

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

Jacopo Bregolin

University of Liverpool

August 24th, 2022
EEA-ESEM 2022, Milano



Communication frictions hinder information flows

Communication frictions hinder information flows

- Not-aligned **incentives** between sender and receiver
- **Language barriers**

Communication frictions hinder information flows

- Not-aligned **incentives** between sender and receiver
- **Language barriers**

To what extent the **cost of language** affects **communication effort**?

This paper

Data from **Stack Overflow**:

→ Q&A website about computer programming (100M+ visitors/month)

- **Sender** is user **answering** the question
- **Receiver** is user **asking** the question

This paper

Data from **Stack Overflow**:

→ Q&A website about computer programming (100M+ visitors/month)

- **Sender** is user **answering** the question
- **Receiver** is user **asking** the question
- Sender writes **higher quality answers** if she can use her **native language rather than a foreign?** How **much?**

This paper

Data from **Stack Overflow**:

→ Q&A website about computer programming (100M+ visitors/month)

- **Sender** is user **answering** the question
- **Receiver** is user **asking** the question
- Sender writes **higher quality answers** if she can use her **native language rather than a foreign?** How **much?**
- Do **incentives** matter?

This paper

Data from **Stack Overflow**:

→ Q&A website about computer programming (100M+ visitors/month)

- **Sender** is user **answering** the question
- **Receiver** is user **asking** the question
- Sender writes **higher quality answers** if she can use her **native language rather than a foreign?** How **much?**
- Do **incentives** matter?
- Does the **quality of the question** matter?

This paper

Data from **Stack Overflow**:

→ Q&A website about computer programming (100M+ visitors/month)

- **Sender** is user **answering** the question
- **Receiver** is user **asking** the question
- Sender writes **higher quality answers** if she can use her **native language rather than a foreign?** How **much?**
- Do **incentives** matter?
- Does the **quality of the question** matter?
- Is there **heterogeneity** across users?

Theoretical framework

- Bob needs some **information to take an action** → asks question with effort E_Q
- Alice **internalizes** a share (γ) of **Bob's utility** → answers question with effort E_A

Theoretical framework

- Bob needs some **information to take an action** → asks question with effort E_Q
- Alice **internalizes** a share (γ) of **Bob's utility** → answers question with effort E_A

Sender best-response effort choice:

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

Where:

- k_A and λ_A are Alice's expertise and language cost respectively
- s is precision of prior

After a drop in the language cost ($\Delta\lambda_A < 0$):

- **Effort increases**

$$\Delta R(E_Q) = -\frac{E_Q\sqrt{\gamma}k_A\Delta\lambda_A}{\lambda_A''\lambda_A'(E_Q+s)} > 0 \quad (1)$$

and:

- the effect's size **depends on the size of the change in the cost of language**:

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

- the effect is positive on the **effort made by the questioner**:

$$\frac{\partial\Delta R(E_Q)}{\partial E_Q} = -\frac{\sqrt{\gamma}k_A\lambda_A''\lambda_A'\Delta\lambda_A s}{[\lambda_A''\lambda_A'(E_Q+s)]^2} > 0 \quad \text{if } \Delta\lambda_A < 0 \quad (3)$$

- the effect is positive on the **degree of incentive alignment**:

$$\frac{\partial\Delta R(E_Q)}{\partial\gamma} = -\frac{E_Qk_A\Delta\lambda_A}{2\sqrt{\gamma}\lambda_A''\lambda_A'(E_Q+s)} > 0 \quad \text{if } \Delta\lambda_A < 0 \quad (4)$$

Empirical strategy

Staggered implementation of languages:



Empirical strategy

Staggered implementation of languages:



Treated users: **natives**

Data

All answers of:

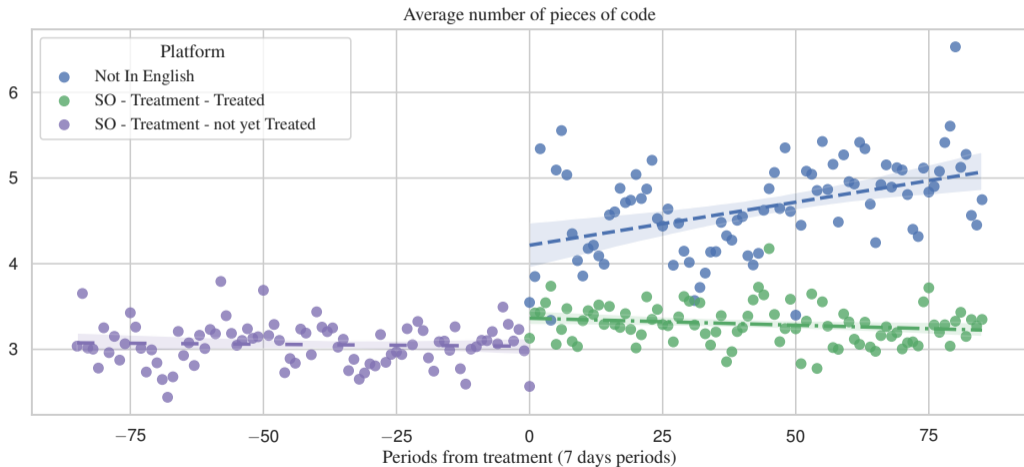
- users participating in both English and non-English languages (**treatment**)
- random sample of users participating only in English (**control**)

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

Variables

- Quality of contributions (**effort**): number of pieces of code in the answer Example
- **Incentives**: amount of auctioned points for answer
- **empathy**: whether questioner speaks the **same language**, questioner's **picture**, questioner has **full name**
- **competition**: number of other answers in same question, number of viewings

Raw data



Effect of a reduction in the cost of language

[Details](#)

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

Standard errors in parentheses

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

[Robustness](#)

Effect is driven by who is "switching" the most

categories: quantiles of $\frac{\# \text{answers no-Eng}}{\# \text{answers after}}$

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

Standard errors in parentheses

Effect increases in questioner's effort

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

Standard errors in parentheses

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

Effect increases in incentives

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

Standard errors in parentheses

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

What trade-off for the platform?

How many languages should **Stack Overflow** have?

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)
- At least 42.8% of non-native English users joined because of the availability of their language (GOOD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)
- At least 42.8% of non-native English users joined because of the availability of their language (GOOD)
- New joiners provide significantly lower quality contributions (BAD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)
- At least 42.8% of non-native English users joined because of the availability of their language (GOOD)
- New joiners provide significantly lower quality contributions (BAD)
- No significant externalities to English website (GOOD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)
- At least 42.8% of non-native English users joined because of the availability of their language (GOOD)
- New joiners provide significantly lower quality contributions (BAD)
- No significant externalities to English website (GOOD)
- Only 11% of programming languages discussed in Stack Overflow are discussed in all websites (BAD)

Implementing multiple languages:

- Quality increases by 24% when writers use their first language (GOOD)
- Answers are 7% more likely to solve the questioner's problem (GOOD)
- At least 42.8% of non-native English users joined because of the availability of their language (GOOD)
- New joiners provide significantly lower quality contributions (BAD)
- No significant externalities to English website (GOOD)
- Only 11% of programming languages discussed in Stack Overflo are discussed in all websites (BAD)
- 33.6% of programming languages discussed in Stack Overflo are discussed in more than one website (BAD)

Conclusion

- **Language barriers** induce substantial **lower quality of communication**
- A **policy** that reduces language barriers is **ineffective if not complemented with incentives and reciprocity**

Conclusion

- **Language barriers** induce substantial **lower quality of communication**
- A **policy** that reduces language barriers is **ineffective if not complemented with incentives and reciprocity**
- A **platform** should implement **additional languages ONLY** if the **community** benefiting is **large enough**

Thank you!

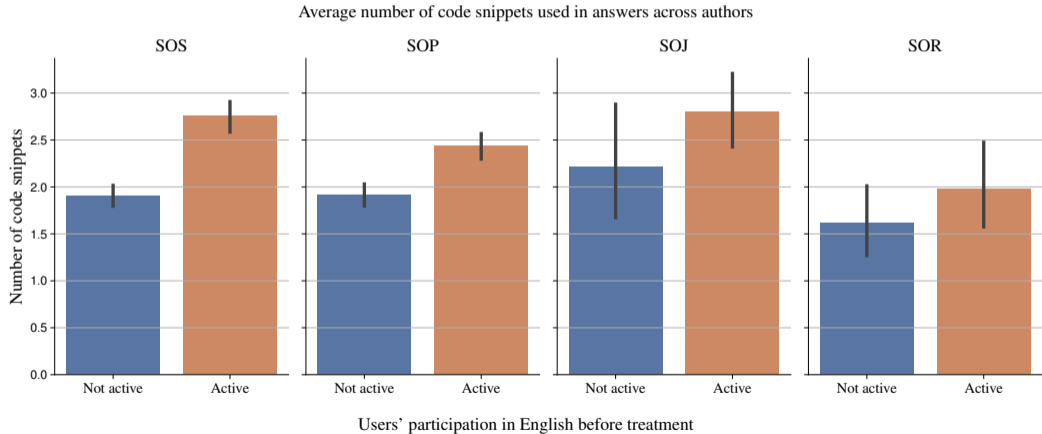
Feedback very welcome: jacopo.bregolin@liverpool.ac.uk

Share of non-native English speakers increases

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

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

New joiners contribute lower quality



Externalities on the English website

Effect on contribution quality in English after treatment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	TWFE	TWFE 1	TWFE 2	TWFE 3	BJS	BJS 1	BJS 2	BJS 3
after × InSo	0.196** (0.0234)	0.186** (0.0347)	0.185** (0.0342)	0.178* (0.0421)	0.203*** (0.0555)	0.209*** (0.0534)	0.216*** (0.0528)	0.203* (0.0948)
Observations	293777	292919	292919	280407	236495	235574	235574	176512
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

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

Inefficiency in aggregating information

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

Table: Number of programming languages for which at least a question has been made in 0, 1, 2, 3, or 4 of the non-English languages. Rows split the sample based on whether the tag appears in the English website (1) or not (0)

Estimation

Let **i** be answers, **j** be users, **t** be weeks.

TWFE:

$$\text{numCodes}_{i(jt)} = \alpha_j + \alpha_t + \beta D_{jt} + \mathbf{W}'_{i(jt)} \boldsymbol{\gamma} + \varepsilon_{i(jt)},$$

Borusyak, Jaravel, Spiess (2021 WP):

[Step 1] $\text{numCodes}_{i(jt)} = \alpha_j + \alpha_t + \mathbf{W}'_{i(jt)} \boldsymbol{\gamma} + \varepsilon_{i(jt)}$ if j not treated at time t ,

[Step 2] $\text{num}\hat{\text{Codes}}_{i(jt)} = \hat{\alpha}_j + \hat{\alpha}_t + \mathbf{W}'_{i(jt)} \hat{\boldsymbol{\gamma}}$ if j treated at time t ,

$\hat{\tau}_{i(jt)} = \text{numCodes}_{i(jt)} - \text{num}\hat{\text{Codes}}_{i(jt)}$ if j treated at time t .

[Step 3] $\hat{\tau} = \frac{1}{N} \sum_{i(jt)|j \text{ treated at time } t} \hat{\tau}_{i(jt)}$.

Estimation: Heterogeneity and 2nd degree effects

Let **c** be some category at either user or answer level.

$$\text{TWFE: } numCodes_{i(jt)} = \alpha_j + \alpha_t + \sum_c \beta_c D_{jt} \mathbf{1}_{c(j)} + \mathbf{W}'_{i(jt)} \boldsymbol{\gamma} + \varepsilon_{i(jt)},$$

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

Robustness: quality as probability that answer is best answer

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

Standard errors in parentheses

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

Back

Example: effort measure

▲ Assumptions:

2



- all rows have the same number of space-delimited 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 ...
    $i=last[i]+0 # replace with the last value seen; "+0" to force undefined to be 0
    last[i]=$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
```

Share Improve this answer Follow

edited 32 mins ago

answered 38 mins ago

 markp-fuso
14k ● 3 ● 11 ● 27

Back