for providing high-quality answers) or her personal preferences, like empathy, and altruism. For what follows, and for simplicity, I will refer to γ as the degree of incentives.[20] She then chooses the answer's quality to minimise the loss function:

$$U_A = \mathbb{E}\left[-\left(\gamma(a^* - \theta)^2 + C_A^2 \Phi_A\right)\right],$$

where C_A is her cost of communication.

I assume that the cost of communication is a function of the cost of using a given language (λ) and the expertise in the subject (k). Specifically:

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

for Bob and Alice, respectively. This functional form captures the fact that with low proficiency in the language used, users have a high cost of communication. Nevertheless, expertise and familiarity with the topic alleviate that cost.

Once the question and the answer are published, Bob receives the information m, which represents a noisy version of the correct solution

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

where the noise terms ε and η are stochastically independent from θ and each other and shrink in the question's and answer's quality. More precisely,

$$\varepsilon \sim \mathscr{N}\left(0, \frac{1}{\Phi_Q}\right); \quad \eta \sim \mathscr{N}\left(0, \frac{1}{\Phi_A}\right).$$

When the quality of the question is higher, the variance of the noise ε is lower and the message is more precise. Similarly, the noise coming from the answer shrinks when the answer's quality is higher.

Finally, Bob uses m to update his prior over the correct solution and picks the optimal solution a^*.

[19] The use of quadratic cost functions, as well as defining the loss function on quadratic distances, is standard in the organisational economics literature, e.g. Calvó-Armengol et al. 2015, Dessein and Santos 2006, Dessein et al. 2016.

[20] Note that, depending on which factors are relevant, Alice's γ may be question-specific or fixed across the questions that Alice answers.

4.2 Optimal choice of answer's quality

For a given question, what is Alice's optimal choice of her answer's quality?[21] Proceeding backwards, given the received message m, Bob selects the action a^* such that:

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

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

$$a^* = \mathbb{E}[\theta|m] = \beta m \quad \text{with} \quad \beta \equiv \frac{\Phi_Q \Phi_A}{\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s}. \tag{1}$$

In words, Bob weighs the message by its expected informativeness.

Anticipating Bob's action, Alice solves the following:

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

and her best response is then given by:

$$BR_A(\Phi_Q) = \frac{\Phi_Q(\sqrt{\gamma} k_A - s\lambda_A)}{\lambda_A(\Phi_Q + s)}. \tag{2}$$

4.3 Implications of the model

This section derives theoretical predictions and testable implications for both *old joiners* who contributed in English before their native language became available and *new joiners* who did not. This is achieved by assuming that all answerers participating in the platform follow the same decision process as Alice.

To formally define *old joiners* and *new joiners* in the context of the theoretical framework, note that Alice contributes an answer if

$$\sqrt{\gamma} k_A > s\lambda_A. \tag{3}$$

This condition says that Alice contributes an answer if she sufficiently internalises Bob's payoff (i.e. γ is high enough), if Bob is sufficiently confused ex-ante about the correct solution (i.e. the precision of the prior, s, is low enough), if her expertise is sufficiently high (i.e. k_A is high enough), and if her cost of using the language is sufficiently low (i.e. λ_A is low enough).

Assuming that Spanish is Alice's native language, it follows that Alice is part of the *old joiners* if the condition holds for both the English and Spanish sites. Conversely, she is part of the *new joiners* if the condition holds for the Spanish site but not for the English one.

[21] Details on the steps are provided in appendix B.

4.3.1 Implications for the *old joiners*: comparative statics on the cost of language

Equation 2 shows how answer quality depends on the cost of language. By taking the partial derivative with respect to the level of cost of language λ_A, we obtain:

$$\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A} = -\frac{\Phi_Q \sqrt{\gamma} k_A}{\lambda_A^2(\Phi_Q + s)} < 0, \tag{4}$$

which suggest that higher costs of language induce lower answer quality. The model provides additional predictions. The cross derivatives show that after a drop in the cost of language, the quality increase is larger for higher question quality and higher degree of incentives:

$$\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A \partial \Phi_Q} = -\frac{\sqrt{\gamma} k_A s}{\lambda_A^2(\Phi_Q + s)^2} < 0, \tag{5}$$

$$\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A \partial \gamma} = -\frac{\Phi_Q k_A}{2\lambda_A^2(\Phi_Q + s)\sqrt{\gamma}} < 0. \tag{6}$$

These comparative statics lead to the following proposition.

Proposition 1 *When users can answer in their native language,*

1. *their answer quality increases,*
2. *the effect increases when the quality of the question is higher, and*
3. *the effect increases when the answerer is more incentivised*

Proposition 1 relies on the assumption that users active in non-English languages are native speakers of those languages and more proficient in those languages than in English.[22] Point 1 directly follows from equation 4, which establishes a negative relationship between the answer's quality and the author's cost of using the language. The use of the native language rather than English induces a decrease in the cost of language and allow users to provide higher quality answers.[23] Point 2 follows from equations 5, which show that the effect of the cost of language on the answer's quality increases with the question's quality. This effect is due to the fact that the model features complementarities between the questioner's and the answerer's quality choices. This is not imposed as an assumption but emerges endogenously. Indeed, question quality and answer quality are complements in determining the informativeness of the communication, as shown in equation 1. Such complementarity is then anticipated and internalised by the answerer's

[22] Supportive evidence includes the fact that 99.8% of users who contribute in both English and other languages contribute in only one non-English language.

[23] Note that a natural extension of point 1 is that the effect is larger for users with a lower proficiency in English, as they experience a larger drop in the cost of language when they use their native language. Section G.1 in the appendix studies heterogeneity of the effect across different possible proxies for the cost of using English.

decision. Finally, point 3 follows from equation 6, which shows that the degree to which the answerer cares about the questioner's utility augments the effect of a change in the cost of language. This effect emerges due to complementarity in the answerer's quality choice between the incentives and the cost of language.

4.3.2 Implications for the *new joiners*

Condition 3 states that, with the incentives (γ) and the prior's precision (s) fixed, Alice's participation is conditional to a level of expertise high enough relative to the cost of language. If she does not have a sufficiently high level of expertise, she does not participate in English but may be able to participate in her native language since lower language costs loosen the condition to participate.

Formally, consider a question on the topic ω and let j index possible answerers for that question. Assume that the answerers differ only on their expertise on ω, k_{A_j}. Define λ_A as their cost of using English and λ'_A as their cost of using the native language, with $\lambda'_A < \lambda_A$. Let $f(k_A)$ be the probability distribution of expertise on ω across answerers and assume it is the same on both the English and native language sites. We can then say that a user is part of the *old joiners* (OJ) or the *new joiners* (NJ) based on the level of expertise:

$$
\begin{aligned}
j \in OJ \quad & \text{if} \quad k_{A_j} > \frac{s\lambda_A}{\sqrt{\gamma}}, \\
j \in NJ \quad & \text{if} \quad \frac{s\lambda'_A}{\sqrt{\gamma}} < k_{A_j} < \frac{s\lambda_A}{\sqrt{\gamma}}.
\end{aligned}
$$

It then follows that, for any j, j' if $j \in OJ$ and $j' \in NJ$:

$$
\mathbb{E}\left[k_{A_j}|\omega\right] \geq \mathbb{E}\left[k_{A_{j'}}|\omega\right]. \tag{7}
$$

In other words, for a given question's topic, answerers who are *old joiners* have higher expertise than answerers who are *new joiners*. Note that this result generalises to the case in which answerers are heterogeneous on their cost of using English.[24]

Note that, from the best response function reported in equation 2, the choice of answer's quality is linear on the answerer's expertise. From equation 7 it follows that, for any j, j' if $j \in OJ$ and $j' \in NJ$:

$$
\frac{\Phi_Q(\sqrt{\gamma}\mathbb{E}\left[k_{A_j}|\omega\right] - s\lambda'_A)}{\lambda'_A(\Phi_Q + s)} \geq \frac{\Phi_Q(\sqrt{\gamma}\mathbb{E}\left[k_{A_{j'}}|\omega\right] - s\lambda'_A)}{\lambda'_A(\Phi_Q + s)}.
$$

This result implies the following proposition.

Proposition 2 *On average, for a given topic, answers from old joiners have higher quality than answers from new joiners.*

[24] See section B.4 in the appendix for more details.

5 Empirical proxies

To test empirically the theoretical predictions, it is necessary to identify observable proxies for the model's variables.

5.1 Measuring the answers' quality

To proxy for the answers' quality, this paper uses the number of separate snippets of code in the text of the answer; I refer to this variable as *numCodes*. More precisely, each answer is an *html* script. To include code snippets in their answer, users add *code* sections (i.e. <*code*>...< /*code*>) to make the code appear in a separate box with a different colour background. The box mimics a programming/statistical software console and makes the code more readable. The proxy of quality is then defined as the number of *code* sections in the answer. The intuition behind this measure is that a typical answer about programming includes some textual explanations and some code snippets to illustrate the solution. The presence of multiple snippets may indicate that the author is either providing several pieces of information (higher informativeness) or explaining one piece of information more clearly with a step-by-step procedure (higher clarity). In both cases, more snippets suggest higher answer quality.[25]

Compared to alternative proxies, this measure has two main advantages. First, it is not directly based on the language itself, which would create a systematic difference across sites by construction. This issue would affect measures based on text or other text characteristics. Second, the number of snippets in an answer is a direct choice of the answerer, as the model assumes it to be.

An alternative way to measure quality relies on community feedback (e.g. the number of *likes* in the question). While measures of this type capture communication quality as a whole rather than the answers' quality, they provide information on the value of the answer to the community. I replicate the analysis with a measure of quality based on whether the questioner has *accepted* the answer. Indeed, questioners can *accept* the answer that solved their problem, if any. I refer to this dummy variable as *Is Best Answer*.[26]

5.2 Measuring the questions' quality

To measure the questions' quality, I adopt the same strategy as for the answers and use the number of separate snippets of code appearing in the text of the question as a proxy. The intuition is that a question with code is clearer as it provides context for the problem and a direct way for the answerer to reproduce the error. More pieces of code provide more examples, different attempts, and so on, all of which can help the answerer

[25] Figure 6 in the appendix shows an example of an answer with two code snippets.

[26] A measure based on *likes* is less preferable as *likes* tend to be more sensitive to timings, order of publication of the answers in a question thread, and community size. Nevertheless, *accepted* answers and answers with more pieces of code have significantly more *likes*, as shown in section D.2 in the appendix.

clearly understand the question.[27]

5.3 Measuring incentives

To capture the degree of incentives, I exploit the so-called *bounty* system implemented by Stack Overlow. The platform allows the questioner to auction a certain amount of his reputation points (i.e. virtual points that the questioner has accumulated through on-site activity) on a question. The questioner commits to allocating these points to the user who provides a satisfactory answer. Once the questioner auctions the points, these are removed from his account, although the questioner is not obliged to designate a recipient. The points may remain unallocated if the questioner considers all received answers unworthy.[28] This feature of Stack Overflow represents a form of virtual payment with which the questioner can provide extra incentives. A larger *bounty* amount implies a more significant stake in the question and a stronger incentive for the answerer.

More precisely, consider an answer i to a question $q(i)$. The proxy for incentives is then the *bounty* amount auctioned on question $q(i)$ and not yet assigned when answer i is published.

6 Empirical predictions

This section presents the empirical counterpart of the theoretical predictions, representing the hypotheses that the empirical analysis addresses. The two sets of empirical predictions correspond to propositions 1 and 2, respectively.

6.1 Treatment effect on the quality of *old joiners*' answers

Proposition 1 states that, everything else equal, the answer quality of *old joiners* increases once they are able to use their native language rather than English only, and that the effect is higher with higher question quality and more incentives for the answerer.

Given the main empirical measures described in section 5, the derived empirical hypotheses are as follows.

Hypothesis 1 *Old joiners include more snippets of code in their answers when they use their native language compared to when they use English.*

Hypothesis 2 *The increase in answer quality due to the use of the native language instead of English is larger when the question's quality is higher.*

[27] The platform advises the inclusion of code for the development of good questions: https://stackoverflow.com/help/how-to-ask.

[28] There are some cases in which the points are allocated even if the questioner does not select a recipient. In addition, other users who did not write the question can also set up *bounties*. More details are available here: https://stackoverflow.com/help/privileges/set-bounties.

Hypothesis 3 *The increase in answer quality due to the use of the native language instead of English is larger when the active bounties auctioned on the question have a higher value.*

6.2 Language costs as screening devices

Proposition 2 shows that, everything else equal, *new joiners* are on average less expert than *old joiners* and consequently write lower-quality answers.

Since the researcher does not observe the answerers' expertise, it is not possible to identify *new joiners* as those who did not participate on the English site strictly because of insufficient expertise. Empirically, I define *new joiners* as users who: 1) registered on the English site before their native language became available; 2) did not contribute in English; and 3) contributed in their native language once it became available. This empirical definition rules out the possibility that these users did not contribute to the English site because they were unaware of the platform, although other explanations remain possible. It follows that the empirical analysis studies the differences in answer quality which are at least partially attributable to differences in expertise.[29]

The testable hypothesis is stated as follows.

Hypothesis 4 *On average, for a question written in their native language, users already active in English before treatment (old joiners) write answers with more pieces of code compared to users who registered before treatment but are not active in English (new joiners).*

7 The effect on the quality of *old joiners*' answers

This section empirically tests hypotheses 1, 2, and 3. It studies whether the introduction of additional languages on the platform has affected the quality of *old joiners*'s answers.

7.1 Data

The data include all answers of users participating on both the English and a non-English site (the *old joiners* or *treatment group*) and all the answers of a random sample of users

[29] Note that in general, other possible explanations do not imply the same prediction, i.e. that *new joiners* contribute lower-quality content compared to *old joiners*. For instance, it is possible that, as discussed in section B.5 in the appendix, topics are systematically different across websites, or users are significantly more empathetic when participating in communities that use their native language. In the former case, *new joiners* may not have participated in English due to insufficient expertise even though they may have the same expertise as *old joiners* when using their native language. Similarly, in the latter case, *new joiners* may not have participated in English because they did not care enough about the questioners' utility in the English community, yet be as empathetic as the *old joiners* in their native language community. Either way, these explanations imply that *new joiners* do not write lower-quality content compared to *old joiners*. This contradicts the theoretical prediction, and it remains an empirical question whether that is true.

who contribute only to the English site (*never-treated group*).[30]

The treatment occurs if the user has the possibility to write in the native language. If a user is part of the *old joiners*, she is *treated* if her native language is available (i.e. if the site using her native language has already launched) and *not-yet-treated* if her native language is not yet available. For the sake of simplicity, I say that an answer is *treated* if its author is *treated*, *not-yet-treated* if its author is *not-yet-treated*, and *never treated* otherwise. Note that all answers written after the launch of the native language site are *treated*, regardless of whether they are in the native language or English.[31]

Table 1 reports the number of observations by the language and treatment status of the author. The data are right-censored at the end of August 2017 as that was the latest available data at the time of the data retrieval.[32]

The observed variables are at the answer level and represent either answer characteristics (quality, number of *likes*, whether the answer has been *accepted*) or characteristics of the question that the answer is answering. The latter characteristics include quality, the *bounty* amount, the number of answers and views received, and the characteristics of the questioner when available, including whether the questioner shares the language with the answerer and has uploaded a profile picture. Table 2 reports summary statistics of the numeric variables by treatment status of the answer's author and the language in which the answer is written.

[30] More precisely, if a user published in both English and, say, Spanish, the user belongs to the *treatment group* if they published at least 1) one answer in English before the Spanish site became available and 2) one answer in Spanish.

[31] The choice of defining the treatment in this way responds to the need to remain close to standard environments, where all observations of *treatment group* users are considered *treated* past the treatment date. An alternative strategy would define an answer as *treated* only when it is written in the native language and exclude English post-treatment answers from the sample. Regression results based on this alternative sample are reported in section E.1 in the appendix.

[32] The data are not left-censored as they have been available since the sites' launch. Nevertheless, some sites had a *proposal* stage for which data are unavailable. For more details, see section A.1 in the appendix.

Sample	Num Answers	Num Authors	Num. Answers per Author mean	 std	 min	 median	 max
English (never treated)	37049	2661	13.92	63.13	1.00	2.00	2175.00
English (not yet treated)	128982	2678	48.16	184.82	1.00	5.00	2848.00
English (treated)	100537	2087	48.17	224.25	1.00	7.00	4894.00
Not In English (treated)	57282	2678	21.39	134.73	1.00	2.00	3759.00
Full sample	323850	5339	60.66	266.49	1.00	7.00	6596.00

Notes. The definitions of the subsamples of answers are as follows. *English (not yet treated)* is the set of answers that users in the *treatment group* write before treatment; *English (treated)* is the set of answers that users in the *treatment group* write after treatment in English; *Not in English (treated)* is the set of answers that users in the *treatment group* write after treatment not in English; and *English (never treated)* is the set of answers written by users in the *never-treated group*. By construction, the same users compose the *English (not yet treated)* and *Not in English (treated)* subsamples. Since some users stop participating in English after treatment, the set of users in the *English (treated)* group is smaller but still contained in the *treatment group* sample.

Table 1: Number of observations in the baseline sample and subsamples based on the language and treatment status of the author.

Sample	Variable	mean	std	min	median	max
English (never treated)	Answer's numCodes	2.01	2.69	0.0	1.0	67.0
	Answer' Score	3.62	20.55	-13.0	1.0	1264.0
	Is Best Answer	0.31	0.46	0.0	0.0	1.0
	Question's numCodes	2.13	2.70	0.0	1.0	41.0
	Bounty Amount	1.88	21.63	0.0	0.0	1200.0
	Question's Num. Answers	3.30	4.64	1.0	2.0	407.0
	Question's Num. Views	15062.10	93963.64	9.0	1090.0	4642034.0
	Same Native language	-	-	-	-	-
	... and Manual Profile Pict.	-	-	-	-	-
English (not yet treated)	Answer's numCodes	2.74	4.02	0.0	2.0	284.0
	Answer' Score	3.45	18.60	-16.0	1.0	1902.0
	Is Best Answer	0.36	0.48	0.0	0.0	1.0
	Question's numCodes	2.23	2.77	0.0	1.0	111.0
	Bounty Amount	1.16	14.70	0.0	0.0	700.0
	Question's Num. Answers	3.28	5.95	0.0	2.0	518.0
	Question's Num. Views	10618.68	72795.79	7.0	1003.0	7465142.0
	Same Native language	0.02	0.12	0.0	0.0	1.0
	... and Manual Profile Pict.	0.01	0.07	0.0	0.0	1.0
English (treated)	Answer's numCodes	3.55	4.65	0.0	2.0	153.0
	Answer' Score	2.33	11.76	-16.0	1.0	1068.0
	Is Best Answer	0.41	0.49	0.0	0.0	1.0
	Question's numCodes	2.69	3.07	0.0	2.0	71.0
	Bounty Amount	1.14	14.33	0.0	0.0	800.0
	Question's Num. Answers	2.52	2.86	0.0	2.0	116.0
	Question's Num. Views	9058.16	87796.96	8.0	603.0	8671208.0
	Same Native language	0.02	0.13	0.0	0.0	1.0
	... and Manual Profile Pict.	0.01	0.09	0.0	0.0	1.0
Not In English (treated)	Answer's numCodes	4.64	6.15	0.0	3.0	186.0
	Answer' Score	2.91	4.57	-5.0	2.0	256.0
	Is Best Answer	0.47	0.50	0.0	0.0	1.0
	Question's numCodes	2.56	3.05	0.0	2.0	87.0
	Bounty Amount	0.95	12.64	0.0	0.0	800.0
	Question's Num. Answers	2.01	1.30	1.0	2.0	27.0
	Question's Num. Views	953.91	3382.53	5.0	177.0	114994.0
	Same Native language	1.00	0.00	1.0	1.0	1.0
	... and Manual Profile Pict.	0.52	0.50	0.0	1.0	1.0

Notes. *Answer's numCodes* is the number of pieces of code appearing in the answer; *answer' score* is the number of 'likes' net of 'dislikes' that the answer received; *is best answer* is a dummy equal to 1 if the questioner 'accepted' the answer as the one solving the problem; *question's numCodes* is the number of pieces of code appearing in the question; *bounty amount* is the number of points auctioned on the question; *question's num. answers* is the number of answers that the question has received, including the one of the observation; *question's num. views* is the number of users that have viewed the question; *same native language* is a dummy equal to 1 if the author of the answer and the questioner speak the same language; and *manual profile picture* is a dummy equal to 1 if the questioner both has a personalised profile picture and speaks the same language as the answerer. The *same native language* and *manual profile picture* variables are not observed for the *never treated* group.

Table 2: Summary statistics at the answer level for the baseline sample by the site's language and the language and treatment status of the author.

7.2 Empirical strategy

The empirical strategy exploits the sequential introduction of websites in Russian, Portuguese, Spanish, and Japanese in addition to the English site. The ability to track users' contributions across sites allows for measuring changes in quality choices for each user before and after their native language site became available, and compare them with the *never-treated* users. In this setting, I can identify the effect of the availability of the native language on answers' quality (the treatment effect on the treated) using a difference-in-difference design with staggered treatment. Section C.1 in the appendix provides details on the identifying assumptions, describes possible violations, and discusses how I address them, including adding control variables in the regressions and through robustness checks.

7.2.1 Estimation method

The literature has proposed several methods to estimate the average treatment effect on the treated (ATT) in staggered difference-in-difference designs. In this paper, I provide estimates under two approaches. The first is the so-called *two-way fixed effect* method (TWFE), which I present as a benchmark since it is the most commonly used method, although I claim it provides biased estimates in this environment. The second is the method proposed by Borusyak et al. (2022).[33]

The *two-way fixed effect* method is an ordinary least square (OLS) model with individual and time fixed effects. More precisely, the estimating equation is:

$$numCodes_i = \alpha_{j(i)} + \delta_{t(i)} + \beta D_{L(j(i),t(i))} + \boldsymbol{W}_i'\boldsymbol{\rho} + \varepsilon_i, \tag{8}$$

where i indexes an answer written by the user $j(i)$ at the time $t(i)$. D is a dummy variable taking a value equal to 1 if the answer i was written when the site in the language L, j's native language, was available; $\boldsymbol{W}$ is a vector of control variables. $\alpha_{j(i)}$ is the author fixed effect, while $\delta_{t(i)}$ is the time fixed effect. The coefficient of interest β then captures the ATT. However, the identification relies on the assumption that all treatment effects for all users and periods are the same.[34] This assumption does not hold as the treatment effect should depend on the size of the change in the cost of language, which is different across users if the users are heterogeneous in the cost of using English.

To allow for less restrictive assumptions, the preferred estimation strategy is the method proposed by Borusyak et al. (2022) (BJS hereafter).[35] The method builds on

[33] I discuss the estimation method using *numCodes* as the dependent variable. However, I execute the same estimation also using *Is Best Answer* as dependent variable.

[34] As discussed in several papers (Callaway and Sant'Anna 2020, de Chaisemartin and D'Haultfœuille 2020, Sun and Abraham 2020, Goodman-Bacon 2021, Borusyak et al. 2022), the TWFE estimation procedure estimates the treatment effect as a weighted average of all possible treatment effects for each user × period cell. The weights sum to 1 but may be negative. If the treatment effects are not homogeneous, asymmetric weighting may cause undesired outcomes.

[35] The literature has proposed other solutions. For example, 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

the intuition that, under the parallel trend assumption, a model trained with the *not-treated* data (i.e. *never-treated* and *not-yet-treated* answers) can predict the potential outcome, i.e. the counterfactual choice of treated users if they had not been treated. More precisely, the BJS method comprises three steps.

Step 1. Estimate with OLS a linear model on the *never-treated* and *not-yet-treated* answers only:

$$numCodes_i = \alpha_{j(i)} + \delta_{t(i)} + \boldsymbol{W}_i'\boldsymbol{\rho} + \varepsilon_i \quad \text{if } j(i) \text{ not treated at time } t(i),$$

with $\alpha_{j(i)}$ and $\delta_{t(i)}$ being the individual and time fixed effects and $\boldsymbol{W}_i$ being additional control variables. This step estimates parameters for individual and time fixed effects for the whole set of individuals and time periods.[36]

Step 2. With the model estimated in step 1, predict the counterfactual quality choice of treated users had they not been treated. Then, compute the treatment effect for each treated user's answer as the difference between the observed quality and the predicted counterfactual quality:

$$\widehat{numCodes}_i = \hat{\alpha}_{j(i)} + \hat{\delta}_{t(i)} + \boldsymbol{W}_i'\hat{\boldsymbol{\rho}} \quad \text{if } j(i) \text{ treated at time } t(i),$$
$$\hat{\tau}_i = numCodes_i - \widehat{numCodes}_i \quad \text{if } j(i) \text{ treated at time } t(i).$$

Step 3. Compute the average treatment effect on the treated (ATT) as the average of all answers' treatment effects. This step can take different forms as the researcher has the flexibility to select the best way to average across treatment effects. The baseline estimates use a simple average with homogeneous weights:

$$\hat{\tau} = \frac{1}{N_{post}} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i, \tag{9}$$

where N_{post} is the number of answers published by the treated users.[37]

7.3 The effect of the cost of language on the answers' quality

Table 3 reports the estimated treatment effect (i.e. *after*) corresponding to $\hat{\beta}$ in the TWFE specification (equation 8) and to $\hat{\tau}$ in the BJS specification (equation 9). Different columns include or exclude certain control variables. *QQuality* identifies the proxy

context of this paper, these 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.

[36] This estimation procedure does not support having time periods with treated observations but no untreated observations in the data. This is a situation that could occur at the end of the (unbalanced) panel.

[37] Imposing homogeneous weights on the weighted average is not optimal if the objective is to estimate an average effect at the user level. Indeed, this overweighs users that contribute more content. Section E.4 in the appendix provides alternative estimates of the ATT by first obtaining the average effect for each user and then taking the average effect across users, as suggested by Baker, Larcker, and Wang (2022).

for question quality as described in section 5.2; *Competition* identifies the variables capturing the number of answers in the question thread and the number of views of the question; and *Empathy* identifies the variables capturing whether the questioner is a native speaker of the same language of the answerer and whether the questioner both displays a personalised profile picture and share the same native language.[38]

The estimates show that when users can write answers in their native language, their answers have significantly higher quality on average, and questioners are more likely to solve their problems. The first set of estimates refers to specifications with *numCodes* as the dependent variable and quantifies the effect shown in figure 2. The preferred specification shows that, on average, answers have 0.586 additional pieces of code. Since the pre-treatment average for users in the *treatment group* is 2.74, as described in table 2 for the sample '*English (not yet treated)*', the preferred specification in column 8 reports a 21.4% increase in quality. In other words, answers written post-treatment, in either English or users' native languages, have 21.4% higher quality than English pre-treatment answers. Section E.1 in the appendix shows that the effect increases to 56.5% when focusing on non-English answers only, while section G.2 in the appendix shows that, in the preferred specification, the effect is not significant when including only English answers. These results suggest that non-English answers drive the overall effect, confirming the theory and hypothesis 1. The second set of estimates refers to specifications where the dependent variable is a dummy equal to 1 if the answer has been *accepted*, and 0 otherwise (*Is Best Answer*). The preferred specification in column 8 shows that the answers' probability of being *accepted* increases by 6.2 percentage points. This effect corresponds to a 17.2% increase from the pre-treatment average (36%) and suggests that a 21% increase in the answers' informativeness leads to a 17% increase in their effectiveness.

[38] Section C.1 in the appendix provides more details on these variables and the rationale for their inclusion.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
D.V.	TWFE	TWFE 1	TWFE 2	TWFE 3	BJS	BJS 1	BJS 2	BJS 3
numCodes	0.378^{*}	0.375^{*}	0.376^{*}	0.217^{*}	0.574^{***}	0.591^{***}	0.596^{***}	0.586^{***}
	(0.0929)	(0.0968)	(0.0970)	(0.0601)	(0.0445)	(0.0426)	(0.0400)	(0.0629)
Is Best Answer	0.0225^{**}	0.0223^{**}	0.0213^{**}	0.00894	0.105^{***}	0.105^{***}	0.0915^{***}	0.0624^{***}
	(0.00365)	(0.00371)	(0.00349)	(0.00391)	(0.00594)	(0.00593)	(0.00512)	(0.00760)
Observations	323850	322992	322992	285943	323850	322919	322919	204541
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QQuality	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

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

Notes. Treatment effect estimates where the dependent variable is the number of pieces of code (*numCodes*) or a dummy equal to 1 if the answer is *accepted* (*Is Best Answer*). The estimates correspond to the average treatment effect on the treated and correspond to the parameter $\hat{\beta}$ or $\hat{\tau}$ when the specification adopted is the TWFE or the BJS, respectively. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 3: Effect of the cost of language on the answers' quality

7.4 Complementarities of the effect with the questioners' choices and incentives

The theoretical predictions 5 and 6 suggest that the positive impact of a decrease in the cost of language is complementary to the questions' quality and the degree of incentives. To test these predictions and the corresponding hypotheses 2 and 3, I run the same estimation strategy used for the baseline estimate but allow the coefficient of interest to vary across the levels of those variables.

Let c index the categories for each of the levels. The specification for the TWFE estimation method is the following:

$$numCodes_i = \alpha_{j(i)} + \alpha_{t(i)} + \sum_c \beta_c D_{L(j(i),t(i))} \mathbf{1}_{c(i)} + \boldsymbol{W}_i' \boldsymbol{\gamma} + \varepsilon_i, \tag{10}$$

where $\mathbf{1}_{c(i)}$ is an indicator function taking a value equal to 1 if the answer i belongs to the level category c. Regarding the BJS estimation method, the only difference is in the last step. After estimating potential outcomes and treatment effects for each treated observation, the ATT results from averaging the treatment effects within each category level rather than across the all treated sample. Formally, the estimated treatment effect for category c is:

$$\hat{\tau}_c = \frac{1}{N_c} \sum_{i|j(i) \text{ treated at time } t(i)} \hat{\tau}_i \mathbf{1}_{c(i)}, \tag{11}$$

where N_c is the number of treated answers that belong to category c.

7.4.1 The treatment effect increases in the quality of the question

To test hypothesis 2, I separately estimate the treatment effect for different levels of question quality. As a proxy for the questions' quality, I use the number of separated snippets of code that the questioner included in the question (the variable *QQuality*). I bin this variable into four levels based on the quartiles of its distribution. This leads the *Low* category to include questions with either 0 or 1 piece of code, the *MediumLow* category questions with 2 pieces of code, the *MediumHigh* category questions with 3 pieces of code, and the *High* category questions with more than 3 and up to 111 pieces of code.

Table 4 reports the estimates. Columns 1–3 contain the $\{\hat{\beta}_c\}_{\forall c}$, while columns 4–6 contain the $\{\hat{\tau}_c\}_{\forall c}$, where c indexes the categories for *QQuality*. Columns 2 and 4 report the results for the specification that includes all the controls. These columns suggest that low-quality questions receive significantly lower-quality answers, while the answers' quality is quite homogeneous across other levels of question quality. Nevertheless, removing the question's quality as a control (columns 4 and 6) provides substantially different estimates and shows that the treatment effect increases with question quality. An intrinsic difference between the questions with zero snippets of code and those with a positive number of pieces of code may drive this change in estimates, as shown in the appendix (section E.5).[39] As a consequence, column 6 of table 4 reports the preferred specification and shows that the treatment effect increases with question quality.

7.4.2 The treatment effect increases in the amount of incentives

To test hypothesis 3, I separately estimate the treatment effect for different levels of incentives. To measure the degree of incentives between the two parties, I use the values of the *bounties* at stake on the question that the answer addresses, as described in section 5.3. I discretise the *bounty* amount into four categories. The *Low* category is just composed 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. In practice, the *MediumLow* category includes questions with 50 points at stake, the *MediumHigh* questions with 100 points at stake, and the *High* category questions with 150 to 1200 points at stake.

The results are reported in table 5 and show that, on average, the treatment effect is higher when authors are more incentivised, as suggested by the theoretical framework.

[39] The mechanism is that questions without any code may not require code in the answer. This fact is discussed in relation to the answers in section D.2 in the appendix. It creates a discontinuity on the number of pieces of code in the answers between questions that have or not have code, which is not addressed by the linearity of the control variable. Section E.5 in the appendix provides additional supporting evidence for this mechanism and the fact that the treatment effect increases with the question's quality. First, it reports the estimates of a model specification that includes question quality as a control variable but excludes questions with no code snippets from the sample. It shows that, in this case, estimates of the treatment effect increase with higher levels of question quality. Second, it reports the estimates obtained including question quality as a control variable and using a dummy equal to 1 if the questioner *accepted* the answer as the dependent variable. In this case, the treatment effect increases with the question's quality.

	(1) TWFE	(2) TWFE 2	(3) TWFE 3	(4) BJS	(5) BJS 2	(6) BJS 3
Low × after	0.121	-0.0394	-0.307*	0.288***	0.317***	0.0697
	(0.118)	(0.0749)	(0.0705)	(0.0626)	(0.0781)	(0.153)
MediumLow × after	0.565**	0.418**	0.343**	0.775***	0.809***	0.736***
	(0.0851)	(0.0562)	(0.0408)	(0.0866)	(0.101)	(0.0841)
MediumHigh × after	0.566**	0.415**	0.478***	0.792***	0.827***	0.881***
	(0.0882)	(0.0485)	(0.0291)	(0.0692)	(0.0910)	(0.124)
High × after	0.596***	0.423***	0.984***	0.896***	0.824***	1.328***
	(0.0546)	(0.0301)	(0.0330)	(0.0411)	(0.0498)	(0.242)
Observations	322992	285943	285943	322919	204541	204545
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls						
QQuality	Yes	Yes	No	Yes	Yes	No
Competition	Yes	Yes	Yes	Yes	Yes	Yes
Empathy	No	Yes	Yes	No	Yes	Yes

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

Notes. The *Low* category includes questions with either 0 or 1 piece of code, the *MediumLow* category questions with 2 pieces of code, the *MediumHigh* category questions with 3 pieces of code, and the *High* category questions with more than 3 and up to 111 pieces of code. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 4: Treatment effect estimates by level of question quality.

The estimates show that, in the absence of incentives via *bounties*, the answers' quality increases by about 21%. Note that *bounties* are not very frequent, which explains why this effect is substantially comparable to the baseline effect reported in section 7.3. The treatment effect increases to 37% when questions have 50 virtual points at stake, to 86% with 100 points at stake, and to 118% with 150 or more points at stake.

8 Language costs as screening devices

This section aims to test the empirical hypothesis 4, which states that *new joiners* contribute lower-quality answers compared to *new joiners*.

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.360^{*}	0.202^{*}	0.578^{***}	0.574^{***}
	(0.0960)	(0.0586)	(0.0400)	(0.0631)
MediumLow × after	1.220^{*}	1.043^{*}	1.511^{***}	1.007^{***}
	(0.273)	(0.255)	(0.198)	(0.191)
MediumHigh × after	2.298^{*}	2.133	2.728^{***}	2.366^{***}
	(0.802)	(0.857)	(0.515)	(0.436)
High × after	3.034^{***}	2.876^{**}	3.431^{***}	3.234^{***}
	(0.266)	(0.308)	(0.463)	(0.394)
Observations	322992	285943	322919	204541
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

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

Notes. The *Low* category includes questions with zero points at stake, the *MediumLow* questions with 50 points at stake, the *MediumHigh* questions with 100 points at stake, and the *High* category questions with 150 to 1200 points at stake. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 5: Treatment effect estimates by level of incentive.

8.1 Data

The data include all answers that satisfy two conditions: 1) they are published in non-English languages, and 2) they answer questions that received at least two answers at the time of the data retrieval. This leads to a final sample of 212,439 answers.[40]

A key categorical variable splits the sample based on the answerers' participation on the English site. While some users never registered on the English site, others registered before their native language became available, and others registered after it became available. Moreover, the users who registered may or may not have contributed in English. Table 6 summarises these possible combinations and reports the sample size for each category. The main groups of interest are those identifying *new joiners* and *old joiners*. The former are answerers who registered on the English site before their native language became available but did not contribute any answer in English before or after the native language site became available. The latter are answerers who registered and contributed answers on the English site before their native language became available.

[40] The exclusion of questions with only one answer implies the exclusion of about 66% of non-English questions and a 44% reduction of the sample of answers, as shown in table 11 in the appendix. The same table shows that the distribution of the number of answers per question is left-skewed. Nevertheless, a significant sample of about 90,000 answers addressed questions that received three or more answers.

		Active on the English site:			
		Only BEFORE	Only AFTER	Both BEFORE and AFTER	Not active
Registered on the English site:	BEFORE	2,589$(**)$	9,517	32,860$(**)$	4,373$(*)$
	AFTER		78,471		50,046
	Not registered				34,583
Total		212439			

Notes. The groups are based on participation on the English site. BEFORE and AFTER refer to the treatment date. Answers of authors considered *old joiners* are marked with the symbol $(**)$. Answers of authors considered *new joiners* marked with the symbol $(*)$.

Table 6: Number of answers in the sample by author group.

8.2 Empirical strategy

By exploiting the fact that each question can receive multiple answers, the empirical approach compares the answers' quality between *old joiners* and *new joiners* within each question's thread. The empirical model corresponds to an OLS regression with the user-group (*old joiners* versus *new joiners*) fixed effects, question fixed effects, order-of-publication fixed effects, and a control variable for seniority on the platform. This design allows us to make the quality comparison conditioning on a given topic and the characteristics of the questioner. In addition, the order-of-publication fixed effects control for the fact that the answers arriving second or later in the thread may systematically have less content as they just complement the information already available in the thread. Finally, the inclusion of the amount of time that the answerer has been registered on the platform as a control variable allows us to account for different degrees of experience. Indeed, if *new joiners* generally registered later, this may induce lower-quality answers as they are less experienced with the platform.

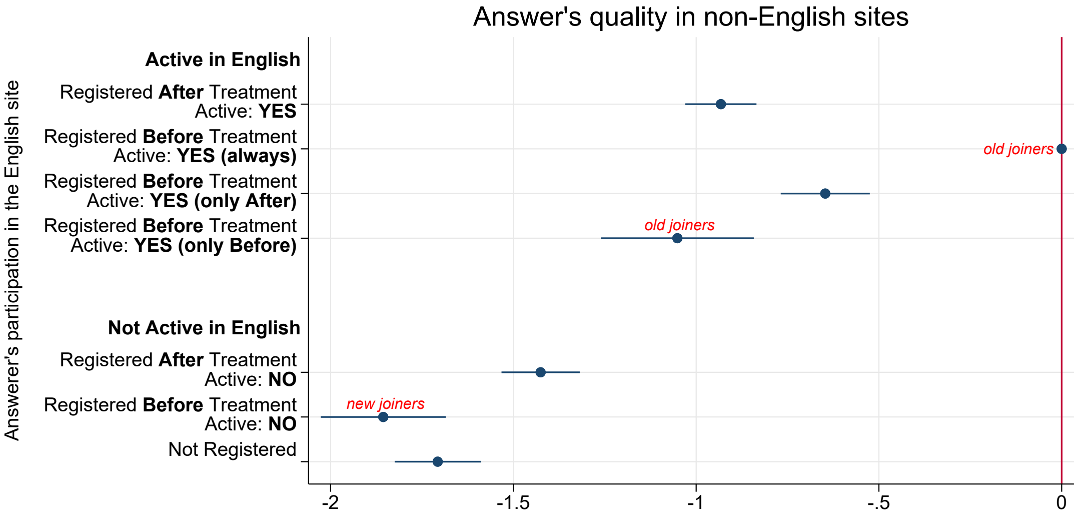
Note that since the analysis only uses data from the non-English sites, the concerns about systematic differences between the English and non-English sites described in section B.5 in the appendix are not applicable.

8.2.1 Estimation method

The unit of analysis is an answer i. Let $j(i)$ denote the author of the answer, $q(i)$ the question it is answering, and $r(i)$ the order by date of publication such that, for instance, $r(i) = 1$ if i is the first answer published for the question $q(i)$. Let G be the set of user types based on whether they registered and eventually participated on the English site. Finally, let $t_{j(i),i}$ be the number of days since the first registration of user j on the Stack Overflow platform until the publication of the answer i. This variable aims to capture the users' experience and knowledge of the platform's functioning. The empirical specification is the following:

$$numCodes_i = \alpha_{q(i)} + \delta_{r(i)} + \sum_{g \in G} \beta_g D_{g(j(i))} + \zeta t_{j(i),i} + \varepsilon_i,$$

where α and δ are the question and order-of-publication fixed effects, while the βs are fixed effect estimates for each group of users.



Notes. Estimates for user group fixed effects on answer quality. The dependent variable is the answers' quality and is proxied with the number of pieces of code included in the answer. The baseline group is users who contributed in English both before and after treatment.

Figure 4: Difference in answer quality between *new joiners* and *old joiners*.

8.3 Results

Figure 4 reports the estimates of $\{\beta_g\}|_{g \in G}$ and the corresponding confidence intervals. It shows that, for a given question, *new joiners* produce significantly lower-quality answers compared to *old joiners*, as the theory suggests.[41] This result highlights the potential role of language costs as barriers to participation for low-expertise users.

9 Managerial Implications

Similar to other knowledge platforms, Stack Overflow needs to decide whether to allow contributions in multiple languages. Several dimensions of reasoning can affect the decision, including ethics, inclusivity, stakeholder preferences, and efficiency. Given the contribution of this paper, this discussion focuses on the latter, particularly on the trade-offs concerning the quality of contributions.[42]

By introducing multiple languages, the platform alleviates communication costs for users native to those languages. First, it allows users who were already contributing to increase their answers' quality. Second, it allows some of the users who were not participating to start contributing.

[41] Detailed regression estimates and robustness exercises with alternative quality measures are available in section F.1 in the appendix.

[42] For further discussions of the trade-offs between efficiency and ethics concerning homogenisation versus diversity of languages, see Ginsburgh and Weber (2011), Blanc and Kubo (2021), and Blouin and Dyer (2022).

The first effect is substantial. On average, when they can use their native language in addition to English, users include 0.586 additional pieces of code, corresponding to a 21.4% increase from the pre-treatment average. This implies a substantial increase in information clarity and the amount of information provided. The increase in quality is most likely beneficial to the platform and may also increase consumer satisfaction as questioners are more likely to find a solution to their problem. The probability that an answer is *accepted* increases by 6.2 percentage points when the answerers are able to use their native language. This means that answers are 17.2% more likely to solve the questioners' problems, suggesting that the increase in quality leads to more effective and valuable communication.

However, the second effect may offset the overall benefits of the platform. The participants who start contributing because of the availability of their native language tend to provide lower-quality answers. Their answers have between 0.94 and 1.96 fewer pieces of code than the answers of users who are active in English. These values correspond to a lower quality of 0.22 to 0.46 standard deviations, respectively. In addition, the probability that the questioners *accept* the answer as a solution to their question is 7.3 to 13.8 percentage points lower. As suggested in this paper, a possible mechanism is that language barriers prevent the participation of low-expertise users. By reducing the cost of communication, the platform loses a screening device and may experience an inflow of lower-quality content.

These results suggest that the platform, when deciding whether to implement additional languages, should take into consideration the sizes of the affected communities. In addition, the results underline how the effective implementation of new languages should be careful about the design of incentives for both questioners and answerers. Indeed, the increase in answer quality is nearly twice as large when the question's quality is in the top quartile rather than in the bottom one. This gap is even bigger with respect to direct incentives to the answerer. When answerers have high virtual rewards at stake, their increase in quality is five times larger than when they have no stakes.[43]

9.1 Other efficiency-related trade-offs

While the analysis in this paper does not allow us to derive the implications on dimensions other than the contributions' quality, the platform faces additional efficiency trade-offs, namely on the number of contributions.

First, if users are time-constrained, they may only contribute to one of the sites. By introducing multiple languages, the platform can experience a reduction in the number of contributions in English. Indeed, in the baseline sample, 22% of answerers who were active in English before treatment stopped contributing in English when their native language became available. For similar reasons, other users may maintain participation on multiple sites but publish fewer answers on each site compared to their contribution in English if only the English site were available. The dispersion of content that would

[43] An additional cost on the platform quality trade-off may arise from negative externalities on the English site. Section G.2 in the appendix shows that is not the case and, if anything, there are positive spillovers.

be in English or other languages may increase search costs for information seekers or even exclude them from some information. Bao, Hecht, Carton, Quaderi, Horn, and Gergle (2012) find evidence of such dispersion on Wikipedia.

Second, the multiplicity of sites may lead to the provision of the same information multiple times. From an efficiency perspective, this is suboptimal as the concentration of effort on a unique site increases the potential for information aggregation.

Third, introducing the availability of multiple languages may affect the overall community size. On the one hand, the availability of languages other than English may allow users who do not know English to join the platform. This is particularly relevant in Q&A platforms where free riding and undersupply of answers may hinder the success of the site. On the other hand, it may fragment the community based on the language used, creating smaller sub-communities. The reduction in community size may directly impact the number of contributions and cause spillovers from the contributing decisions of the remaining users, as Zhang and Zhu (2011) find for Wikipedia.

Finally, the availability of multiple languages may allow the platform to attract users from competing platforms that use those languages (Jeon, Jullien, and Klimenko 2021).

10 Conclusion

Knowledge platforms like Stack Overflow aim to create a valuable source of information by aggregating knowledge across the internet. Compared to traditional resources like paper encyclopaedias, these platforms face the challenge of ensuring that crowdsourced information is reliable and of high quality.

This paper studied the quality trade-offs knowledge platforms face when making their website available in one or multiple languages. The results show that the benefits of allowing contributions in multiple languages in addition to English are substantial. Users native to those languages experience a reduction in their cost of communication, leading to higher-quality, more effective contributions. At the same time, these users do not reduce the quality of their contributions in English. However, as language costs act as barriers to participation, the availability of more languages also induces contributions from users who, on average, have lower expertise.

While this analysis is specific to the context of Stack Overflow, the results may apply to different environments. Extensive literature has addressed communication costs as a major constraint to efficient economic activities; however, to my knowledge, it has not yet quantified the problem. This is relevant to 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 hierarchical structure of the company, managers need to evaluate the advantages of hiring employees who can act as *translators* and allow synergies between teams that use different specialised languages (Crémer et al. 2007). Finally, national states may want to understand the benefits of setting a homogeneous language before limiting individual freedom and the cultural traits of minorities.

Finally, the results provide insights on how the introduction of external technologies, such as live translators and search engines that allow for searches across languages, may impact information quality on the internet (Brynjolfsson, Hui, and Liu 2019). Such technologies, if sufficiently effective, would reduce language barriers. This may benefit knowledge platforms, as they would not have to implement new languages in their design. However, it prevents them from using language barriers as a tool to screen participants.

Future research should be devoted to understanding the external validity of the results in different environments.

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. 22

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. 31

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

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. 46

Blanc, G. and M. Kubo (2021). Schools, language, and nations: Evidence from a natural experiment in France. *Working paper*. 5, 29

Blouin, A. and J. Dyer (2022). How cultures converge: An empirical investigation of trade and linguistic exchange. *Working paper*. 29

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 (2022). Revisiting event study designs: Robust and efficient estimation. *Working Paper*. 4, 21, 47, 61, 62, 63

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. 32

Burtch, G., A. Ghose, and S. Wattal (2014). Cultural Differences and Geography as Determinants of Online Prosocial Lending. *MIS Quarterly 38*(3), 773–794. 46

Callaway, B. and P. H. C. Sant'Anna (2020, December). Difference-in-Differences with multiple time periods. *Journal of Econometrics*. 4, 21

Calvó-Armengol, A., J. de Martí, and A. Prat (2015). Communication and influence. *Theoretical Economics 10*, 649–690. 9, 11

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. 5

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. 31

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, 31

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, 21

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill. 37

Dessein, W., A. Galeotti, and T. Santos (2016, June). Rational inattention and organizational focus. *American Economic Review 106*(6), 1522–36. 9, 11

Dessein, W. and T. Santos (2006). Adaptive Organizations. *Journal of Political Economy 114*(5), 956–995. 9, 11

Dewatripont, M. and J. Tirole (2005). Modes of communication. *Journal of Political Economy 113*(6), 1217–1238. 2, 9, 10

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. 5, 29

Ginsburgh, V. and S. Weber (2020, June). The Economics of Language. *Journal of Economic Literature 58*(2), 348–404. 46

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. 21

Grinblatt, M. and M. Keloharju (2001). How Distance, Language, and Culture Influence Stockholdings and Trades. *The Journal of Finance 56*(3), 1053–1073. 46

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. 31

Jin, C. (2022). Does competition improve information quality: Evidence from the security analyst market. *Working Paper*. 49

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. 31, 46

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, 21

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. 5

Vives, X. (2008). *Information and Learning in Markets: The Impact of Market Microstructure*. Princeton University Press. 37

Wang, G. A., H. J. Wang, J. Li, A. S. Abrahams, and W. Fan (2014, December). An Analytical Framework for Understanding Knowledge-Sharing Processes in Online Q&A Communities. *ACM Transactions on Management Information Systems 5*(4), 18:1–18:31. 49

Ye, J., S. Han, Y. Hu, B. Coskun, M. Liu, H. Qin, and S. Skiena (2017, November). Nationality Classification Using Name Embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, New York, NY, USA, pp. 1897–1906. Association for Computing Machinery. 52

Zhang, T. and F. J. Oles (2001, April). Text Categorization Based on Regularized Linear Classification Methods. *Information Retrieval 4*(1), 5–31. 53

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. 31

Additional material

Appendix A Details about Stack Overflow

A.1 The introduction of new websites

The creation of new Stack Overflow websites follows a specific process. The main objective is to ensure, before the launch, a sufficiently active community base that will guarantee the website's growth and long-term sustainability. First of all, the website is proposed in an ad-hoc platform called *Area 51* where registered users can support the proposal and start publishing questions and answers. If the website idea receives enough attention and contributions, it proceeds to the *beta* period, gets its URL, and becomes accessible as an independent site. The *beta* period is split into two steps. First, in the so-called *private beta*, only users who actively supported 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 vary slightly. For example, some *privileges* are reachable with different amounts of points, with generally lower requirements in earlier stages.[44]

Data is available starting from the *private beta* period. Table 7 reports the dates for the start of each stage for websites in different languages.

Platform	Proposal	Private beta	Public beta	Graduation
English		01/08/2008	-	15/09/2008
Russian	01/06/2012	27/03/2015	27/03/2015	11/12/2015
Japanese	-	29/09/2014	16/12/2014	[not graduated]
Spanish	02/08/2012	01/12/2015	15/12/2015	17/5/2017
Portuguese	05/11/2010	12/12/2013	29/01/2014	15/5/2015

Table 7: Dates in which the platforms passed the different development stages.

Appendix B Details about the theoretical framework

I solve the model by backward induction.

B.1 Solving the model: second stage

In the second stage, the questioner observes the message m and chooses the action to address his problem and maximise his expected utility based on the message received.

[44] https://meta.stackexchange.com/questions/58587/reputation-requirements-compared

Note that the expectation is taken with respect to θ.

$$\begin{aligned}
& a^* \equiv \arg\max_a \mathbb{E}[-\left((a-\theta)^2 + C_Q^2\Phi_Q\right)|m] \\
\iff & a^* \equiv \arg\max_a -a^2 - \mathbb{E}[\theta^2|m] + 2a\mathbb{E}[\theta|m] + C_Q^2\Phi_Q \\
\iff & -2a^* + 2\mathbb{E}[\theta|m] = 0 \\
\iff & a^* = \mathbb{E}[\theta|m].
\end{aligned}$$

Since both θ and m are normally distributed, we can use Bayesian updating for normal random variables to compute the expectation.[45][46]

Recall that $m = \theta + \varepsilon + \eta$, $\theta \sim \mathcal{N}(0, \frac{1}{s})$, $\varepsilon \sim \mathcal{N}(0, \frac{1}{E_Q})$, $\eta \sim \mathcal{N}(0, \frac{1}{E_A})$, and $\{\theta, \varepsilon, \eta\}$ are independent. It follows that $(\theta, m) \sim \mathcal{N}(\mu, \Sigma)$ with:

$$\mu = \begin{bmatrix} \mu_\theta \\ \mu_m \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{\theta,\theta} & \Sigma_{\theta,m} \\ \Sigma_{m,\theta} & \Sigma_{m,m} \end{bmatrix} = \Sigma = \begin{bmatrix} \frac{1}{s} & \frac{1}{s} \\ \frac{1}{s} & \frac{1}{s} + \frac{1}{\Phi_Q} + \frac{1}{\Phi_A} \end{bmatrix}.$$

We then have that:

$$\begin{aligned}
a^* &= \mathbb{E}[\theta|m] \\
&= \mu_\theta + \frac{\Sigma_{\theta,m}}{\Sigma_{m,m}}(m - \mu_m) \\
&= \frac{\frac{1}{s}}{\frac{1}{s} + \frac{1}{E_Q} + \frac{1}{E_Q}} m \\
&= \beta m \quad \text{with} \quad \beta \equiv \frac{\Phi_Q\Phi_A}{\Phi_Q\Phi_A + \Phi_Q s + \Phi_A s}.
\end{aligned}$$

B.2 Solving the model: first stage

The answerer chooses the quality level to maximise her expected utility, where the expectation is with respect to θ, ε, and η, as the message is not yet realised.

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

[45] Note that to avoid introducing cumbersome notation, I am using m to identify both the random variable and its realisation.

[46] The result is reported in Vives (2008)'s technical appendix (section 10.2.1, page 376) and shown in DeGroot (1970). Consider two normal random variables $(\theta, s) \sim \mathcal{N}(\mu, \Sigma)$ such that the mean vector and the variance-covariance matrix correspond to $\mu = \begin{bmatrix} \mu_\theta \\ \mu_s \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{\theta,\theta} & \Sigma_{\theta,s} \\ \Sigma_{s,\theta} & \Sigma_{s,s} \end{bmatrix}$. Then, the conditional density of θ given s is $(\theta \mid s) \sim \mathcal{N}\left(\mu_\theta + \Sigma_{\theta,s}\Sigma_{s,s}^{-1}(s - \mu_s), \Sigma_{\theta,\theta} - \Sigma_{\theta,s}\Sigma_{s,s}^{-1}\Sigma_{s,\theta}\right)$.

Based on the action expected to be chosen by Bob, the problem rewrites as follows:

$$\begin{aligned}
&\max_{\Phi_A \geq 0} -\gamma \mathbb{E}[(\beta m - \theta)^2] - C_A^2 \Phi_A \\
\iff &\max_{\Phi_A \geq 0} -\gamma \mathbb{E}[\beta m - \theta]^2 - \gamma \mathbb{V}[\beta m - \theta] - C_A^2 \Phi_A \quad \text{(by property of the variance)} \\
\iff &\max_{\Phi_A \geq 0} -\gamma \mathbb{V}[\beta m - \theta] - C_A^2 \Phi_A \quad \text{(since } m \text{ and } \theta \text{ have zero mean)} \\
\iff &\max_{\Phi_A \geq 0} -\gamma \left(\beta^2 \mathbb{V}[m] + \mathbb{V}[\theta] - 2\beta \mathbb{V}[\theta]\right) - C_A^2 \Phi_A \\
&\quad \text{Note that } \mathbb{V}[m] = \frac{\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s}{s \Phi_Q \Phi_A} = \frac{1}{\beta s} \text{ since } \beta \equiv \frac{\Phi_Q \Phi_A}{\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s}, \text{ so:} \\
\iff &\max_{\Phi_A \geq 0} -\gamma \left(\beta^2 \frac{1}{\beta s} + \frac{1}{s} - 2\beta \frac{1}{s}\right) - C_A^2 \Phi_A \\
\iff &\max_{\Phi_A \geq 0} -\gamma \left(\beta \frac{1}{s} + \frac{1}{s} - 2\beta \frac{1}{s}\right) - C_A^2 \Phi_A \\
\iff &\max_{\Phi_A \geq 0} -\gamma \left(\frac{1}{s}(1 - \beta)\right) - C_A^2 \Phi_A = -\gamma \frac{1}{s} + \gamma \frac{1}{s} \beta - C_A^2 \Phi_A.
\end{aligned}$$

By first-order condition, the best response quality level satisfies:

$$\begin{aligned}
\frac{\partial}{\partial \Phi_A} \left(-\gamma \frac{1}{s} + \gamma \frac{1}{s} \frac{\Phi_Q \Phi_A}{\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s} - C_A^2 \Phi_A\right) &= 0 \\
\frac{\gamma \Phi_Q^2}{(\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s)^2} &= C_A^2 \equiv \left(\frac{\lambda_A}{k_A}\right)^2 \\
(\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s)^2 &= \frac{\gamma \Phi_Q^2 k_A^2}{\lambda_A^2} \\
\Phi_Q \Phi_A + \Phi_Q s + \Phi_A s &= \sqrt{\frac{\gamma \Phi_Q^2 k_A^2}{\lambda_A^2}} \\
\Phi_A(\Phi_Q + s) &= \frac{\sqrt{\gamma} \Phi_Q k_A}{\lambda_A} - \Phi_Q s \\
\Phi_A(\Phi_Q + s) &= \frac{\Phi_Q(\sqrt{\gamma} k_A - s\lambda_A)}{\lambda_A}.
\end{aligned}$$

The best response is then given by:

$$BR_A(\Phi_Q) = \frac{\Phi_Q(\sqrt{\gamma} k_A - s\lambda_A)}{\lambda_A(\Phi_Q + s)}. \tag{12}$$

Note that in the first-order condition, a second solution of the quadratic equation is excluded as it takes only negative values, which is not allowed by the model assumptions (i.e. the quality level is weakly positive).

B.3 Comparative statics

Under the condition that quality is positive (i.e. under the assumption that $\sqrt{\gamma}k_A > s\lambda_A$)), and everything else held constant, a marginal increase in language cost affects the quality choice in the following way:

$$\begin{aligned}\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A} &= \frac{\partial}{\partial \lambda_A}\left(\frac{\Phi_Q(\sqrt{\gamma}k_A - s\lambda_A)}{\lambda_A(\Phi_Q + s)}\right) = \frac{\partial}{\partial \lambda_A}\left(\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A(\Phi_Q + s)} - \frac{\Phi_Q s}{\Phi_Q + s}\right)\\ &= -\frac{(\Phi_Q + s)\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^2(\Phi_Q + s)^2}\\ &= -\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^2(\Phi_Q + s)} < 0.\end{aligned}$$

The second degree effect is given by:

$$\begin{aligned}\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A^2} &= \frac{\partial}{\partial \lambda_A}\left(-\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^2(\Phi_Q + s)}\right)\\ &= -\frac{2\lambda_A(\Phi_Q + s)(-\Phi_Q\sqrt{\gamma}k_A)}{\lambda_A^4(\Phi_Q + s)^2}\\ &= \frac{2\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^3(\Phi_Q + s)} > 0.\end{aligned}$$

The change in quality due to a marginal change in the language cost depends on the quality of the question:

$$\begin{aligned}\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A \partial \Phi_Q} &= \frac{\partial}{\partial \Phi_Q}\left(-\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^2(\Phi_Q + s)}\right)\\ &= -\frac{\sqrt{\gamma}k_A(\lambda_A^2(\Phi_Q + s)) - \lambda_A^2(\Phi_Q\sqrt{\gamma}k_A)}{\lambda_A^4(\Phi_Q + s)^2}\\ &= -\frac{\sqrt{\gamma}k_A s}{\lambda_A^2(\Phi_Q + s)^2} < 0.\end{aligned}$$

The change in quality due to a marginal change in the language cost depends on the incentive degree:

$$\begin{aligned}\frac{\partial BR_A(\Phi_Q)}{\partial \lambda_A \partial \gamma} &= \frac{\partial}{\partial \gamma}\left(-\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A^2(\Phi_Q + s)}\right)\\ &= \left(-\frac{\Phi_Q k_A}{\lambda_A^2(\Phi_Q + s)}\right)\frac{1}{2\sqrt{\gamma}}\\ &= -\frac{\Phi_Q k_A}{2\lambda_A^2(\Phi_Q + s)\sqrt{\gamma}} < 0.\end{aligned}$$

The change in quality due to a change in γ is given by:

$$\frac{\partial BR_A(\Phi_Q)}{\partial \gamma} = \frac{\partial}{\partial \gamma}\left(\frac{\Phi_Q(\sqrt{\gamma}k_A - s\lambda_A)}{\lambda_A(\Phi_Q + s)}\right) = \frac{\partial}{\partial \gamma}\left(\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A(\Phi_Q + s)} - \frac{\Phi_Q s}{\Phi_Q + s}\right)$$
$$= \frac{\Phi_Q k_A}{\lambda_A(\Phi_Q + s)}\frac{1}{2\sqrt{\gamma}} > 0.$$

The change in quality due to a change in k_A is given by:

$$\frac{\partial BR_A(\Phi_Q)}{\partial k_A} = \frac{\partial}{\partial k_A}\left(\frac{\Phi_Q(\sqrt{\gamma}k_A - s\lambda_A)}{\lambda_A(\Phi_Q + s)}\right) = \frac{\partial}{\partial k_A}\left(\frac{\Phi_Q\sqrt{\gamma}k_A}{\lambda_A(\Phi_Q + s)} - \frac{\Phi_Q s}{\Phi_Q + s}\right)$$
$$= \frac{\Phi_Q\sqrt{\gamma}}{\lambda_A(\Phi_Q + s)} > 0.$$

B.4 Details and theoretical derivation for how *old joiners* have higher expertise than *new joiners*

Section 4.3.2 provides theoretical predictions that rely on the result that *old joiners* have higher expertise than *new joiners*. This section aims to provide additional details and show that this relationship holds even if answerers differ in their cost of using English.

Let $j \in (0, 1)$ index answerers participating on a topic ω. Let k_j be the expertise of user j on the topic ω.[47] In addition, let $\lambda \in \Lambda$ define the cost of using English and λ' be the cost of using the native language (assumed fixed across users), with $\lambda' < \lambda \ \forall \lambda \in \Lambda$. Moreover, let NJ identify the set of *new joiners*, and OJ the set of *old joiners*, such that:

$$j \in NJ \quad \text{if} \quad \frac{s\lambda'}{\sqrt{\gamma}} < k_j < \frac{s\lambda_j}{\sqrt{\gamma}}$$
$$j \in OJ \quad \text{if} \quad k_j > \frac{s\lambda_j}{\sqrt{\gamma}},$$

where k_j is j's expertise on the topic ω and γ and s are assumed to be constants. Let $f(k)$ and $F(k)$ define, respectively, the probability density function and the cumulative density function over levels of expertise and across answerers. Assume that these distributions are fixed across different cost levels for using English (λ).

B.4.1 Case where all answerers have the same cost of using English

This is the case discussed in the main text. Since answerers are only allowed to differ in their expertise (k_j), we have that

$$\mathbb{E}\left[k_j|\omega, j \in OJ\right] \geq \mathbb{E}\left[k_j|\omega, j \in NJ\right]$$

by direct implication of the definition of *old joiners* and *new joiners*.

[47] For the sake of clarity, compared to the main text, I have dropped the subscript A on all notation that refers to expertise and language cost, as it is not needed for the exposition of this proof.

B.4.2 Case with two possible cost levels for the use of English

To provide a more tractable example, consider the case where there are two types of answerers: one has a high cost of using English ($\overline{\lambda}$) while the other has a low cost of using English ($\underline{\lambda}$). Figure 5 represents the setting in the particular case that f is the density of a normal distribution. On the English site, contributions would come from a share $1-F\left(\frac{s\underline{\lambda}}{\sqrt{\gamma}}\right)$ of low-cost answerers and $1-F\left(\frac{s\overline{\lambda}}{\sqrt{\gamma}}\right)$ of high-cost answerers. On the native language site, additional contributions would come from a share $F\left(\frac{s\underline{\lambda}}{\sqrt{\gamma}}\right)-F\left(\frac{s\lambda'}{\sqrt{\gamma}}\right)$ of low-cost answerers and a share $F\left(\frac{s\underline{\lambda}}{\sqrt{\gamma}}\right)-F\left(\frac{s\lambda'}{\sqrt{\gamma}}\right)$ of high-cost answerers. For the sake of readability, I use κ, $\overline{\kappa}$, $\underline{\kappa}$, and κ' to mean $\frac{s\lambda}{\sqrt{\gamma}}$, $\frac{s\overline{\lambda}}{\sqrt{\gamma}}$, $\frac{s\underline{\lambda}}{\sqrt{\gamma}}$, and $\frac{s\lambda'}{\sqrt{\gamma}}$, respectively. We then have that:

$$
\begin{aligned}
\mathbb{E}\left[k_j|\omega, j\in OJ\right] &= \frac{\int_{\underline{\kappa}}^{\infty} kf(k)dk + \int_{\overline{\kappa}}^{\infty} kf(k)dk}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]} \\
&= \frac{1-F(\underline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\underline{\kappa}}^{\infty} kf(k)dk}{1-F(\underline{\kappa})} + \frac{1-F(\overline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\overline{\kappa}}^{\infty} kf(k)dk}{1-F(\overline{\kappa})} \\
&= \frac{1-F(\underline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\left(\frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf(k)dk}{1-F(\underline{\kappa})} + \frac{\int_{\overline{\kappa}}^{\infty} kf(k)dk}{1-F(\underline{\kappa})}\right) + \frac{1-F(\overline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\overline{\kappa}}^{\infty} kf}{1-F} \\
&= \frac{1-F(\underline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\left(\frac{F(\overline{\kappa})-F(\underline{\kappa})}{1-F(\underline{\kappa})}\frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf(k)dk}{F(\overline{\kappa})-F(\underline{\kappa})} + \frac{1-F(\overline{\kappa})}{1-F(\underline{\kappa})}\frac{\int_{\overline{\kappa}}^{\infty} kf(k)dk}{1-F(\overline{\kappa})}\right) \\
&+ \frac{1-F(\overline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\overline{\kappa}}^{\infty} kf(k)dk}{1-F(\overline{\kappa})} \\
&= \frac{F(\overline{\kappa})-F(\underline{\kappa})}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf(k)dk}{F(\overline{\kappa})-F(\underline{\kappa})} + \frac{2[1-F(\overline{\kappa})]}{[1-F(\underline{\kappa})]+[1-F(\overline{\kappa})]}\frac{\int_{\overline{\kappa}}^{\infty} kf(k)dk}{1-F(\overline{\kappa})} \\
&= \frac{[F(\overline{\kappa})-F(\underline{\kappa})]B + 2[1-F(\overline{\kappa})]C}{[F(\overline{\kappa})-F(\underline{\kappa})]+2[1-F(\overline{\kappa})]},
\end{aligned}
$$

where B is the average expertise for levels between $\underline{\kappa}$ and $\overline{\kappa}$ while C is the average expertise for levels above $\overline{\kappa}$ (i.e. of areas $\mathscr{B}$ and $\mathscr{C}$ in figure 5). For the *new joiners*, we have that:

$$
\begin{aligned}
\mathbb{E}\left[k_j|\omega, j \in NJ\right] &= \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk + \int_{\kappa'}^{\overline{\kappa}} kf(k)dk}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \\
&= \frac{F(\underline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\underline{\kappa}) - F(\kappa')} + \frac{F(\overline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\kappa'}^{\overline{\kappa}} kf}{F(\overline{\kappa}) -} \\
&= \frac{F(\underline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\underline{\kappa}) - F(\kappa')} \\
&+ \frac{F(\overline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \left(\frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\overline{\kappa}) - F(\kappa')} + \frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf(k)dk}{F(\overline{\kappa}) - F(\kappa')} \right) \\
&= \frac{F(\underline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\underline{\kappa}) - F(\kappa')} \\
&+ \frac{F(\overline{\kappa}) - F(\kappa')}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \left(\frac{F(\underline{\kappa}) - F(\kappa')}{F(\overline{\kappa}) - F(\kappa')} \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\underline{\kappa}) - F(\kappa')} + \frac{F(\overline{\kappa}) - F(\underline{\kappa})}{F(\overline{\kappa}) - F(\kappa')} \frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf(k)}{F(\overline{\kappa}) - F} \right. \\
&= \frac{2[F(\underline{\kappa}) - F(\kappa')]}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\kappa'}^{\underline{\kappa}} kf(k)dk}{F(\underline{\kappa}) - F(\kappa')} + \frac{F(\overline{\kappa}) - F(\underline{\kappa})}{[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\kappa')]} \frac{\int_{\underline{\kappa}}^{\overline{\kappa}} kf}{F(\overline{\kappa}) -} \\
&= \frac{2[F(\underline{\kappa}) - F(\kappa')]A + [F(\overline{\kappa}) - F(\underline{\kappa})]B}{2[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\underline{\kappa})]},
\end{aligned}
$$

where A is the average expertise for levels below $\underline{\kappa}$ (i.e. of area $\mathscr{A}$ in figure 5).

Note that $B < C$. Indeed:

$$
\begin{aligned}
\int_{\underline{\kappa}}^{\infty} kf(k)dk &= \overline{\kappa} + \int_{\underline{\kappa}}^{\overline{\kappa}} (k - \overline{\kappa})f(k)dk + \int_{\overline{\kappa}}^{\infty} (k - \overline{\kappa})f(k)dk \\
&\leq \overline{\kappa} + \int_{\overline{\kappa}}^{\infty} (k - \overline{\kappa})f(k)dk \\
&= \overline{\kappa} + [1 - F(\overline{\kappa})] \frac{\int_{\overline{\kappa}}^{\infty} (k - \overline{\kappa})f(k)dk}{1 - F(\overline{\kappa})} \\
&\leq \overline{\kappa} + \frac{\int_{\overline{\kappa}}^{\infty} (k - \overline{\kappa})f(k)dk}{1 - F(\overline{\kappa})} = C,
\end{aligned}
$$

where the first inequality holds as it removes something weakly negative, while the second inequality holds because $0 \leq [1 - F(\overline{\kappa})] \leq 1$ and $\int_{\overline{\kappa}}^{\infty}(k - \overline{\kappa})f(k)dk > 0$. Similarly,

$$\begin{aligned}
\int_{\underline{\kappa}}^{\infty} kf(k)dk &= \overline{\kappa} + \int_{\underline{\kappa}}^{\overline{\kappa}}(k - \overline{\kappa})f(k)dk + \int_{\overline{\kappa}}^{\infty}(k - \overline{\kappa})f(k)dk \\
&\geq \overline{\kappa} + \int_{\underline{\kappa}}^{\overline{\kappa}}(k - \overline{\kappa})f(k)dk \\
&= \overline{\kappa} + [F(\overline{\kappa}) - F(\underline{\kappa})]\frac{\int_{\underline{\kappa}}^{\overline{\kappa}}(k - \overline{\kappa})f(k)dk}{F(\overline{\kappa}) - F(\underline{\kappa})} \\
&\geq \overline{\kappa} + \frac{\int_{\underline{\kappa}}^{\overline{\kappa}}(k - \overline{\kappa})f(k)dk}{F(\overline{\kappa}) - F(\underline{\kappa})} = B,
\end{aligned}$$

where the first inequality holds as it removes something weakly positive, while the second inequality holds because $0 \leq [F(\overline{\kappa}) - F(\underline{\kappa})] \leq 1$ and $\int_{\overline{\kappa}}^{\infty}(k - \overline{\kappa})f(k)dk < 0$. It follows that:

$$B < \int_{\underline{\kappa}}^{\infty} kf(k)dk < C \implies B < C.$$

Similarly, it is possible to show that $A < B$. It follows that $A < B < C$.

Since:

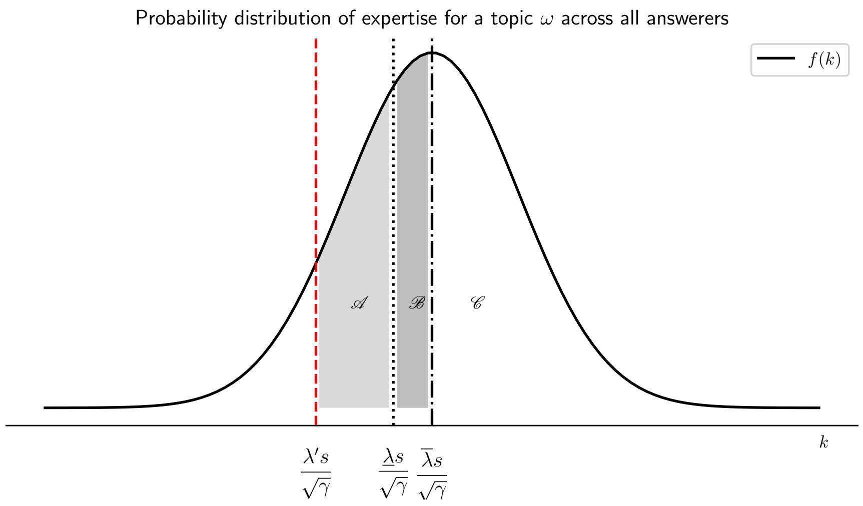
$$\begin{aligned}
\mathbb{E}\left[k_j|\omega, j \in OJ\right] &= \frac{[F(\overline{\kappa}) - F(\underline{\kappa})]B + 2[1 - F(\overline{\kappa})]C}{[F(\overline{\kappa}) - F(\underline{\kappa})] + 2[1 - F(\overline{\kappa})]} \geq C \\
\mathbb{E}\left[k_j|\omega, j \in NJ\right] &= \frac{2[F(\underline{\kappa}) - F(\kappa')]A + [F(\overline{\kappa}) - F(\underline{\kappa})]B}{2[F(\underline{\kappa}) - F(\kappa')] + [F(\overline{\kappa}) - F(\underline{\kappa})]} \leq C,
\end{aligned}$$

then:

$$\mathbb{E}\left[k_j|\omega, j \in OJ\right] \geq \mathbb{E}\left[k_j|\omega, j \in NJ\right].$$

B.4.3 Case with L possible cost levels for the use of English

Consider the general case where there are L possible cost levels (i.e. $\#\Lambda = L$). Define a superscript $l = 1, ..., L$ to index possible cost levels ranked in descending order, such that $\max(\Lambda) = \lambda^1$ and $\min(\Lambda) = \lambda^L$. For the sake of readability, I use the notation κ^l



Notes. When the cost of language decreases to λ', new answerers with a cost of using English equal to $\underline{\lambda}$ and expertise levels corresponding to the light grey area start to participate. By contrast, the new answerers with a cost of using English equal to $\overline{\lambda}$ have expertise corresponding to both grey areas. Indeed, users in the dark grey area participate in English if they have a low cost of English and do not participate in English if they have a high cost of English. These users may have higher expertise compared to some answerers with a low cost of using English who were already active in English.

Figure 5: Probability distribution of expertise across answerers.

to mean $\frac{s\lambda^l}{\sqrt{\gamma}}$ and define $K \equiv \{\kappa^l\}_{\forall l}$. We then have that:

$$\begin{aligned}
\mathbb{E}\left[k_j | \omega, j \in OJ\right] &= \sum_{\kappa \in K} \left[\frac{\int_\kappa^\infty k f(k) dk}{\sum_{\kappa \in K}(1 - F(\kappa))} \right] \\
&= \sum_{\kappa \in K} \left[\frac{\int_\kappa^{\kappa^1} k f(k) dk}{\sum_{\kappa \in K}(1 - F(\kappa))} + \frac{1 - F(\kappa^1)}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^1}^\infty k f(k) dk}{1 - F(\kappa^1)} \right] \\
&= \sum_{\kappa \in K \setminus \{\kappa^1\}} \left[\frac{\int_\kappa^{\kappa^1} k f(k) dk}{\sum_{\kappa \in K}(1 - F(\kappa))} \right] + \frac{L(1 - F(\kappa^1))}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^1}^\infty k f(k) dk}{1 - F(\kappa^1)} \\
&= \sum_{\kappa \in K \setminus \{\kappa^1\}} \left[\frac{\int_\kappa^{\kappa^2} k f(k) dk}{\sum_{\kappa \in K}(1 - F(\kappa))} + \frac{F(\kappa^1) - F(\kappa^2)}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^2}^{\kappa^1} k f(k) dk}{F(\kappa^1) - F(\kappa^2)} \right] \\
&+ \frac{L(1 - F(\kappa^1))}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^1}^\infty k f(k) dk}{1 - F(\kappa^1)} \\
&= \sum_{\kappa \in K \setminus \{\kappa^1, \kappa^2\}} \left[\frac{\int_\kappa^{\kappa^2} k f(k) dk}{\sum_{\kappa \in K}(1 - F(\kappa))} \right] + \frac{(L-1)(F(\kappa^1) - F(\kappa^2))}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^2}^{\kappa^1} k f(k) dk}{F(\kappa^1) - F(\kappa^2)} \\
&+ \frac{L(1 - F(\kappa^1))}{\sum_{\kappa \in K}(1 - F(\kappa))} \frac{\int_{\kappa^1}^\infty k f(k) dk}{1 - F(\kappa^1)}.
\end{aligned}$$

By iterating the reasoning, we can derive the general statement:

$$\begin{aligned}\mathbb{E}\left[k_j|\omega, j \in OJ\right] &= \sum_{l=1,\dots,L-1}\left[\frac{(L-l)(F(\kappa^l)-F(\kappa^{l+1}))}{\sum_{\kappa\in K}(1-F(\kappa))}\frac{\int_{\kappa^{l+1}}^{\kappa^l} kf(k)dk}{F(\kappa^l)-F(\kappa^{l+1})}\right] + \frac{L(1-F(\kappa^1))}{\sum_{\kappa\in K}(1-F(\kappa))}\frac{\int_{\kappa^1}^{\infty} kf(k)dk}{1-F(\kappa^1)}\\ &= \sum_{l=1,\dots,L-1}\left[\frac{(L-l)(F(\kappa^l)-F(\kappa^{l+1}))}{\sum_{\kappa\in K}(1-F(\kappa))}B^l\right] + \frac{L(1-F(\kappa^1))}{\sum_{\kappa\in K}(1-F(\kappa))}C,\end{aligned}$$

where $\{B^l\}_{l=1,\dots L-1}$ is the sequence of average expertise values for ranges of k determined by intermediate cost levels of using English. In other words, by letting λ^1 be equal to $\overline{\lambda}$ in the binary cost case and λ^L be equal to $\underline{\lambda}$, the area $\mathscr{B}$ in figure 5 is now split in $L-1$ areas defined by the thresholds $\{\kappa^l\}_{\forall l}$. The values $\{B^l\}_{l=1,\dots L-1}$ are the average values for each area. Meanwhile, C is the average expertise of *old joiners* that have a cost of using English equal to λ^1.

Similarly, for *new joiners* and $\kappa' \equiv \frac{s\lambda'}{\sqrt{\gamma}}$, we have that:

$$\begin{aligned}\mathbb{E}\left[k_j|\omega, j \in NJ\right] &= \sum_{\kappa\in K}\left[\frac{\int_{\kappa'}^{\kappa} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\right]\\ &= \sum_{\kappa\in K}\left[\frac{\int_{\kappa^L}^{\kappa} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))} + \frac{F(\kappa^L)-F(\kappa')}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa'}^{\kappa^L} kf(k)dk}{F(\kappa^L)-F(\kappa')}\right]\\ &= \sum_{\kappa\in K\setminus\{\kappa^L\}}\left[\frac{\int_{\kappa^L}^{\kappa} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\right] + \frac{L(F(\kappa^L)-F(\kappa'))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa'}^{\kappa^L} kf(k)dk}{F(\kappa^L)-F(\kappa')}\\ &= \sum_{\kappa\in K\setminus\{\kappa^L\}}\left[\frac{\int_{\kappa^{L-1}}^{\kappa} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))} + \frac{F(\kappa^{L-1})-F(\kappa^L)}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa^L}^{\kappa^{L-1}} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa^{L-1})-F(\kappa^L))}\right]\\ &+ \frac{L(F(\kappa^L)-F(\kappa'))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa'}^{\kappa^L} kf(k)dk}{F(\kappa^L)-F(\kappa')}\\ &= \sum_{\kappa\in K\setminus\{\kappa^L,\kappa^{L-1}\}}\left[\frac{\int_{\kappa^{L-1}}^{\kappa} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\right] + \frac{(L-1)(F(\kappa^{L-1})-F(\kappa^L))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa^L}^{\kappa^{L-1}} kf(k)dk}{\sum_{\kappa\in K}(F(\kappa^{L-1})-F(\kappa^L))}\\ &+ \frac{L(F(\kappa^L)-F(\kappa'))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa'}^{\kappa^L} kf(k)dk}{F(\kappa^L)-F(\kappa')}.\end{aligned}$$

By iterating the reasoning, we can derive the following general statement:

$$\begin{aligned}\mathbb{E}\left[k_j|\omega, j \in NJ\right] &= \sum_{l=1,\dots,L-1}\left[\frac{l(F(\kappa^l)-F(\kappa^{l+1}))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa^{l+1}}^{\kappa^l} kf(k)dk}{F(\kappa^l)-F(\kappa^{l+1})}\right] + \frac{L(F(\kappa^L)-F(\kappa'))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}\frac{\int_{\kappa'}^{\kappa^L} kf(k)dk}{F(\kappa^L)-F(\kappa')}\\ &= \sum_{l=1,\dots,L-1}\left[\frac{l(F(\kappa^l)-F(\kappa^{l+1}))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}B^l\right] + \frac{L(F(\kappa^L)-F(\kappa'))}{\sum_{\kappa\in K}(F(\kappa)-F(\kappa'))}A,\end{aligned}$$

where A is the average expertise of *new joiners* with a cost of using English equal to κ^L.

Using the same reasoning as in the binary cost case, it is possible to show that $A < B^{L-1} < ... < B^1 < C$. In addition, as was the case in the binary cost case, it can be noticed that the average expertise of *old joiners* is the weighted average of $\{B^l\}_{l=1,...L-1}$ and A, while the average expertise for *new joiners* is the weighted average of $\{B^l\}_{l=1,...L-1}$ and C. Since the weights are weakly larger on larger values for the *old joiners* and weakly larger on smaller values for the *new joiners*, it follows that:

$$\mathbb{E}\left[k_j|\omega, j \in OJ\right] \geq \mathbb{E}\left[k_j|\omega, j \in NJ\right],$$

as was the case in the binary cost case.

B.5 Remark on the endogenous self-selection of questioners across sites

The theoretical predictions discussed in sections 4.3.1 and 4.3.2 rely on the assumption that the questions published on Stack Overflow are comparable across sites and the language used is the only substantial difference. This implies that Alice expertise k_A and incentives γ are fixed across English and Spanish questions.

This assumption may not hold in reality. Indeed, the choice of publishing a question may depend on site-specific characteristics, leading questioners to self-select on certain sites. A leading factor is the number of participants, which is smaller on non-English sites and affects the chances of receiving an answer and the speed of answer arrival. Another factor is the availability of expertise in specific topics. Users who are native Spanish speakers may, for instance, be more expert in specific tasks than other users, attracting questioners with questions on those tasks to the Spanish site.

This non-random distribution of questioners and questions across sites may cause systematically different expertise (k_A) and degree of incentives (γ) for the answerer. For instance, if the cultural pool of participants on the Spanish site is more homogeneous and Alice perceives them as culturally closer to her, she may care more for them to solve their problems. In this case, on average, Alice would have a higher γ on the Spanish site than on the English one.[48] At the same time, if the questioners' endogenous choice of websites depends on culture-specific skills, the topics of questions on the Spanish website may be more familiar to Alice than the topics of questions published in English. Her expertise may then be systematically higher on the Spanish website than on the English one.

The data do not allow for testing these hypotheses directly. In the empirical sections, I discuss the implications for identification and develop robustness tests.

[48] The literature has shown that in various environments, people are more willing to interact with culturally closer individuals because of empathy, trust, and other reasons. For instance, Grinblatt and Keloharju (2001) show that geographical proximity and language drive investors' decisions, while Burtch, Ghose, and Wattal (2014) show that cultural proximity affects borrowing choices on a crowdfunding platform. In addition, Lyons (2017) shows that teams with the same nationality are more productive, while BenYishay and Mobarak (2019) find that group identity affects communication effectiveness. Ginsburgh and Weber (2020) provide an extensive review of research on how language affects behaviour and economic interactions.

Appendix C Details on estimation

C.1 Details on identification and possible violations of the parallel trend assumption for the estimation of the *old joiners*' treatment effect

Consider the example for which Alice is a native Spanish speaker.[49] Let λ_A^{ENG} be her cost of communicating in English when English is her only available language. In addition, let λ_A^{SPA} be her cost of communicating in Spanish. For an answer i published by Alice on the Spanish website at the time t, the treatment effect is given by:

$$\mathbb{E}\left[\Phi_{Ait}\left(\lambda_A^{SPA}\right) - \Phi_{Ait}\left(\lambda_A^{ENG}\right)\right], \tag{13}$$

where $\Phi_{Ait}(\lambda_A^{ENG})$ is the counterfactual quality choice that Alice would have made if she had answered the same question without the option of using Spanish (i.e. the *potential outcome*). While $\Phi_{Ait}(\lambda_A^{SPA})$ is observed, $\Phi_{Ait}(\lambda_A^{ENG})$ is not. The identification of the treatment effect requires the parallel trend assumption and the assumption of no anticipation to hold. More formally, let j index the answers' authors (i.e. Alice, or any other user providing answers) and $\Phi_{Ai(j)t}$ be the quality of the answer i at the time t made by user j. The parallel trend assumption states that $\mathbb{E}\left[\Phi_{Ai(j)t}\left(\lambda_A^{ENG}\right)\right] = \alpha_j + \delta_t$. It implies that, in the absence of the treatment, Alice's quality choice is determined by a time-invariant component specific to her (α_j) and a time-specific component fixed across users (δ_t). In other words, $\mathbb{E}\left[\Phi_{Ai(j)t}\left(\lambda_A^{ENG}\right) - \Phi_{Ai(j)t'}\left(\lambda_A^{ENG}\right)\right]$ is the same across all users (i.e. $\forall j$) for all periods t and t' (whenever $i(j)t$ and $i(j)t'$ are observed). The no-anticipation assumption imposes that when the Spanish website is not yet available, Alice's quality choice is not affected by the fact that it will be.

In the setting of this paper, the no-anticipation assumption is naturally satisfied. In the absence of the Spanish website, Alice can only write in English and her cost of language is not affected even if she anticipates the arrival of the Spanish website. The satisfaction of the parallel trend assumption is more challenging. Indeed, as the theoretical framework shows, several factors affecting the quality choice are answer-specific and are not necessarily fixed across users and time. These include systematic differences across sites on questions' quality, community size, and the kind of topics discussed.

First of all, the question's quality may depend on the questioner's cost of language. Assuming that the questioners writing on the Spanish site are native Spanish speakers, there is the possibility that their questions' average quality in Spanish is higher than on the English site. The theoretical framework would then suggest that Alice's quality choice is significantly higher on the Spanish site independently of changes in her cost of language. To address this problem, I include the question's quality ($QQuality$) in the regression as a control variable.

[49] For most of the paper, I use Borusyak et al. (2022) as the main reference for the econometric modelling. While I discuss the empirical strategy and identification using Spanish as the non-English language, the discussion extends to other languages.

A second issue that may violate the assumption is that the pool of questioners may be structurally different across websites, as questioners choose where to participate based on site characteristics.[50] This may lead to the following two main confounding effects.

1. Alice's parameter γ may be systematically higher on the Spanish site.
   γ represents the degree to which Alice cares for the questioner to solve his problem. If γ is higher on the Spanish site, Alice will choose higher quality in Spanish than in English independently of changes in her cost of language.[51] To address this issue, I exploit variation on the English site over Alice's ability to identify the questioner as a culturally close individual. Indeed, Alice can observe the profile page of the questioner. Since users are free to decide whether to include informative items on their profile page, including their name, place of residence, and profile picture, Alice may or may not learn the cultural identity of the questioner.

   Using the questioners' profile page data, I construct two proxies of cultural proximity. The first proxy takes a value equal to 1 if the questioner is identified as having the same native language as the answerer, and 0 otherwise. For instance, the first proxy takes a value equal to 1 if the questioner lives in Spain's capital or displays a Spanish name. The second proxy takes a value equal to 1 if the questioner, besides sharing the same native language, displays a manually uploaded profile picture. The latter variable relies on the assumption that the personalised profile picture contains traits supposedly relatable to the answerer if they share the same native language. The analysis includes these proxies as control variables named *empathy*.[52]

2. Alice's expertise (k_A) may be systematically higher on the Spanish site.
   This may happen if the topics addressed on the Spanish site are systematically different and more familiar to Alice than the topics discussed on the English site.[53] To address this issue, I execute a test that compares the words in the questions' titles across sites and identifies those that are site-specific. This procedure allows us to exclude answers that may be addressing site-specific topics and check if the results depend on those.[54]

A third issue that may harm identification is the difference in community size across websites. As discussed, while this may have an indirect impact through the self-selection of questioners, it can also directly impact the degree of competition. A larger number

[50] Section B.5 in the appendix discusses this issue.

[51] The first derivative of the answerer's best response quality is shown in section B.3 in the appendix.

[52] Section D.3 in the appendix provides details on the construction of these variables. Note that these variables are not available for the *never-treated* users as it is impossible to guess their native language.

[53] Systematic differences in the topics discussed can also be problematic as they may induce mechanical differences on the proxy measure of quality across sites. For example, some topics may not require pieces of code in the answers. This possibility is discussed in section D.2 in the appendix, while robustness analysis is in section E.3 in the appendix. Note that, in this case, there is no prior intuition on the direction of the potential bias.

[54] Details of the procedure and the robustness checks are in sections D.4 and E.2 in the appendix, respectively.

of competing answerers for the same question can affect the quality of Alice's answer in several ways and in potentially opposite directions. The quality may decrease if the higher competition makes Alice feel more rushed or increase if the competition creates pressure to produce the best answer. In addition, Alice's answer may just complement already existing answers, adding little information. The proxy for quality would be low in this case since it does not incorporate complementarities between answers.[55]

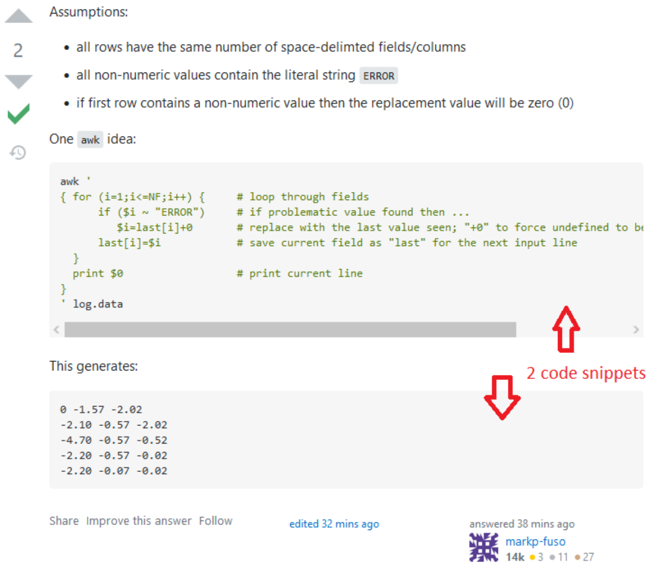
This issue cannot be explained through the model as the model abstracts from strategic behaviour that depends on other answerers' strategies and network effects. To avoid mismeasurement due to this issue, I include measures of competition and active community size as control variables. More precisely, for each answer i in the sample, let $q(i)$ be the question it addresses. The set of control variables referred to as *Competition* includes 1) the number of published answers for question $q(i)$ (including i) and 2) the number of views received by $q(i)$.

Appendix D Details about the data and the measures

D.1 Quality measure based on code snippets

Figure 6 provides an example of how the quality measure is computed based on code snippets.

[55] For a different question-and-answer site, Wang, Wang, Li, Abrahams, and Fan (2014) and the references therein find that the number of answers in the thread affects the probability that the questioner solves his problem. While they evaluate the impact on outcomes at the thread level, their evidence suggests a relationship between the answers' quality and the number of answers to the question. In a different context, Jin (2022) shows that competition in information provision with rewards on relative accuracy induces incentives for accuracy but also incentives to differentiate from competitors.



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
```

Notes. In this example, there are two snippets of code, as identified by the red arrows. The proxy for quality would then be equal to 2.

Figure 6: Example of an answer in Stack Overflow.

D.2 Correlation between proxy for quality and *likes*

Table 8 reports OLS estimates at the answer level to show the correlation between the variables used as a proxy for quality and the *likes* received. More precisely, the dependent variable in the linear model is the score equal to the number of *likes* minus the number of *dislikes* that the answer received at the time the data were retrieved. The explanatory variables are either dummies equal to 1 if the answer has a corresponding number of pieces of code in the text (columns 1–4) or a dummy equal to 1 if the answer is *accepted* (column 5). Columns 1 and 5 report estimates that use the whole sample. These columns show that, overall, answers with more pieces of code and *accepted* answers have a higher score. Columns 2 and 3 run the same specification as column 1 but limit the sample to only English and non-English answers, respectively. Column 4 replicates column 3 but excludes answers with no pieces of code. These estimates indicate that while the positive correlation between the number of pieces of code and the score is confirmed in the sample from the English site, it is not in the sample from the non-English site. This is caused by the answers without code. Indeed, excluding those, the positive correlation is re-established.

One possible explanation for these results is that some questions on non-English sites may not require pieces of code in the answer. In that case, answers receive a positive response from the community even if they do not include code. Section E.3 provides robustness analysis to address this issue.

	(1)	(2)	(3)	(4)	(5)
	All sample	Only English	non-English	non-English exl. zero	Score
Number of pieces of code:					
0	0	0	0		
	(.)	(.)	(.)		
1	0.0731	0.211*	-0.914***	0	
	(0.0816)	(0.0953)	(0.0694)	(.)	
[2,3]	0.537***	0.728***	-0.617***	0.297***	
	(0.0794)	(0.0939)	(0.0632)	(0.0570)	
[4,5]	0.903***	1.169***	-0.386***	0.528***	
	(0.0963)	(0.117)	(0.0695)	(0.0634)	
[6,284]	2.046***	2.441***	0.585***	1.498***	
	(0.0885)	(0.109)	(0.0626)	(0.0565)	
Answer is *accepted*:					
No					0
					(.)
Yes					2.330***
					(0.0548)
Constant	2.401***	2.300***	3.126***	2.213***	2.117***
	(0.0602)	(0.0703)	(0.0512)	(0.0448)	(0.0343)
Observations	323850	266568	57282	49440	323850

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

Notes. OLS estimates at the answer level. The dependent variable is the *score* of the answer (i.e. the number of *likes* received minus the number of *dislikes*). In columns 1–4, the explanatory variables are dummies with a value of 1 if the number of pieces of code appearing in the answer falls in a specific range. Estimates show that answers with more pieces of code tend to have a higher *score*. In column 5, the explanatory variable is a dummy equal to 1 if the answer is *accepted* by the questioner as the solution to the problem. The estimates show that answers *accepted* by the questioner tend to have a higher *score*.

Table 8: Correlation between different proxies of quality

D.3 Construction of variables to capture empathy

***Same-native-language* variable**. For each English answer in the sample, I retrieve the corresponding question and the profile page of the questioner asking that question. I then infer the language of the questioner using two sources: the location and the name displayed.

To infer the language from the location information, I first use the *geopy* Python package to retrieve a standardised format of the location. This process is necessary as users are free to fill the location field in their profile page as they wish. The algorithm attempts to identify the location provided and outputs the name of the country. I then use data from the web to map each country to the main language used.[56]

To infer the language from the name, I proceed as follows. First, I only select users' displayed names that are composed of at least two words, starting with an upper case letter and then at least one lower case letter. I then infer the nationality and the language using the NamePrism API (Ye, Han, Hu, Coskun, Liu, Qin, and Skiena 2017), which was kindly made available to me by Prof. Steven Skiena and co-authors.[57]

For answers published in a non-English language, I assume that the questioner's native language is the one used on the site. I also assume that the native language of the answerers is the language used on the non-English site where they contribute. This assumption does not help to assign the native language to *never treated* users, for whom I am unable to compute this proxy.

For a given answer, the *same-native-language* proxy is then equal to 1 if 1) the answer is on a non-English site, 2) the questioner's country-based language is the same as the answerer's, or 3) the questioner's name-based language is the same as the answerer's. For all other observations, the proxy is equal to 0.

***Manual-profile-picture* variable**. Similarly to the earlier discussion, I retrieve the question and profile page of the questioner for each English answer in the sample. The data include the URL of the profile picture. If the URL includes the word *gravatar*, then the user is using the default avatar provided by the platform. Otherwise, the user is using a personalised profile picture.

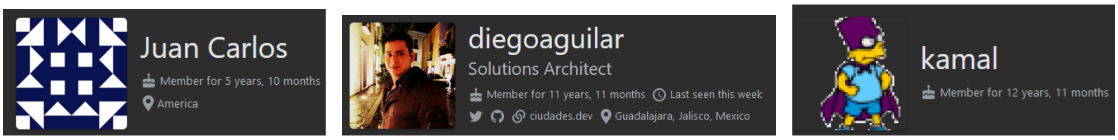
It is plausible to think that, in general, users upload pictures which either represent a photo of the user or an avatar with features representative of the user's identity. Even if this is not the case, the choice of pictures that users upload may still be correlated to their culture. Under this assumption, the presence of a manually uploaded image can increase empathy in the answerer if she relates with it.

For a given answer, the variable *manual-profile-picture* is a dummy that takes a value equal to 1 if the author of the question addressed by the answer displays a personalised picture **AND** if the *same-native-language* variable is equal to 1, and takes a value equal to 0 otherwise.

Examples. Figure 7 provides some examples of user profile page headlines which were extracted from the English site. Consider an answer i written in English by an

[56] The data are retrieved from the website https://www.internetworldstats.com/languages.htm, where the main language is the first one in the list provided for each country.

[57] For more details, see https://name-prism.com/api.



Juan Carlos
Member for 5 years, 10 months
America

diegoaguilar
Solutions Architect
Member for 11 years, 11 months Last seen this week
ciudades.dev Guadalajara, Jalisco, Mexico

kamal
Member for 12 years, 11 months

Notes. The algorithms used in the paper identify the left and centre users as Spanish native speakers using their name and location, respectively. The left user uses a default avatar, while the centre and right users use personalised images.

Figure 7: Examples of user profile page headlines extracted from the English site.

author who is also active on the Spanish site. If i answers a question written by the author on the left or the author on the centre of figure 7, then the variable *same-native-language* would be equal to 1 for answer i. This is because the algorithm has identified those two users as Spanish native speakers from the name and the location, respectively. The variable would be equal to 0 if the question's author were the user on the right. Concerning the value of the *manual-profile-picture* variable for answer i, it is equal to 1 only if the author of the question is the user in the centre. This is because that user is the only user with the same native language and a manually uploaded picture. The user on the left displays the default avatar, while the user on the right displays a manually uploaded image; however, the guessed language is different from the guessed language of the answerer (i.e. the *same-native-language* variable is equal to 0).

D.4 Endogenous selection of question's topics' across sites

To test for the endogenous selection of the questions' topics across sites, I compare the words used in the questions' titles and test whether they are systematically related to a non-English language rather than English. The rationale for using the titles is that, compared to the questions' bodies, they are shorter and less likely to depend on a specific language from a linguistic perspective. This allows for more reliable comparisons across languages. Moreover, titles generally contain critical information about the topic and the kind of question. In addition, this method allows us to identify which words may be site-specific, allowing an ex-post interpretation of what may drive the selection.

The procedure I follow builds on the intuition of the tests for selection into treatment, which are sometimes used in the literature on field experiments.[58] This corresponds to a joint orthogonality test for a set of observable characteristics, where the dependent variable captures the categories in which the selection may occur. In practice, my approach relies on a logistic regression at the title level. The dependent variable is a dummy equal to 1 if the title is from a non-English site and 0 if it is from the English site. By contrast, the explanatory variables are word frequencies for each word appearing across titles. The regression is run separately for each non-English language site. This method is inspired by the natural language processing (NLP) literature on text classification (e.g. Zhang and Oles (2001)).

[58] For instance, see: https://blogs.worldbank.org/impactevaluations/tools-trade-joint-test-orthogonality-when-testing-balance.

For the sake of clarity, I present the detailed procedure using Spanish as an example; the same process has been applied to the other non-English languages.

First, the method pre-processes the data and constructs the regression matrices with the following steps:

1. For all the English and Spanish answers in the sample, retrieve the titles of the corresponding questions.

2. Translate into English all the titles extracted from the Spanish site. To do this, I use Google translate through the Python package *deep_translator*.[59]

3. Parse, separate, and clean the titles' words. This process includes five steps: 1) the selection of only nouns and adjectives via the tokenisation and tagging of the titles; 2) the exclusion of words with less than two characters and more than 15; 3) the transformation of words to lowercase; 4) the exclusion of words commonly considered not meaningful (so-called *stopwords*); and 5) the use of only word roots, which are identified with the Porter stemmer (e.g. *manually* becomes *manual*).[60]

4. Construct a word frequency matrix (explanatory variables) X and the vector of the dependent variable Y. Let i index the questions' titles, N_{eng} be the number of questions from the English site, and N_{spa} be the number of questions from the Spanish site. In addition, let $j = 1, ..., W$ index the unique words extracted from all English and Spanish titles after the pre-processing steps described above. The matrix X has dimensions $(N_{eng} + N_{spa}, W)$ and cell $[i, j]$ reports the frequency of the word j in the title i. The vector Y has dimensions $(N_{eng} + N_{spa}, 1)$ and the element $[i, 1] = 1$ if the title i is from a Spanish question, and 0 otherwise.

With the constructed dataset, the procedure estimates a logit regression model with *l1* regularisation and the liblinear solver.[61] Using the estimated coefficients, the model predicts the probability that a title belongs to the Spanish site. Formally, the predicted probability that a title belongs to the Spanish site is:

$$\hat{p}_i(Y_i = 1|X_i) = \frac{1}{1 + \exp(-X_i\hat{\boldsymbol{\beta}} - \hat{\beta}_0)}.$$

A title i is considered problematic if 1) it appears on the Spanish site and 2) the predicted probability that it belongs to the Spanish site is significantly greater than the prior probability. Formally:

$$i \quad \text{is problematic} \iff (Y_i = 1) \quad \text{and} \quad \hat{p}_i^{lb}(Y_i = 1|X_i) > \frac{N_{spa}}{N_{eng} + N_{spa}},$$

[59] Credit to Nidhal Baccouri: https://deep-translator.readthedocs.io/en/latest/index.html.

[60] The first part is carried out with the Gensim software and the *simple_preprocess* function. The *stopwords* list is retrieved from the NLTK Python package. Stemming is carried out with the NLTK package and the Porter stemming algorithm (https://tartarus.org/martin/PorterStemmer/).

[61] See https://www.csie.ntu.edu.tw/ cjlin/liblinear/.

where $\hat{p}_i^{lb}(Y_i = 1|X_i)$ is the lower bound of a 95% confidence interval around $\hat{p}_i(Y_i = 1|X_i)$ which is computed via bootstrapping. In other words, a Spanish title is problematic if the correct prediction of its site from the logit classifier is significantly better than the prediction from a random classifier.

Table 9 reports the share of questions with problematic titles, conditional on the question's answer(s) being part of the baseline sample. It shows that between 30% and 50% of questions answered in the baseline sample may relate to topics significantly different from those discussed in English. While these numbers are significant, they are upper bounds. Indeed, the prediction exercise that the procedure follows is in-sample. Alternative approaches that test the prediction power of word frequencies out-of-sample would have worse performance. In addition, an assessment of the logit classifier via measures of precision and recall shows a relatively low performance, as shown in figure 8. The dots in figure 8 represent the so-called precision-recall curve for the logit classifier. A performant classifier would produce a curve close to the top-right corner of the graph. By contrast, a random classifier would produce a horizontal curve at the level of the prior probability (the intermittent line in the graphs).[62] From the graphs, it is possible to infer that, while the model predictions are better than those of a random classifier, it is not achieving good performance. The curve becomes substantially comparable to the curve produced by a random classifier when the problematic questions are removed from the sample. The black star and the thick cross identify the values of precision and recall if we set that:

$$\hat{Y}_i = \begin{cases} 1 & \text{if} \quad \hat{p}_i(Y_i = 1|X_i) > \frac{N_{spa}}{N_{eng}+N_{spa}} \\ 0 & \text{otherwise} \end{cases} \quad .$$

It is possible to notice that at that threshold, the precision value is low while the recall is relatively high. This is in part explainable by the fact that $N_{spa} << N_{eng}$, so the denominator of the recall measure is relatively small.

[62] Let a positive prediction be a prediction that a title belongs to the Spanish site. The measure of precision is the ratio between the number of correct positive predictions and the total number of positive predictions. The measure of recall is the ratio between the number of correct positive predictions and the total number of titles belonging to the Spanish site. Since the classifier outputs a probability for the positive prediction, a positive prediction occurs if the predicted probability is higher than an arbitrary threshold. The precision-recall curve computes the measures of precision and recall for a set of threshold values in $(0, 1)$.

	Num. Titles	Share problematic
Language		
Russian	8293	49.52%
Japanese	3053	28.07%
Spanish	12583	39.9%
Portuguese	25555	46.32%

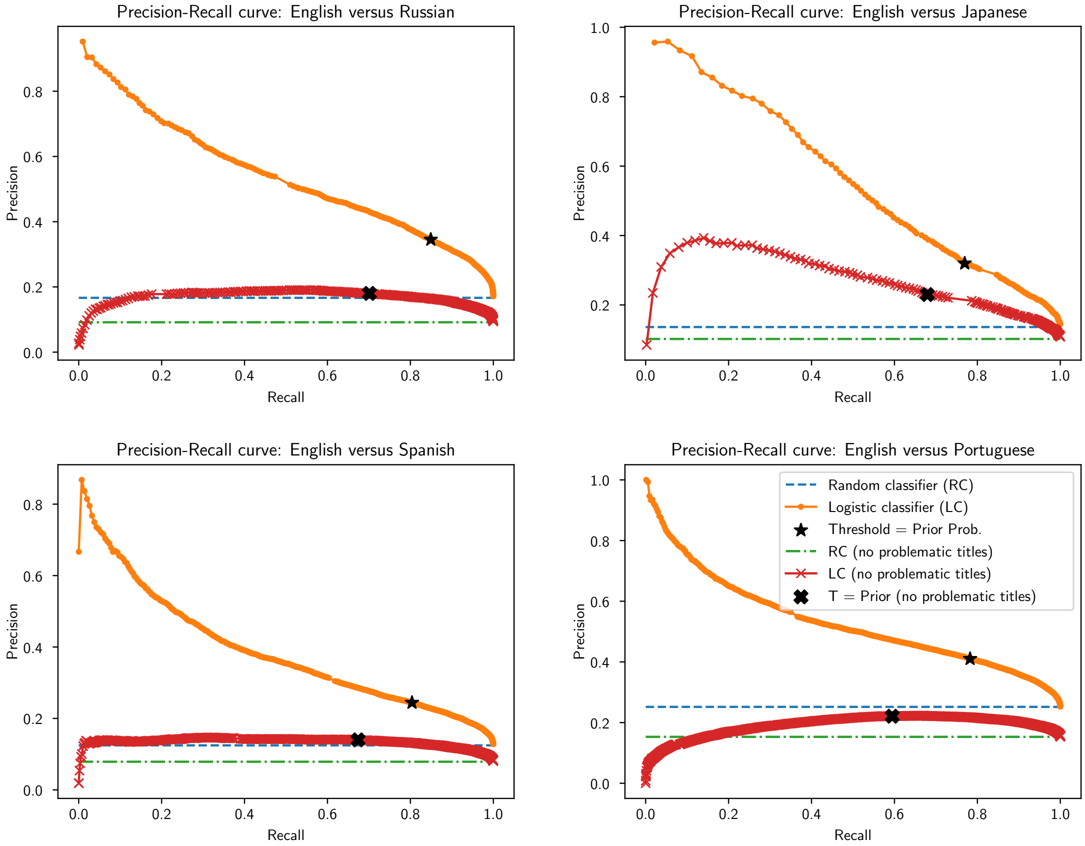
Notes. Share of titles that the logit classifier has correctly predicted to be in the respective language, doing significantly better than a random classifier. The second column reports the number of observations in each language, which corresponds to the number of questions whose answers appear in the baseline regression sample.

Table 9: Share of problematic titles.

Russian	Japanese	Spanish	Portuguese
literatur	rubymin	mercadopago	cpf
uwp	monaca	conda	kotlin
swear	ffmpeg	mercado	monetari
yandex	acquisit	formvalid	nfe
ru	hpack	duda	pagseguro
vk	activerecord	devexpress	bank
phalcon	hoge	androidstudio	made
russian	licens	know	accent
equip	created_at	windowsbuild	demoisel
everyon	casperj	chartj	mandatori
cyril	electron	sii	brazilian
afraid	judgment	navigationview	mount
filestream	bulk	content_main	spservic
memo	countermeasur	lua	sqlsrv
tcpclient	collectionview	exercis	portugues

Notes. List of the 15 words whose frequency vectors have the highest positive coefficient estimates.

Table 10: Most language-specific words.



Notes. Precision-recall curves for the logit classifier and a random classifier for each language. Each point on the curves corresponds to the classifiers' values of recall and precision for different thresholds. The threshold (t) is an arbitrary value for the probability of the title being assigned to the non-English site. It sets the discriminant of the predicted probability $\hat{p}$ such that if $\hat{p}_i > t$, then $\hat{Y}_i = 1$, and 0 otherwise. The curves are computed for the full sample and the restricted sample that excludes problematic titles. A good classifier should produce a concave curve that bends towards the top-right corner. The figure shows that the logit classifier performs poorly: by removing the problematic titles, it is closely comparable to a random classifier.

Figure 8: Precision-recall curves.

D.5 Number of answers per question on the native language sites

Table 11 reports the share of questions published on the native language sites with a certain number of answers and the share and number of answers relevant to those questions. For instance, the first column specifies that 65.77% of questions have only one answer. There are 167,582 of those answers (i.e. answers to questions with only one reply), or 44.1% of the total sample of answers.

Appendix E Robustness for DiD analysis

Num. answers per question	1	2	3	4	5+
Share of questions	65.77%	24.2%	7.01%	2.01%	1.02%
Num. Answers in sample	167582	123338	53592	20460	15049
Share of answers in sample	44.1%	32.46%	14.1%	5.38%	3.96%

Notes. Distribution of the number of answers per question for the sample of all non-English questions. The first row provides the share of questions with a given number of answers, while the other rows refer to the answers to those questions, that is, the share of the sample concerned.

Table 11: Share of the data by thread length.

E.1 Removing answers written in English after treatment by *treatment group* users

Table 12 reports estimates for the main effect of interest after removing *treated* answers written in English by *treatment group* users.

	(1) TWFE	(2) TWFE 1	(3) TWFE 2	(4) TWFE 3	(5) BJS	(6) BJS 1	(7) BJS 2	(8) BJS 3
after	1.231^{**}	1.248^{**}	1.254^{**}	1.340^{**}	1.371^{***}	1.411^{***}	1.412^{***}	1.547^{***}
	(0.176)	(0.157)	(0.159)	(0.143)	(0.0334)	(0.0313)	(0.0288)	(0.0167)
Observations	223313	222815	222815	185766	223313	222805	222805	151997
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QQuality	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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Baseline regression estimates where the dependent variable is the number of pieces of code. The sample exclude *treated* English answers. The estimates correspond to the average treatment effect and correspond to the parameters $\hat{\beta}$ or $\hat{\tau}$ when the specification adopted is the TWFE or the BJS, respectively. The standard errors are clustered (*cse*) at the native language level.

Table 12: *Old joiners*' treatment effect, excluding English answers post-treatment.

E.2 Removing answers on site-specific topics

Section D.4 shows a procedure to test whether the answers' topics are more likely to appear in a particular language. The procedure estimates the probability that a question belongs to a non-English site using word frequency vectors, where the words are extracted from the questions' titles. The procedure labels a question as *problematic* if the predicted probability that the question correctly belongs to a non-English site is larger than the probability arising from a random prediction. Answers responding to non-*problematic* questions address topics that, according to the test, are not specifically prevalent in non-English languages.

To ensure that language-specific topics are not driving the results, I estimate the baseline regression from a sample that excludes answers to *problematic* questions. After imposing the sample restriction, the sample is adjusted to ensure that the users in the remaining sample are active both before and after treatment. Table 13 reports the results.

	(1) TWFE	(2) TWFE 1	(3) TWFE 2	(4) TWFE 3	(5) BJS	(6) BJS 1	(7) BJS 2	(8) BJS 3
after	0.382^{**}	0.379^{**}	0.380^{**}	0.242^{*}	0.453^{***}	0.465^{***}	0.471^{***}	0.438^{***}
	(0.0654)	(0.0700)	(0.0702)	(0.0543)	(0.0502)	(0.0480)	(0.0453)	(0.0806)
Observations	279181	278402	278402	241353	279181	278333	278333	179516
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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Baseline regression estimates where the dependent variable is the number of pieces of code. The sample exclude answers that respond to *problematic* questions. The estimates correspond to the average treatment effect and correspond to the parameters $\hat{\beta}$ or $\hat{\tau}$ when the specification adopted is the TWFE or the BJS, respectively. The standard errors are clustered (*cse*) at the native language level.

Table 13: *Old joiners*' treatment effect, excluding language-specific threads.

E.3 Removing answers with zero pieces of code

Table 14 reports regression results comparable to the estimation reported in table 3 after dropping all answers with zero pieces of code and selecting users that, given the remaining answers, were 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	0.421^{*}	0.415^{*}	0.415^{*}	0.222	0.670^{***}	0.695^{***}	0.693^{***}	0.773^{***}
	(0.124)	(0.129)	(0.129)	(0.0810)	(0.0351)	(0.0338)	(0.0309)	(0.0521)
Observations	249787	249361	249361	223034	249787	249355	249355	156412
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QQuality	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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. Baseline regression estimates where the dependent variable is the number of pieces of code. The sample excludes answers with zero pieces of code. The estimates correspond to the average treatment effect and correspond to the parameters $\hat{\beta}$ or $\hat{\tau}$ when the specification adopted is the TWFE or the BJS, respectively. The standard errors are clustered (cse) at the native language level.

Table 14: *Old joiners*' treatment effect, excluding answers with zero pieces of code.

E.4 Alternative way to compute the average treatment effect

The baseline estimation based on Borusyak et al. (2022) computes the treatment effect for each *treated* answer and obtains the average treatment on the treated by averaging the answers' treatment effects with uniform weighting. Since the panel is unbalanced, this approach implies that users who contribute more post-treatment have a larger weight on the final average treatment on the treated.

To obtain an estimate that weighs equally all *treatment group* users, it is possible to average the answers' treatment effects first within each author and then across authors. To apply this approach, it is necessary to modify the third step in the estimation process followed by Borusyak et al. (2022), which is discussed in section 7.2.1. Let $j \in J$ index *treatment group* users and t index time. In addition, define I_j as the set of answers published by the user j after treatment (i.e. when j's native language was already available). If $\hat{\tau}_i$ is the treatment effect for answer i, the average treatment effect is:

$$\text{[Step 3]} \quad \hat{\tau} = \frac{1}{J} \sum_{j} \left(\frac{1}{\#I_j} \sum_{i \in I_j} \hat{\tau}_i \right).$$

Table 15 reports the estimates obtained with this approach and using the baseline sample, while table 16 reports the estimates that follow this approach but exclude all *treated* answers in English.[63] It is possible to see that by setting equal weights across users, the effect is much smaller, if significant at all. By focusing only on the non-English answers, the effect is again positive and significant, even if it is smaller.

These results suggest that there is important heterogeneity across users, which is

[63] The baseline estimates that use this latter sample are shown in table 12.

	(1) BJS	(2) BJS 1	(3) BJS 2	(4) BJS 3
after	-0.0110	0.00171	0.0101	0.0792***
	(0.0255)	(0.0247)	(0.0230)	(0.00957)
Observations	323850	322919	322919	204541
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes
Empathy	No	No	No	Yes

Standard errors in parentheses

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

Notes. Point estimates for the treatment effect using the method employed by Borusyak et al. (2022). It differs from the baseline estimation in how it computes the final average treatment effect. Compared to the baseline, this approach weighs the users equally rather than overweighing users with more contributions post-treatment.

Table 15: *Old joiners*' treatment effect at the author level.

compatible with the results reported in section G.1. Indeed, these results suggest that the treatment effect is larger for users contributing more to the native language site.

	(1) BJS	(2) BJS 1	(3) BJS 2	(4) BJS 3
after	0.212^{***}	0.227^{***}	0.243^{***}	0.298^{***}
	(0.0253)	(0.0242)	(0.0211)	(0.0234)
Observations	223313	222805	222805	151997
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	No	Yes	Yes	Yes
Competition	No	No	Yes	Yes
Empathy	No	No	No	Yes

Standard errors in parentheses

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

Notes. Point estimates for the treatment effect using the method employed by Borusyak et al. (2022) and excluding answers written in English post-treatment. Compared to table 12, this approach weighs the users equally rather than overweighing users with more contributions post-treatment.

Table 16: *Old joiners*' treatment effect at the author level, excluding English answers post-treatment.

E.5 The role of zero-snippets-of code questions for the complementarity of the effect with questions' quality

Section 7.4.1 tests the hypothesis for which the main effect increases when the question quality is higher. It discusses the possibility that questions with no pieces of code require answers that do not necessarily include pieces of code, and vice-versa. This section provides supporting evidence that the treatment effect increases with higher question quality by addressing this specific issue. First, it shows that by removing questions with no pieces of code, the treatment effect is substantially heterogeneous across levels of question quality. Second, it shows that the probability that the answer is accepted increases with the question's quality.

A possible way to address intrinsic differences between questions with and without code is to remove those without code from the analysis. Table 18 reports estimates for the same models used in section 7.4.1. The sample is different as I remove all answers that address a question with zero pieces of code and ensure that the remaining answerers were active both before and after treatment. Table 17 reports the new quality levels for the question. The results show that answer quality increases by 16.8% when the question has only one piece of code and by 32.2% when the question has four or more pieces of code.

A second approach is to use a measure of the answers' quality which does not rely on the amount of code. Table 19 reports estimates for the same specifications as in section 7.4.1 but using as dependent variable a dummy equal to 1 if the answer is *accepted* as the solution by the questioner. The table shows that the probability that the questioner *accepts* the answer increases by 13.8% when the question's quality is low and by 22.6%

when the question's quality is high.

	Number of snippets of code in the question
Low	{1}
MediumLow	2
MediumHigh	3
High	(3,111]

Table 17: Categories for the quality level of the question.

Appendix F Robustness for the empirical analysis of the study of selection on expertise

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.224	0.0644	0.378***	0.460***
	(0.153)	(0.0957)	(0.0414)	(0.0618)
MediumLow × after	0.403*	0.246*	0.589***	0.662***
	(0.114)	(0.0703)	(0.0289)	(0.0452)
MediumHigh × after	0.445*	0.283*	0.658***	0.757***
	(0.111)	(0.0576)	(0.0318)	(0.0497)
High × after	0.576**	0.393**	0.893***	0.883***
	(0.0759)	(0.0323)	(0.0375)	(0.0420)
Observations	245367	218173	245302	152708
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

Standard errors in parentheses

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

Notes. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 18: Estimates by question quality level after dropping observations with zero snippets of code in the question.

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.00877	-0.00332	0.0764***	0.0497***
	(0.00325)	(0.00278)	(0.00527)	(0.00932)
MediumLow × after	0.0269**	0.0151*	0.0967***	0.0706***
	(0.00456)	(0.00380)	(0.00790)	(0.00485)
MediumHigh × after	0.0307***	0.0189*	0.101***	0.0629***
	(0.00323)	(0.00486)	(0.00650)	(0.0122)
High × after	0.0353**	0.0214	0.109***	0.0813***
	(0.00766)	(0.00857)	(0.00554)	(0.00529)
Observations	322992	285943	322919	204541
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

Standard errors in parentheses

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

Notes. The dependent variable is a dummy equal to 1 if the answer is *accepted* as the solution by the questioner. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 19: Estimates by question quality level.

F.1 Alternative quality measures to study selection on expertise

Table 20 reports estimate results for the following model (discussed in section 8.2.1):

$$Quality_i = \alpha_{q(i)} + \delta_{r(i)} + \sum_{g \in G} \beta_g D_{g(j(i))} + \zeta t_{j(i),i} + \varepsilon_i.$$

The dependent variable is a measure of quality for answer i. Different measures use different proxies, including the number of pieces of code included in the answer, a dummy equal to 1 if the answer is *accepted* as the solution to the question, and the number of upvotes received, net of the downvotes. The explanatory variables are, from left to right in the specification, a question fixed effect, an order-of-publication fixed effect, the fixed effect of the authors' group based on their participation on the English site, and the number of days between the author's registration and the publication date.

The results show that, on average, *new joiners* write answers of lower quality than *old joiners*, which is consistent across different quality measures.

Appendix G Additional results

	(1)	(2)	(3)	(4)	(5)
	numCodes	STDnumCodes	Score	STDScore	IsAcceptedAnswer
Registered After Active	-0.932***	-0.218***	-0.832***	-0.244***	-0.0684***
	(0.0497)	(0.0116)	(0.0380)	(0.0111)	(0.00629)
Registered Before Always Active	0	0	0	0	0
	(.)	(.)	(.)	(.)	(.)
Registered Before Active After	-0.647***	-0.151***	-0.495***	-0.145***	-0.0606***
	(0.0622)	(0.0145)	(0.0476)	(0.0139)	(0.00787)
Registered Before Active Before	-1.051***	-0.246***	-0.810***	-0.237***	-0.0687***
	(0.107)	(0.0249)	(0.0816)	(0.0239)	(0.0135)
Registered After Not Active	-1.426***	-0.333***	-1.098***	-0.322***	-0.122***
	(0.0546)	(0.0128)	(0.0418)	(0.0122)	(0.00692)
Registered Before Not Active	-1.856***	-0.433***	-1.197***	-0.351***	-0.147***
	(0.0873)	(0.0204)	(0.0668)	(0.0196)	(0.0110)
Not Registered	-1.707***	-0.399***	-1.274***	-0.373***	-0.144***
	(0.0601)	(0.0140)	(0.0459)	(0.0135)	(0.00760)
DaysFromOldestReg	0.000483***	0.000113***	0.000321***	0.0000939***	0.0000506***
	(0.0000263)	(0.00000615)	(0.0000201)	(0.00000590)	(0.00000333)
Observations	189963	189963	189963	189963	189963
Question Fixed Effects	Yes		Yes		Yes
Order-of-Publication Fixed Effects	Yes		Yes		Yes
Standardized Dep. Var.	No	Yes	No	Yes	No

Standard errors in parentheses

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

Notes. Regression estimates with different proxies for answer quality. The dependent variable in the specification of column 1 is the number of pieces of code in the answer; in column 2, it is the same variable as in column 1 but standardised. In column 3, it is the number of upvotes minus the downvotes received by the answer; in column 4, it is the same as in column 3 but standardised. In column 5, it is a dummy equal to 1 if the answer has been *accepted*, and 0 otherwise.

Table 20: Difference in answer quality between *old joiners* and *new joiners*.

G.1 Heterogeneity on *old joiners*' treatment effect

Section 7 identified a beneficial effect of introducing multiple languages for *old joiners*. From a managerial perspective, it is relevant to understand whether specific groups of users may be driving the treatment effect.

Proposition 1, which provides the theoretical rationale for the effect, states that the cost of using English is a critical dimension of heterogeneity. Indeed, even if users have the same native language, their cost of using English can differ. By assuming that users have the same cost of using the native language, users with a higher cost of using English face a larger drop in language cost once they use the native language. The comparative statics of the model suggest that the increase in answer quality should be larger for these users.

The researcher does not observe a precise empirical measure for the cost of using English. Nevertheless, this section aims to provide insights into heterogeneity by looking at dimensions that may correlate with it. It investigates heterogeneity based on three considerations: the degree to which the users shifted their contribution to the native language once available, the specific native language, and the intensity of participation in English before the native language became available.

G.1.1 Rate of adoption of the native language

Users differentiate themselves in how they shift contributions to the non-English site post-treatment. They may remain active mainly in English, with few contributions in their native language, or prefer to use their native language and eventually abandon the English site. The users' choice may reflect some extra benefit or cost that they receive by participating on the native language site compared to participating in the English one. While many factors can drive this change in utility, variation in the cost of using English could in part explain differences in behaviour.

For a given user, let κ be the number of answers in the native language over the total number of answers written post-treatment. I group users into four categories by their value of κ, using the 25^{th}, 50^{th}, and 75^{th} quantiles as intermediate boundaries. Table 21 reports each category's resulting range of values. I then implement the same empirical strategy used in section 7.4 to estimate a category-specific treatment effect. Table 22 reports the estimates for the four categories. The parameters correspond to $\{\hat{\beta}_c\}_{\forall c}$ in equation 10 for the TWFE columns and to $\{\hat{\tau}_c\}_{\forall c}$ in equation 11 for the BJS columns. Columns 1–4 report the estimates using the whole sample as described in section 7, while columns 5–8 report the estimates excluding the answers written in English.

Columns 1–4 are more consistent with the definition of treatment used in the baseline regressions. Nevertheless, in this context, these columns are harder to interpret as different categories of users participate on the English site with different intensities. In particular, the users in the lower categories maintain a stronger presence on the English site after their native language becomes available compared to users in the higher categories. This may confound the interpretation of the results as the users in the lower categories may face a smaller reduction in the cost of language because they mostly keep writing in English after treatment, and not because they have a lower cost of using English. The estimates in columns 5–8 correct for this confounding effect by excluding the answers written in English after treatment.

The estimates show that the effect is largely driven by users who switch to the native language website. This result supports the possibility that users who are less proficient in English have a larger benefit from the introduction of their native language.

	Share of answers not in English in the after-period
Low	(0,0.143]
MediumLow	(0.143,0.425]
MediumHigh	(0.425,0.875]
High	(0.875,1]

Table 21: Categories for the rate of adoption of the native language.

	All sample				Sample excludes English Answers Post-Treament			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	TWFE	TWFE 2	BJS	BJS 2	TWFE	TWFE 2	BJS	BJS 2
Low × after	0.0870	0.128	0.145^{**}	0.142	0.483	0.572	0.240^{***}	0.0918
	(0.0972)	(0.104)	(0.0548)	(0.0993)	(0.552)	(0.570)	(0.0498)	(0.0528)
MediumLow × after	0.213	0.105	0.375^{***}	0.200^{**}	1.275^{**}	1.358^{*}	1.224^{***}	0.734^{***}
	(0.108)	(0.107)	(0.0434)	(0.0714)	(0.239)	(0.251)	(0.0345)	(0.0277)
MediumHigh × after	0.648^{*}	0.259	0.469^{***}	0.561^{***}	1.084^{*}	1.168^{*}	0.682^{***}	0.705^{***}
	(0.179)	(0.134)	(0.0451)	(0.0491)	(0.272)	(0.277)	(0.0502)	(0.0456)
High × after	1.462^{***}	0.870^{**}	1.797^{***}	2.088^{***}	1.601^{**}	1.676^{**}	1.859^{***}	2.163^{***}
	(0.143)	(0.147)	(0.0286)	(0.0361)	(0.216)	(0.197)	(0.0279)	(0.0370)
Observations	322992	285943	322919	204541	222815	185766	222805	151997
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QQuality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes	No	Yes	No	Yes

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

Notes. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 22: Estimates of the average treatment effect by the author's share of post-treatment native language answers over the total answers.

G.1.2 Language-specific effects

The treatment effect may differ across different native languages. For instance, the cost of using English may be lower for users who are native to languages closer to English. To estimate language-specific effects, I use the baseline OLS specification from equation

8 but apply it separately for each language. Since each regression is specific to one of the languages, there is only one treatment date and the standard TWFE method can be used more reliably.

Table 23 reports the language-specific estimates for two samples. The first sample (columns 1–4) is the same sample used for the baseline results: it includes all answers published post-treatment by both the *treatment group* and never-treated users. The second set of columns (5–8) reports the treatment effect estimates excluding the answers posted in English by already treated users. While the former sample is more consistent with the previous analysis, the latter sample allows for better comparison across languages since the treatment effect only considers native language answers.

Comparing the two regression sets, the Spanish language stands out, suggesting that Spanish native speakers have different behaviours from the other groups when participating in English post-treatment. The results in columns 5–8 suggest that the effect is slightly larger for Spanish and Portuguese. Since one could think that Spanish and Portuguese are more similar to English, at least to the extent that they have the same alphabet, this result does not support the hypothesis that differences across languages are driven by variations in the cost of using English.[64]

	All sample				Sample excludes English Answers Post-Treament			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Spanish	Portuguese	Russian	Japanese	Spanish	Portuguese	Russian	Japanese
after	0.0831	0.443***	0.298*	0.360***	1.108***	1.177***	0.987***	0.922***
	(0.0799)	(0.0858)	(0.120)	(0.0812)	(0.0898)	(0.0974)	(0.138)	(0.110)
Observations	141917	144402	87589	60231	115766	113326	53116	51754
Controls								
QQuality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Empathy	No	No	No	No	No	No	No	No

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

Table 23: Estimates of the average treatment effect for each native language.

G.1.3 Intensity of participation

When it comes to online communities, a small group of users usually contributes the majority of the content. This trend is also evident on Stack Overflow. Table 24 reports different category levels for the number of answers published by *treatment group* users

[64] Note that a scenario where the cost of using English drives heterogeneity can still be consistent with the results. Let the distribution of the cost of using English across users have large weights on the tails for Russian and Japanese users while being more normally distributed for Spanish and Portuguese users. This suggests that only Russian and Japanese users who are proficient in English participate in pre-treatment as the other users find the cost of using English prohibitively high. However, this pattern does not hold for Spanish and Portuguese users. Consequently, conditionally on the researcher observing their contributions, Russian and Japanese users may experience a smaller reduction in the cost of language once their native language becomes available.

in English before treatment. The quartiles of the distribution over the *treatment group* users define each group. The table shows that, before treatment, 50% of the users contributed five or fewer answers. By contrast, 25% of the users contributed between 21 and 2,848 answers. The cost of using English may be one reason that drives these participation differences.

Intuitively, users participating more in English pre-treatment may be more proficient in English. The empirical predictions would then suggest that the effect is lower for higher categories of participation. Table 25 reports the estimates for each category of contribution intensity. The results suggest that the effect is non-monotonic on the degree answerers were active in English pre-treatment and, overall, do not support the hypothesis.

	Number of answers published in English before treatment
Low	{1,2}
MediumLow	(2,5]
MediumHigh	(5,21]
High	(21,2848]

Table 24: Categories for the number of answers published before treatment.

	(1) TWFE	(2) TWFE 2	(3) BJS	(4) BJS 2
Low × after	0.178	-0.187	0.612^{***}	0.395^{***}
	(0.139)	(0.0844)	(0.0533)	(0.0975)
MediumLow × after	0.352	0.0196	0.831^{***}	0.691^{***}
	(0.206)	(0.198)	(0.0342)	(0.0483)
MediumHigh × after	0.262	0.0123	0.648^{***}	0.605^{***}
	(0.130)	(0.0769)	(0.0220)	(0.0306)
High × after	0.393^{*}	0.253^{*}	0.559^{***}	0.615^{***}
	(0.0980)	(0.0586)	(0.0436)	(0.0716)
Observations	322992	285943	322919	204541
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls				
QQuality	Yes	Yes	Yes	Yes
Competition	Yes	Yes	Yes	Yes
Empathy	No	Yes	No	Yes

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

Notes. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 25: Estimates of average treatment effect by author's intensity of contribution in English before treatment.

G.2 Externalities on the English site

The theoretical framework's assumptions do not allow for quality spillovers across sites. Indeed, the choice of quality on the English site does not depend on the choice of quality on the native language site or any other native language site characteristics. There is no straightforward reason why the existence of a native language site may impact the users' quality choices on the English site. Nevertheless, behavioural factors may create such spillovers if the users consider the native language and English sites complementary or substitutable.

To test for externalities on the English site, I estimate the *old joiners*' treatment effect excluding non-English contributions. In other words, I replicate the baseline analysis in section 7 but use a sample that only includes answers written in English. Table 26 reports the estimates for β in equation 8 in the context of TWFE regression and for τ in equation 9 in the context of BJS estimation. The results suggest that the introduction of multiple languages had positive spillovers on the English website even though these were not significant in the preferred specification. 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	0.0449	0.0512	0.0505	0.0677	0.117^{*}	0.119^{*}	0.125^{*}	0.147
	(0.0506)	(0.0541)	(0.0541)	(0.0804)	(0.0576)	(0.0554)	(0.0530)	(0.0976)
Observations	261771	260950	260950	223901	261771	260887	260887	176268
cse	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang	Nat-lang
Controls								
QQuality	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
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes. The standard errors are clustered (*cse*) at the native language level, i.e. at the treatment level.

Table 26: Estimates of the treatment effect on the quality of English answers.