Modelling Segregation Dynamics in Linguistically Diverse Communities

Name: Linhao Chen
Date: September 2nd, 2025
Supervisor: Kenneth Y. Wertheim

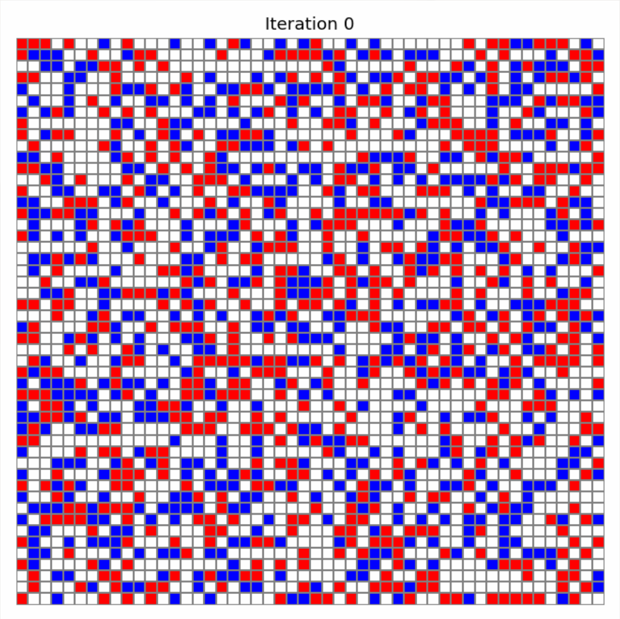
Supported by China Scholarship Council (CSC)

The Schelling Model and segregation dynamics

- **Thomas Schelling's classic segregation model (1971)**

- **Simple rules determine segregation**

    1. Agents make friends with neighbours that belong to same group
    2. When proportion of friends falls below thresholds (e.g., 50%), agent is unhappy
    3. Unhappy agents move
    4. Back to 1.



- **A typical agent-based model.**

- **Useful for revealing emergent phenomena in a social group.**

The Schelling Model and segregation dynamics

- **Limitations of the original Schelling Model:**

  rules only based on binary divisions (e.g., nationality, gender)

  Segregation or mixing have different patterns (more or fewer clusters)

- **Real-life social segregation:**

  involve many factors (economics, cultural, educational…)

  some of them are continuous (e.g., language)

- **Our motivation:**

  extend the model by incorporating linguistic factors to describe

  segregation dynamics in linguistically diverse community (university campus)

Additions made to Schelling's model

Agent attribute: language ability

- **Group1** comprises students or academics from a foreign country
- **Group2** comprises local students and staff members from the host country (UK)
- Normal distribution for each group's language ability

μ_1^{La} (0, 0.5) μ_2^{La} fixed at 0.5

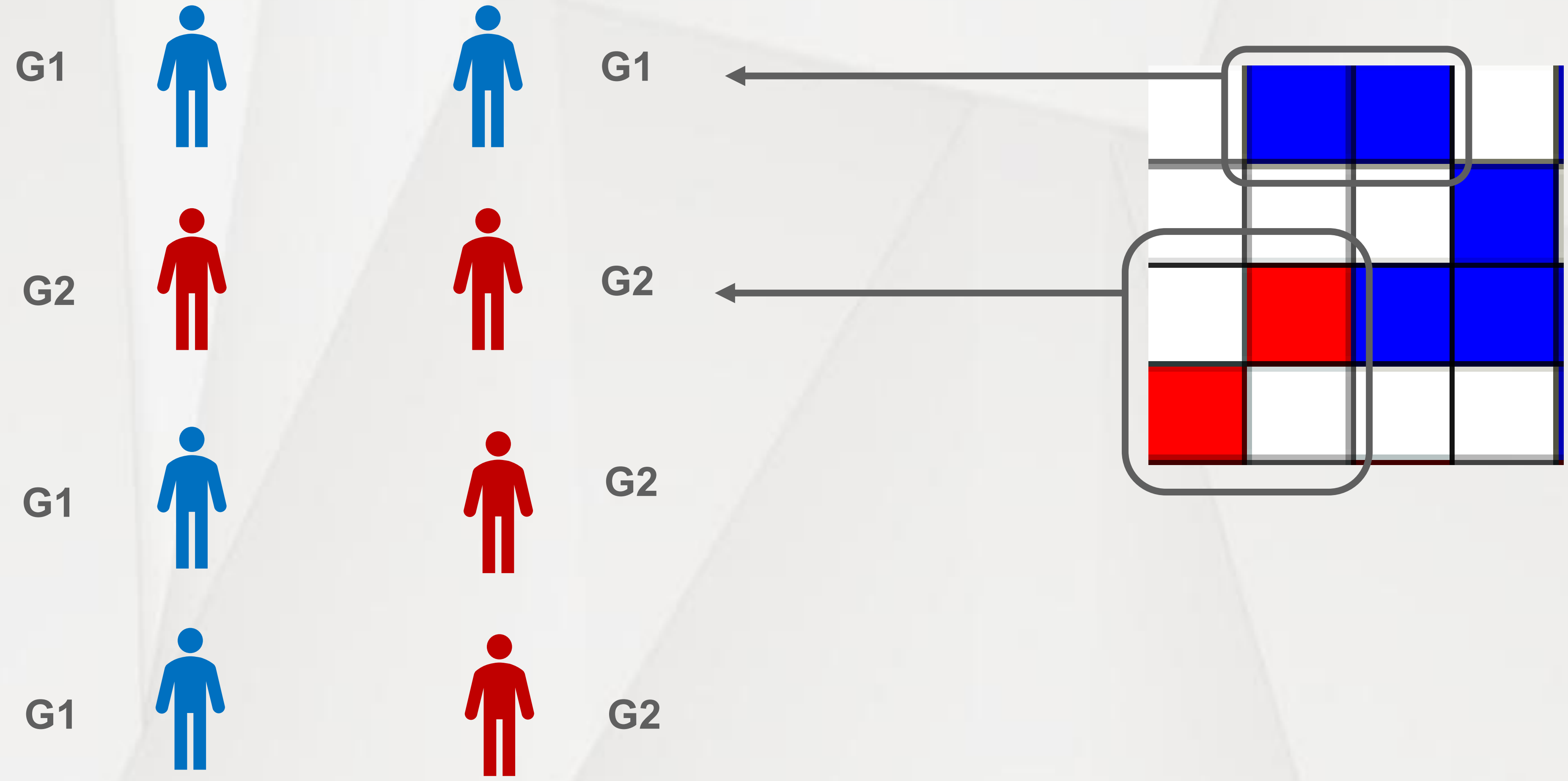
$\sigma_{1\&2}^{La}$ (0.05, 0.25)

IELTS passed,
High school graduated,
Socially awkward,
New to the country
…..

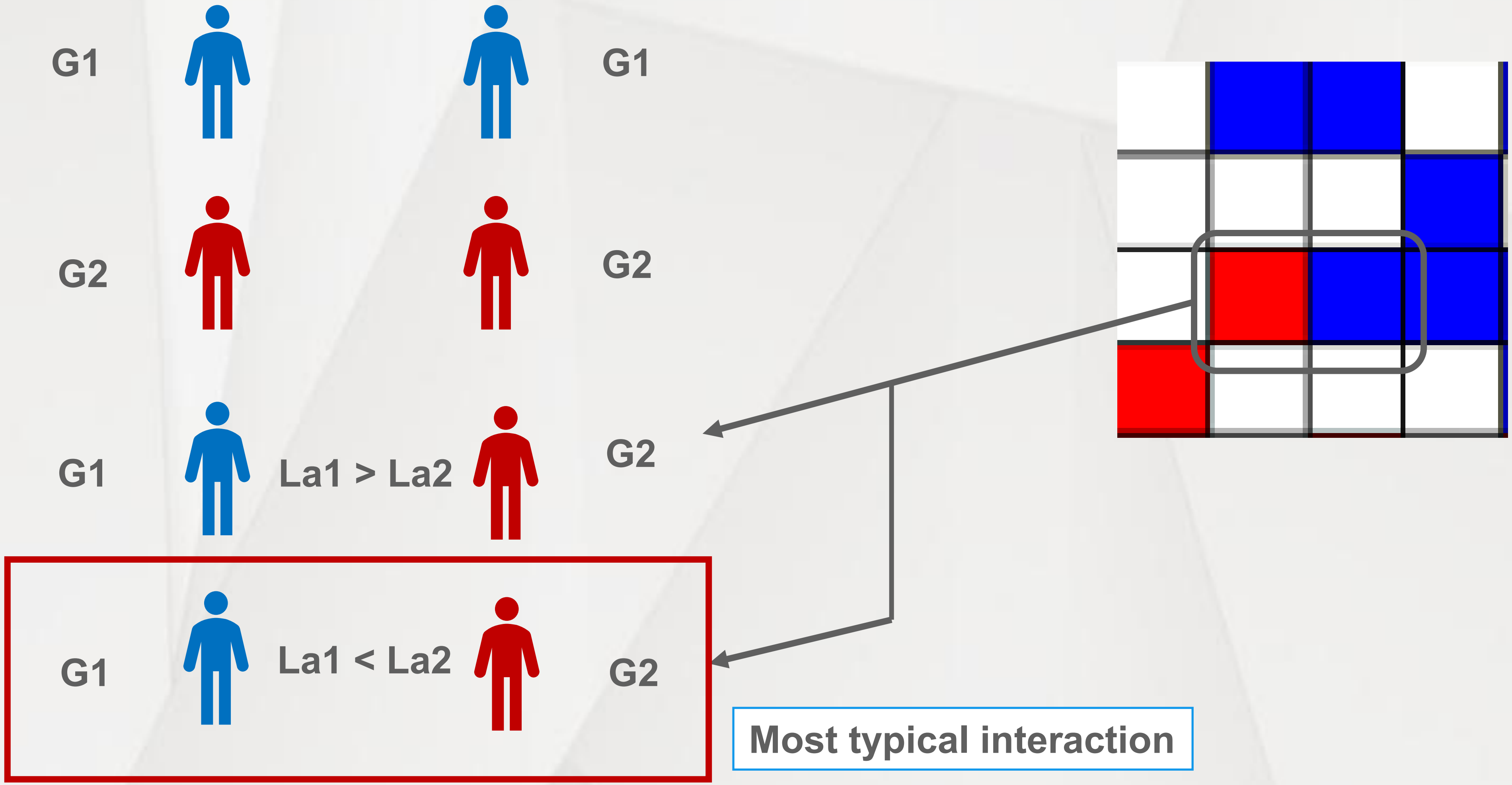
Second-year undergraduate born and raised in the UK.

Language expert (know every word, every grammar rule, and so on),
Every domain of knowledge,
Every social situation,
Familiar with the local culture
…..

Four types of agent interactions



Four types of agent interactions



Agent attribute: language ideology (tolerance)

G1 $La_1 < La_2$ G2

$Lt_{1<2} \neq Lt_{2>1}$

Most typical interaction

Language gap

if

$|La_1 - La_2| < Lt_{1<2}$

and $|La_1 - La_2| < Lt_{2>1}$

then

friends

else

not friends

The other three interactions are governed by similar parameters.

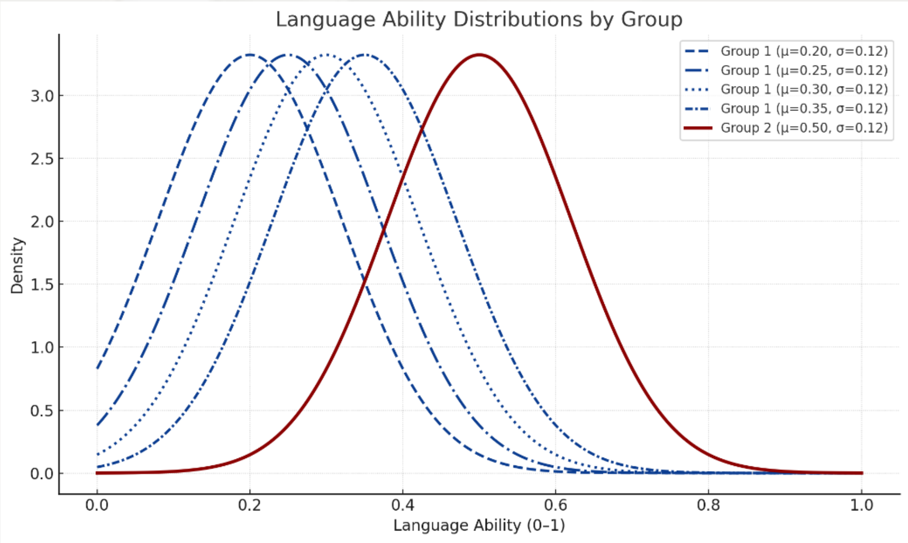
Simulating segregation in artificial societies in four scenarios

Artificial society is a set of 10 probability distributions

Two agent types (G1 and G2).

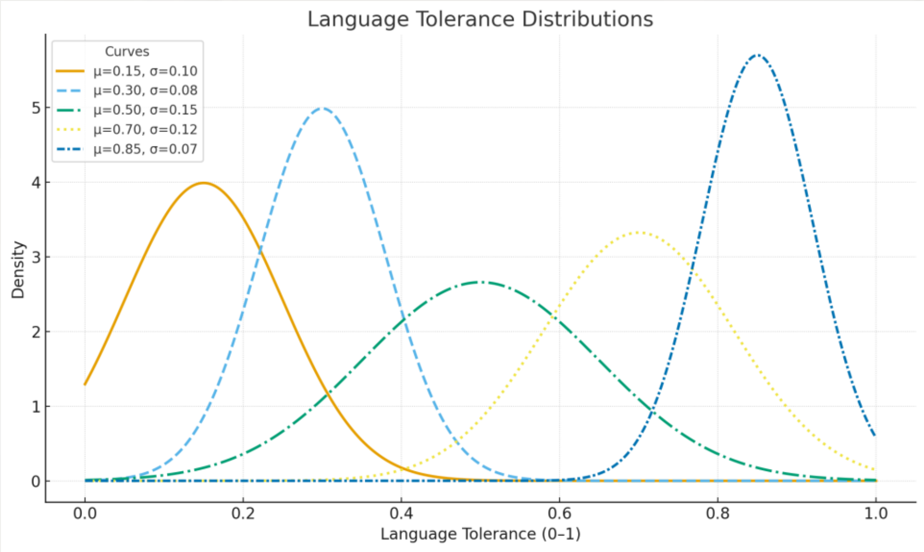
5 attributes per agent type:

- Language ability
- Language tolerance x 4



Therefore, one artificial society:

- 9 means (μ_2^{La} fixed at 0.5)
- 10 standard deviations.



Comparing four scenarios

Schelling's Model

Criterion of friendship:
(1) Same group

One scenario (G).

Modified Model with Linguistic Factors

Criterion of friendship:
(1) Same group
(2) Language gap is smaller than level of tolerance

Three scenarios:

(1) and (2): GL

(2) only: L

(2) with an extra constraint: L^B

Comparing four scenarios

Scenario G

if
    same group
then
    friends
else
    not friends

Scenario GL

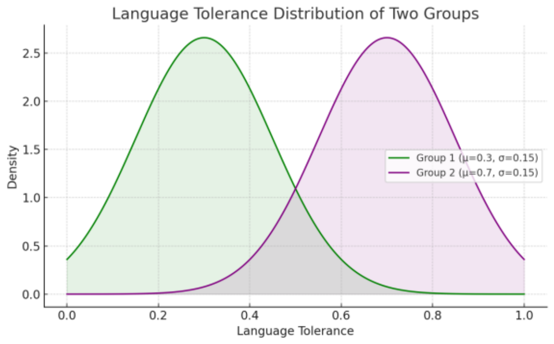
if
    same group
then
    friends
elif
    language gap < language tolerance
then
    friends
else
    not friends

Scenario L

if
    language gap < language tolerance
then
    friends
else
    not friends

Scenario L^B

Extra constraint:



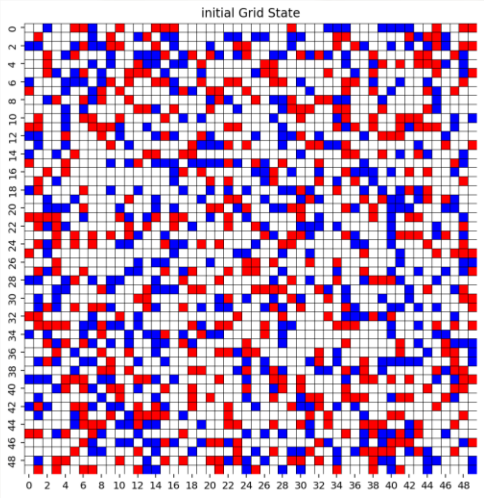
Linguistically biased scenario

Simulation configurations

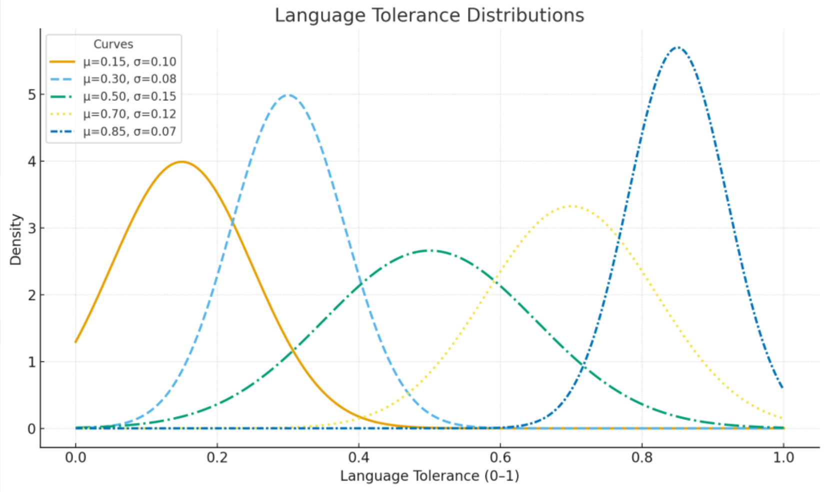
- 50,000 artificial societies for each scenario

- Fixed configurations:

50 rows and 50 columns
Half of the pixels are empty
Minimum fraction of friends is 0.5



- Latin hypercube sampling of 9 means and 10 SDs
- 10 distributions per artificial society.



Simulation algorithm

1. Agents try to make friends. The criterion or criteria depend on the scenario the artificial society is in.

2. Evaluate each agent to determine if they are happy.

3. Move every unhappy agent to a new pixel stochastically.

4. Terminate or back to step 1.

Post-simulation processing

- Three termination conditions

  1. **Equilibrium Condition**:
     The society equilibrates when all agents are satisfied

  2. **Quasi steady state Condition**:
     The society does not change much

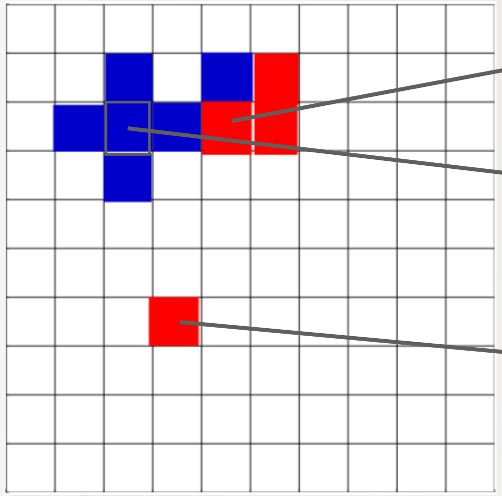
  3. **Maximum Iteration Condition (100 steps)**
     The model would stop after run 100 steps

- 20 runs per artificial society

- Average outputs over 20 runs to obtain ensemble averages for one society

Four outputs for each artificial society

- Segregation (average over agents)



$$\frac{\text{number of same group}\ \ (2)}{\text{number of neighbours}\ (4)}$$

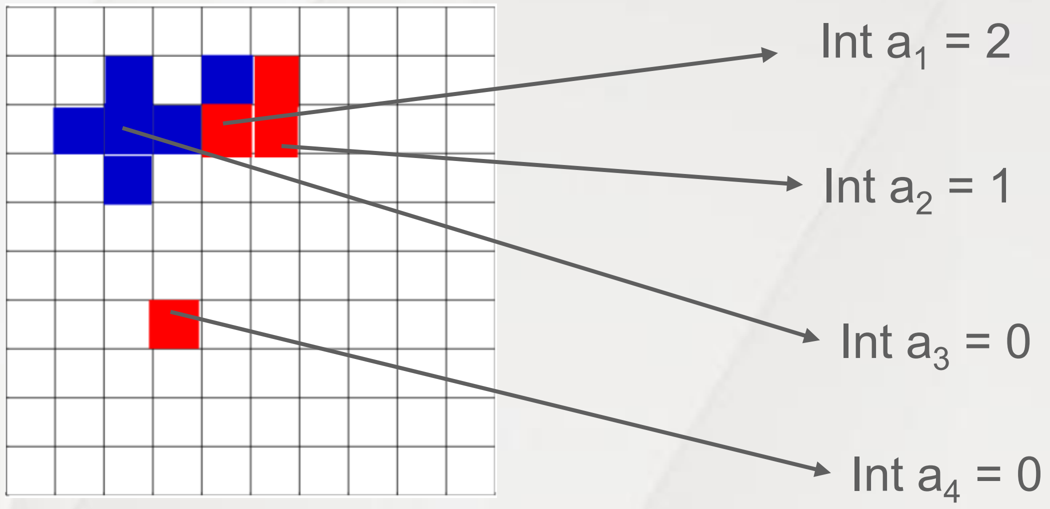
1

0

$$\boxed{\text{Segregation} = \frac{\text{Sum of these fractions}}{\text{Number of agents}}}$$

Four outputs for each artificial society

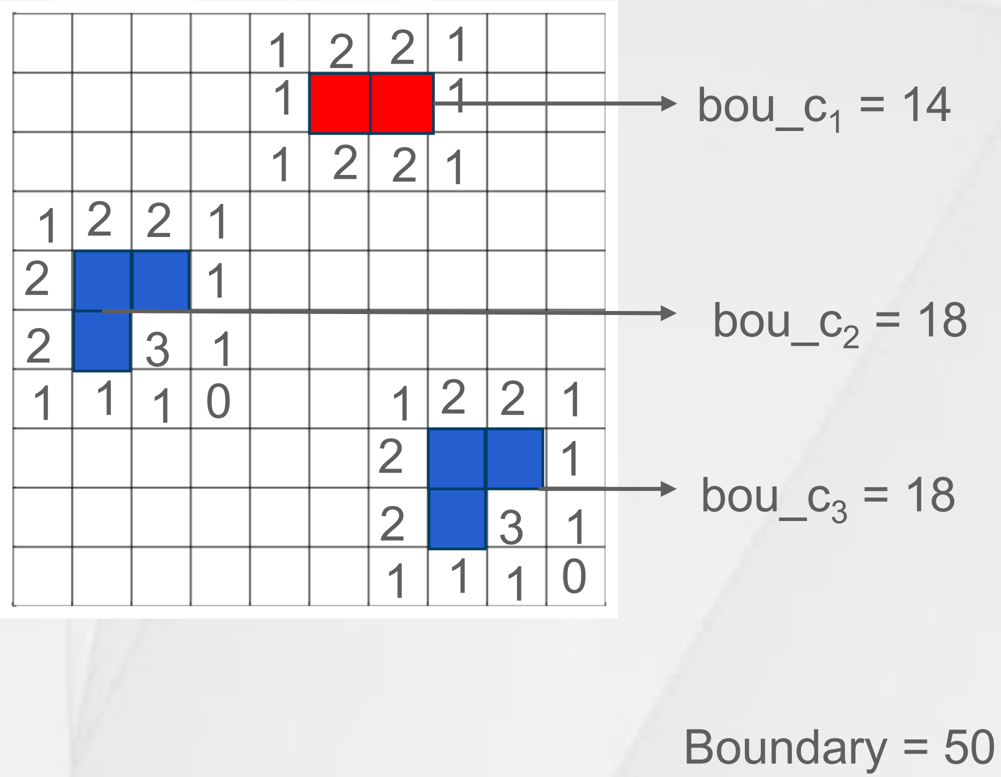
- Interface (sum over agents)

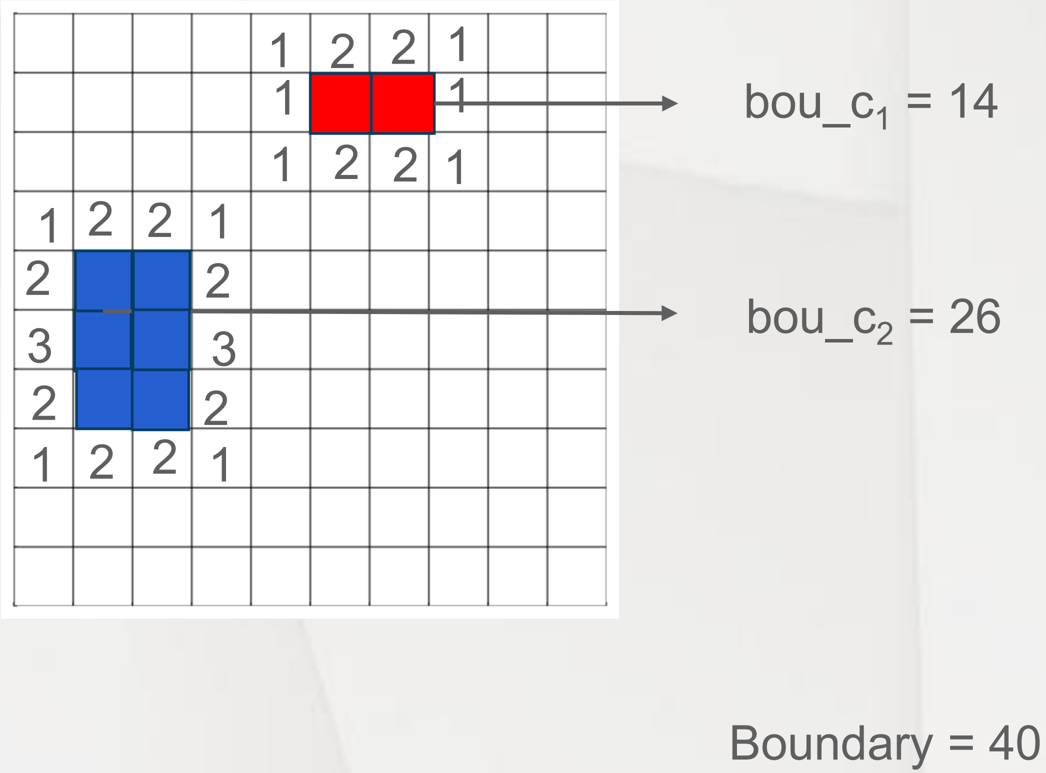


$$\boxed{\text{Interface} = \Sigma \text{ Int } a_n}$$

Four outputs for each artificial society

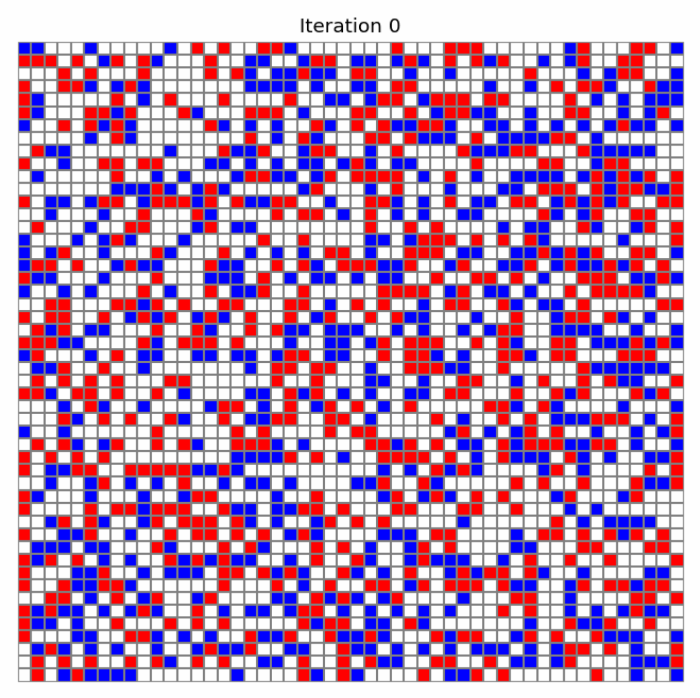
- Boundary (sum over agents) = $\boxed{\Sigma\ bou\_c_n}$





Four outputs for each artificial society

- Iterations till termination (transient period)
- It measures the agents' willingness to meet new people



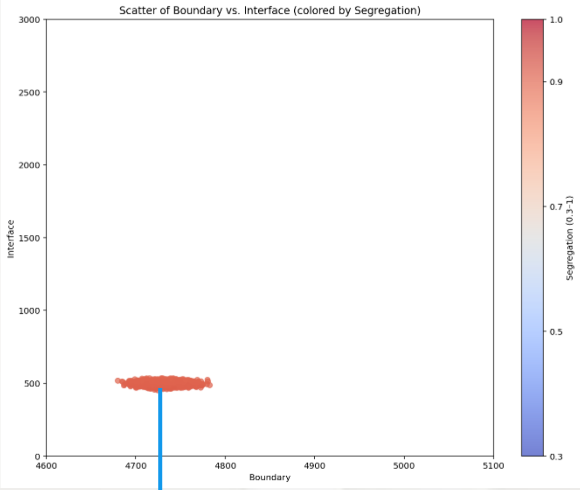
Results and Discussion

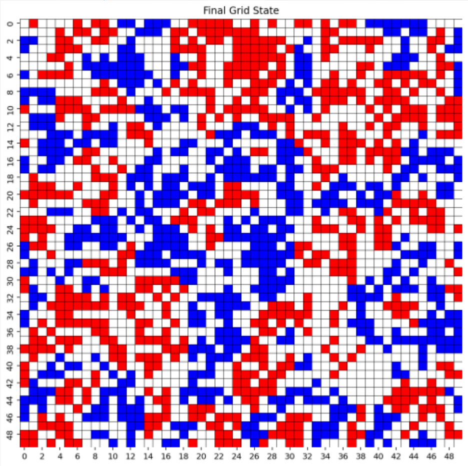
- Relationship between four outputs
- Relationship between four outputs and 19 inputs
- Regime-specific properties

Linguistic factors result in a complicated relationship between mixing and social fragmentation

Relationship between 4 outputs

Scenario G





x-axis: Boundary (4600~5100)

y-axis: Interface (0~3000)

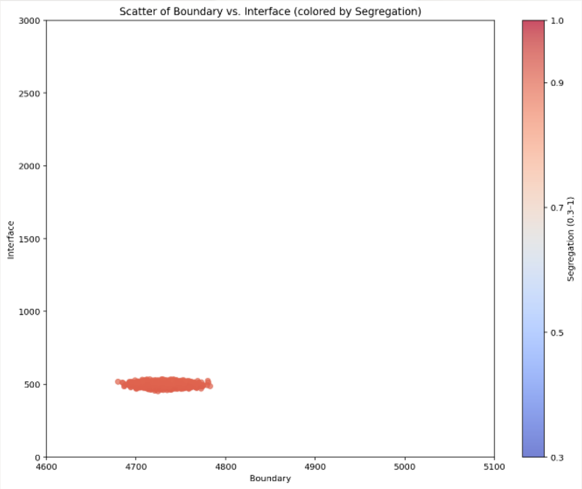
Color: Segregation (0.3~1)

Each dot is an artificial society (ensemble average of 20 runs).

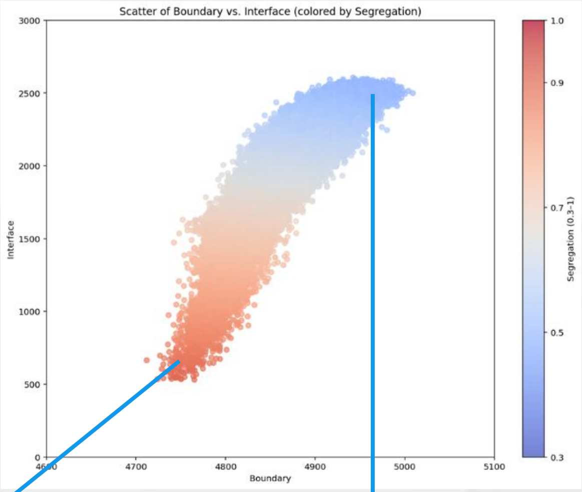
High segregation occurs in relatively big clusters.

Relationship between 4 outputs

Scenario G

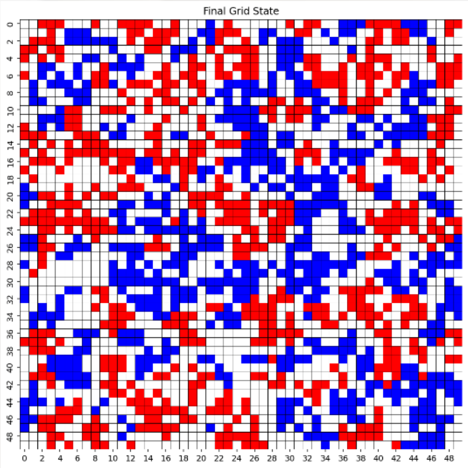


Scenario GL

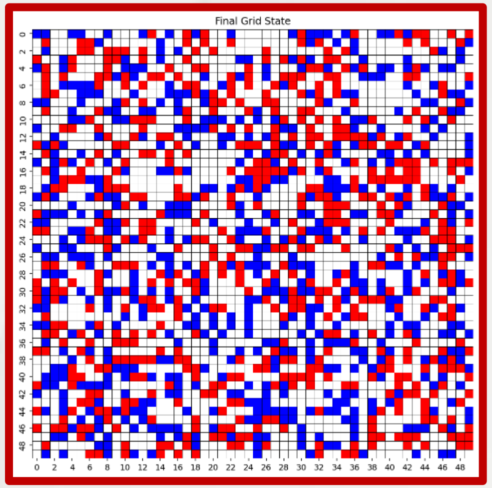


$$boundary \ \alpha \ interface$$

$$segregation \ \alpha \ \frac{1}{interface}$$



seg=0.87 bou= 4700



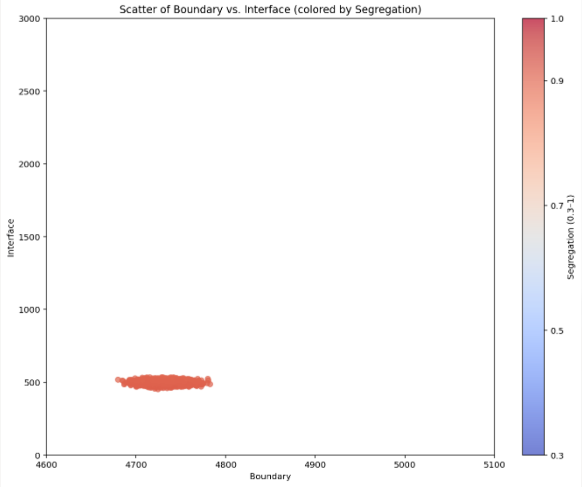
seg=0.53 bou= 4946

When in-group preferences are diluted by linguistic factors, intergroup mixing occurs in many small clusters

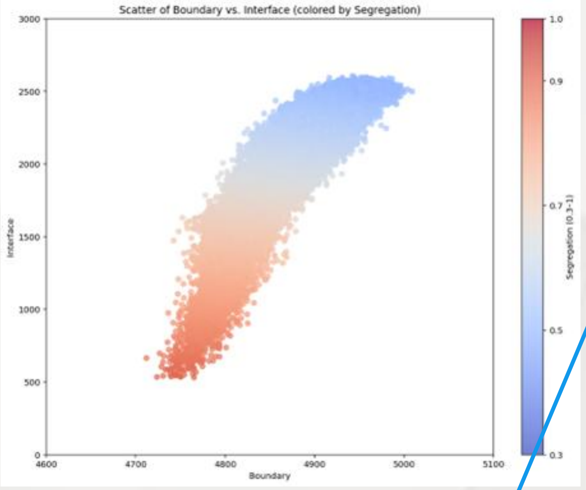
Our first finding!

Relationship between 4 outputs

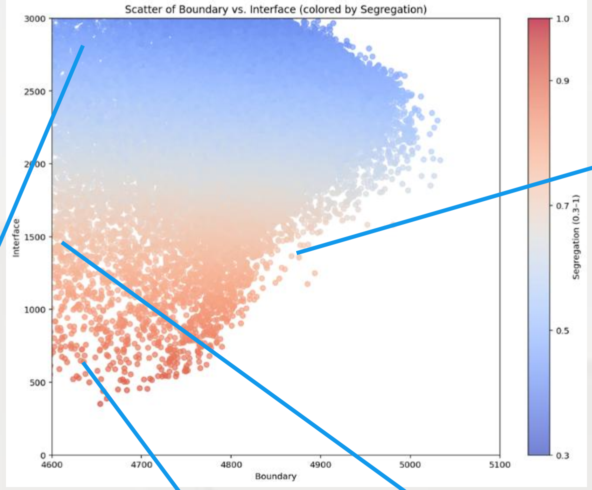
Scenario G

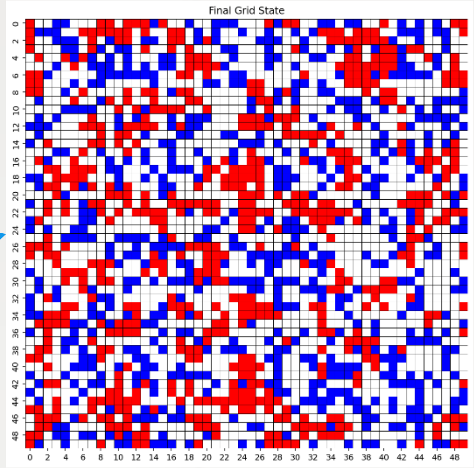


Scenario GL

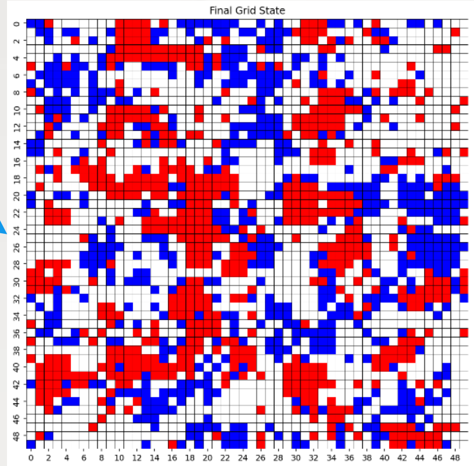


Scenario L



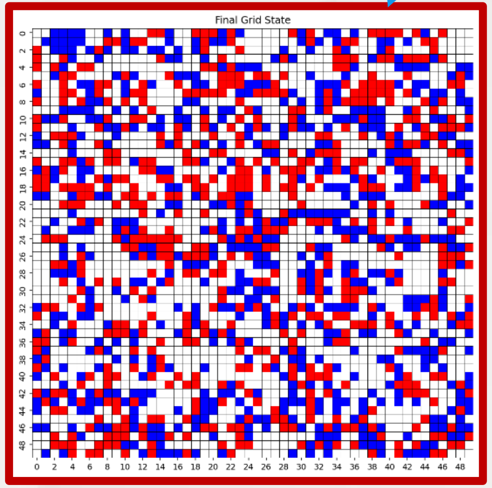


Seg=0.72, bou= 4834

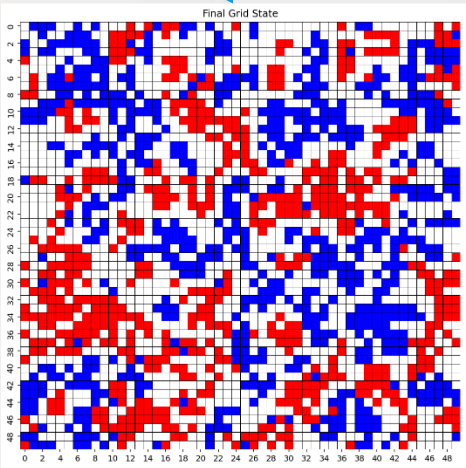


Seg=0.70, bou=3716

When only linguistic factors matter, intergroup mixing could occur in fewer but bigger clusters.



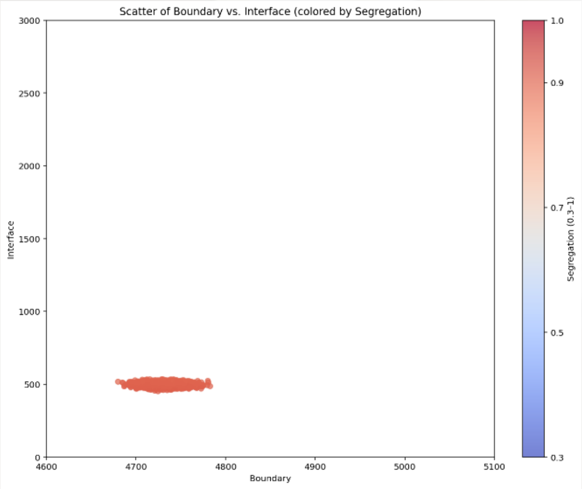
Seg=0.46, bou= 4466



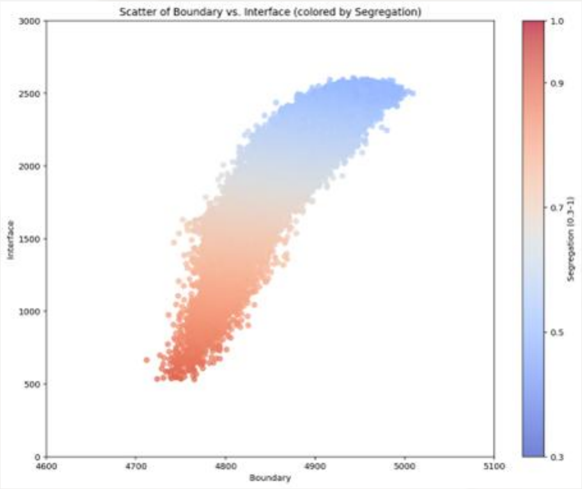
Seg=0.87, bou= 4490

Relationship between 4 outputs

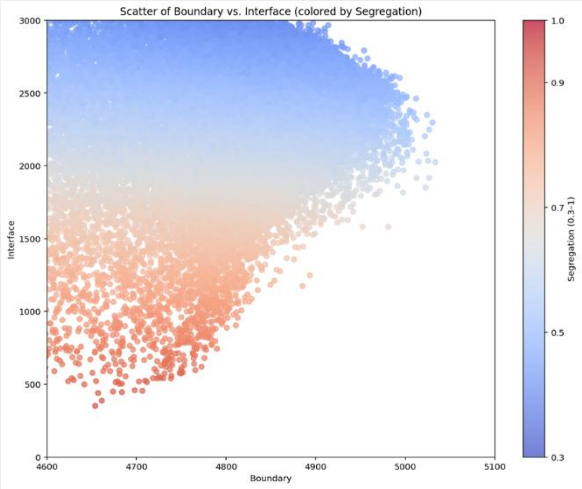
Scenario G



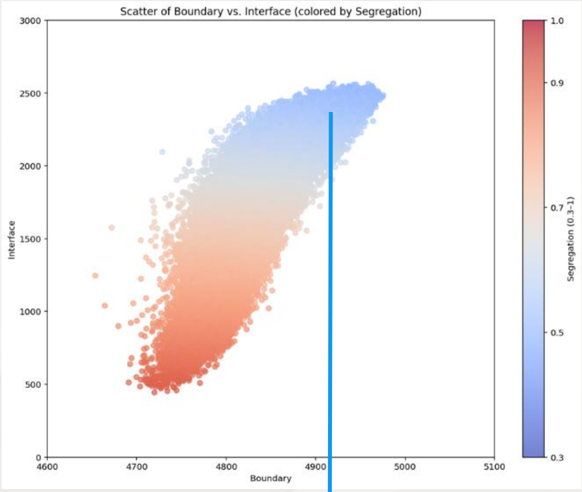
Scenario GL

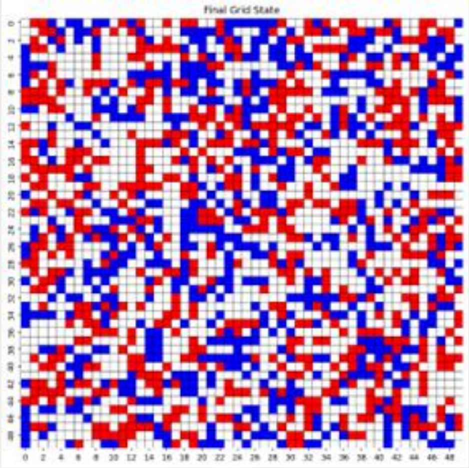


Scenario L



Scenario L^B



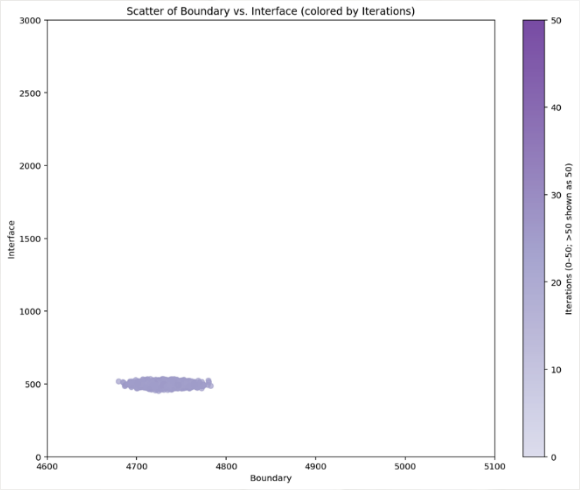


When linguistic factors hide in-group preferences, such preferences still affect intergroup mixing.

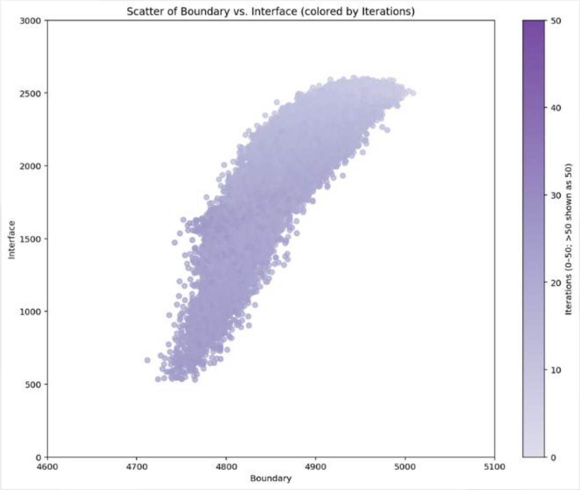
Intergroup mixing occurs in many small clusters again!

Relationship between 4 outputs

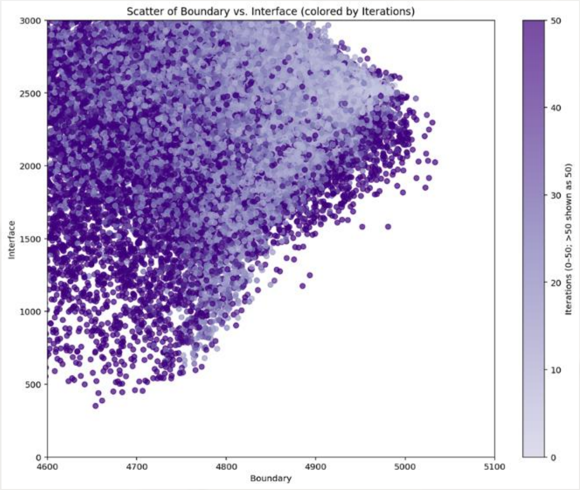
Scenario G



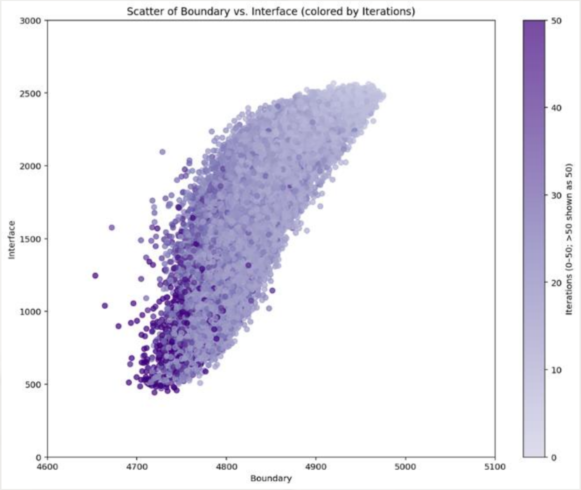
Scenario GL



Scenario L



Scenario L^B



x-axis: Boundary

y-axis: Interface

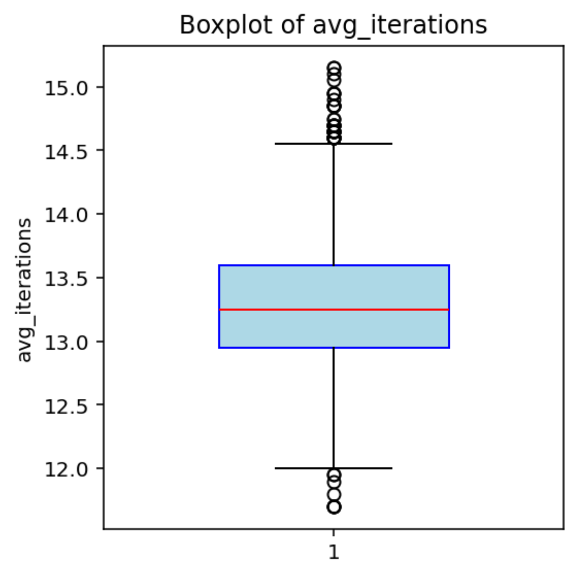
Color: Iterations

Each dot is an artificial society (ensemble average of 20 runs).

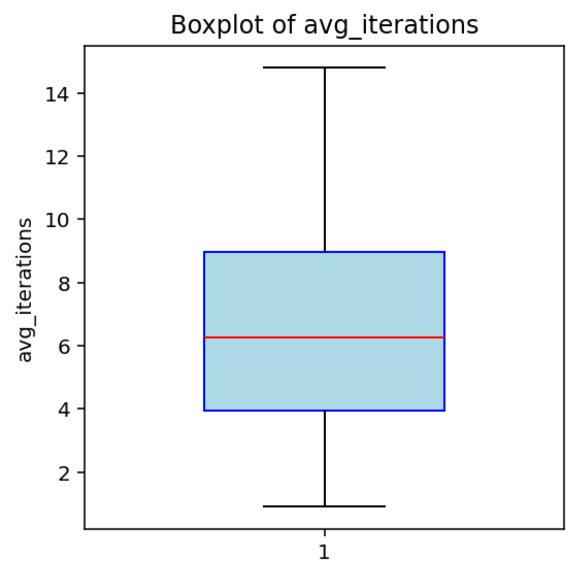
However, when in-group preferences are hidden and not explicit, the transient dynamics are different.

Relationship between 4 outputs

Scenario G

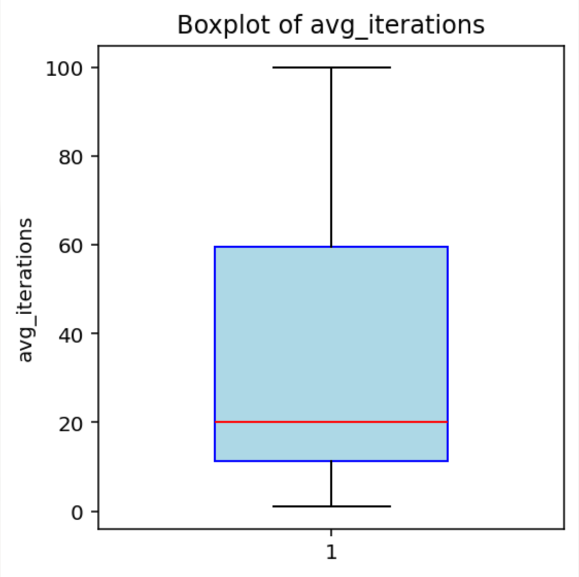


Scenario GL

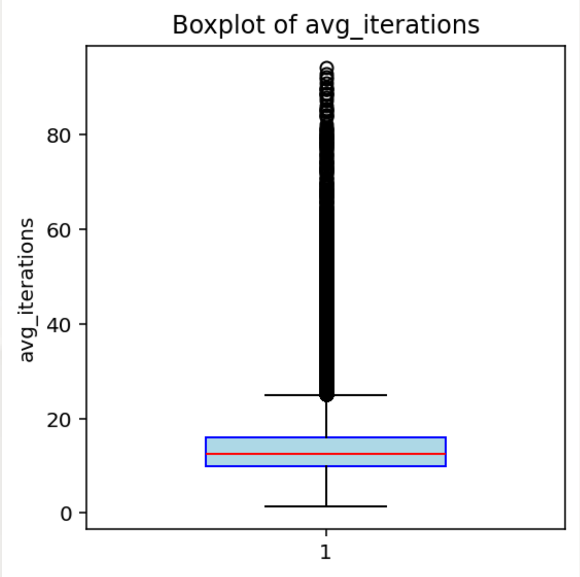


On average, six iterations.

Scenario L



Scenario L^B



On average, 20 iterations.

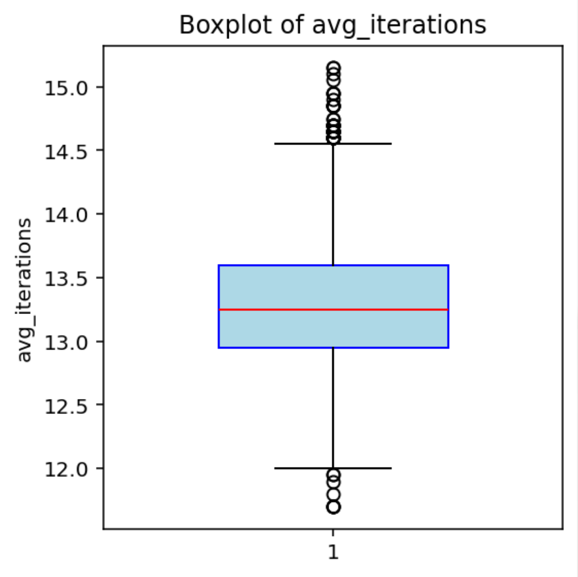
Our second finding!

When only linguistic factors matter, agents are more willing to meet new people before mixing or segregating.

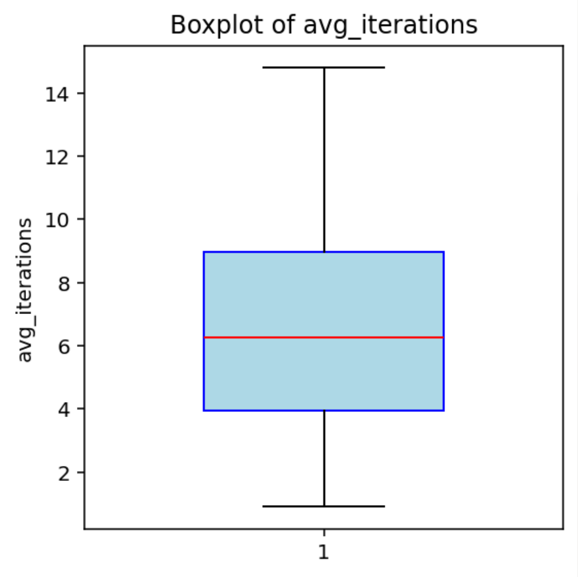
At least, they try despite their hidden in-group preferences!

Relationship between 4 outputs

Scenario G

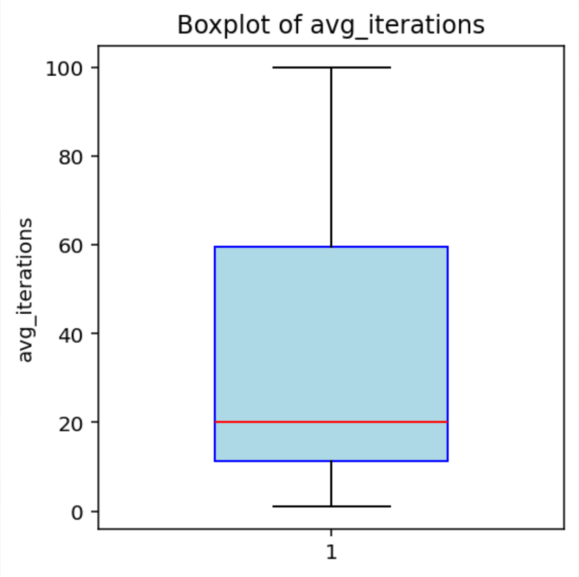


Scenario GL

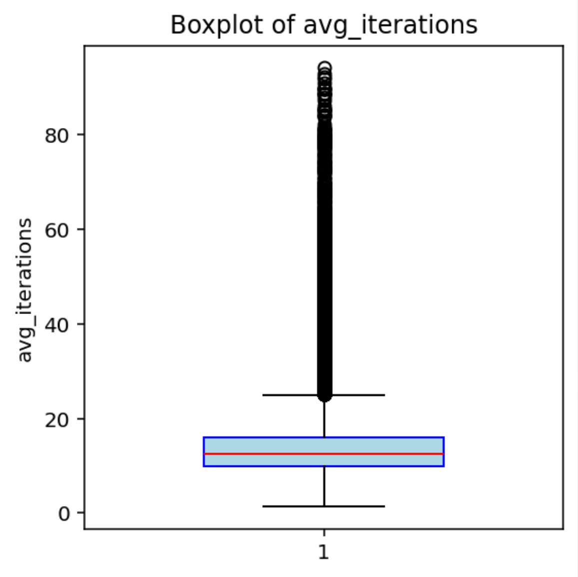


On average, six iterations.

Scenario L



Scenario L^B



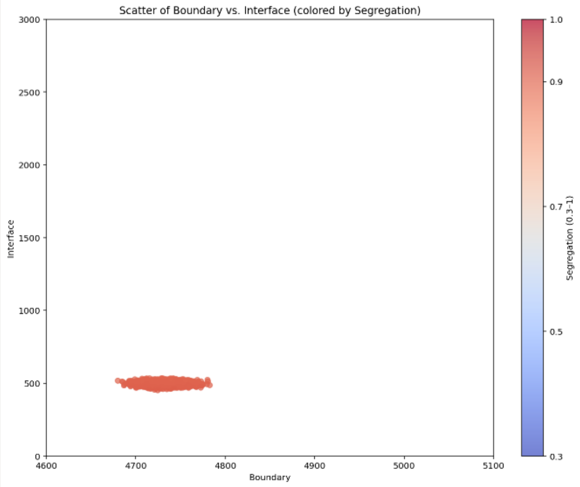
On average, 20 iterations.

GL: Two ways to become friends – group and language

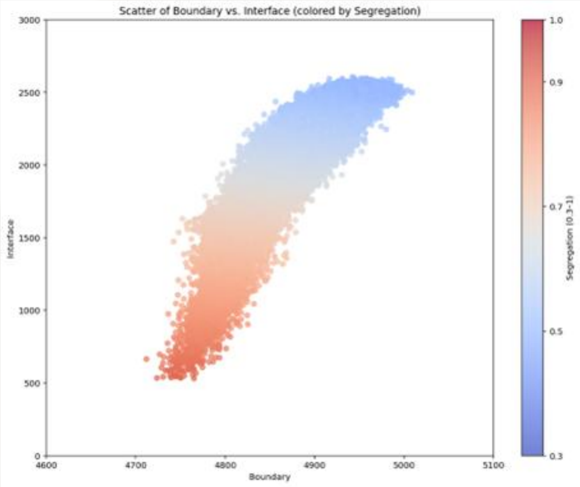
L and **L^B** : One way to become friends – language

Relationship between 4 outputs

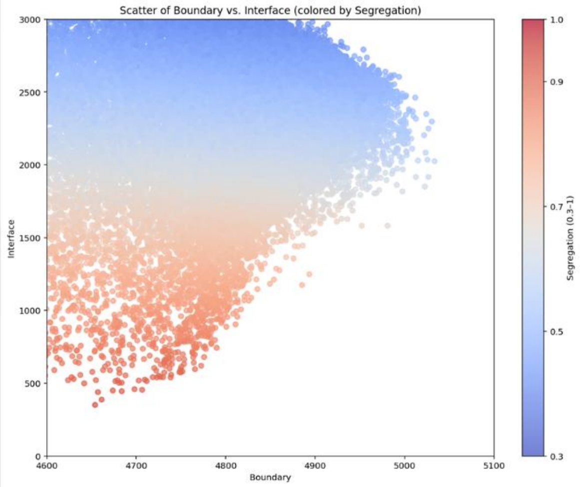
Scenario G



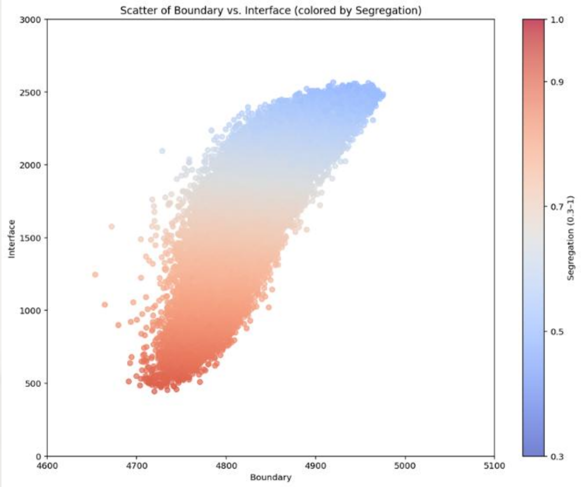
Scenario GL



Scenario L



Scenario L^B



Overall, the modified model with linguistic factors gave us more diverse and complex results.

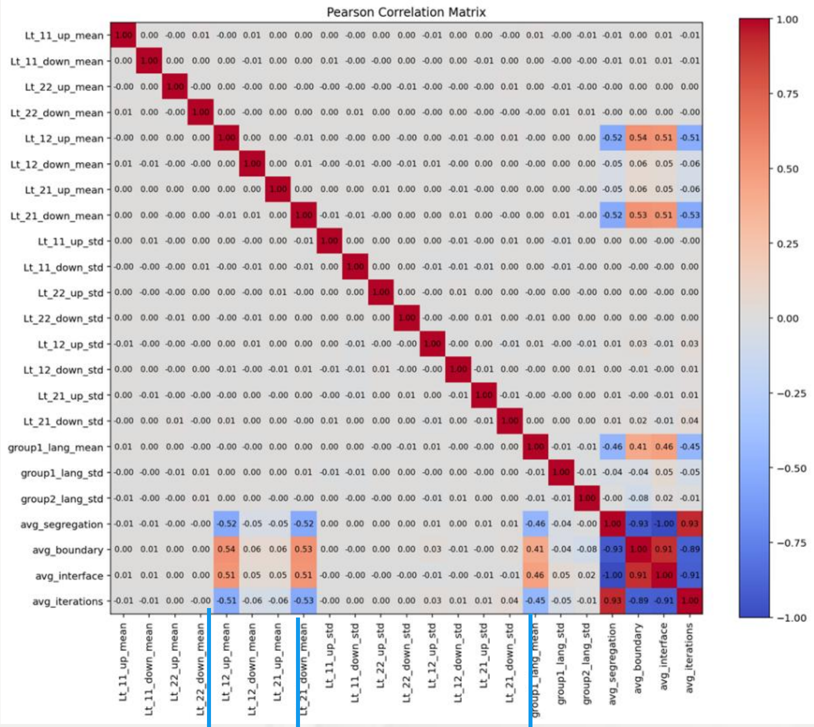
Results and Discussion

- Relationship between four outputs
- Relationship between four outputs and 19 inputs
- Regime-specific properties

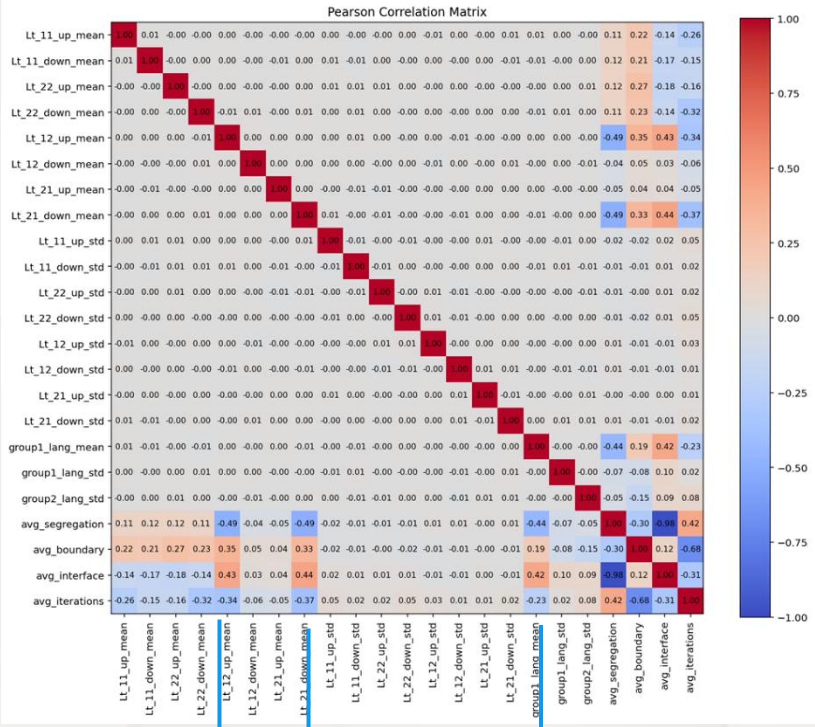
Typical interaction should be the top target for policymakers

Three inputs dominate in all three scenarios

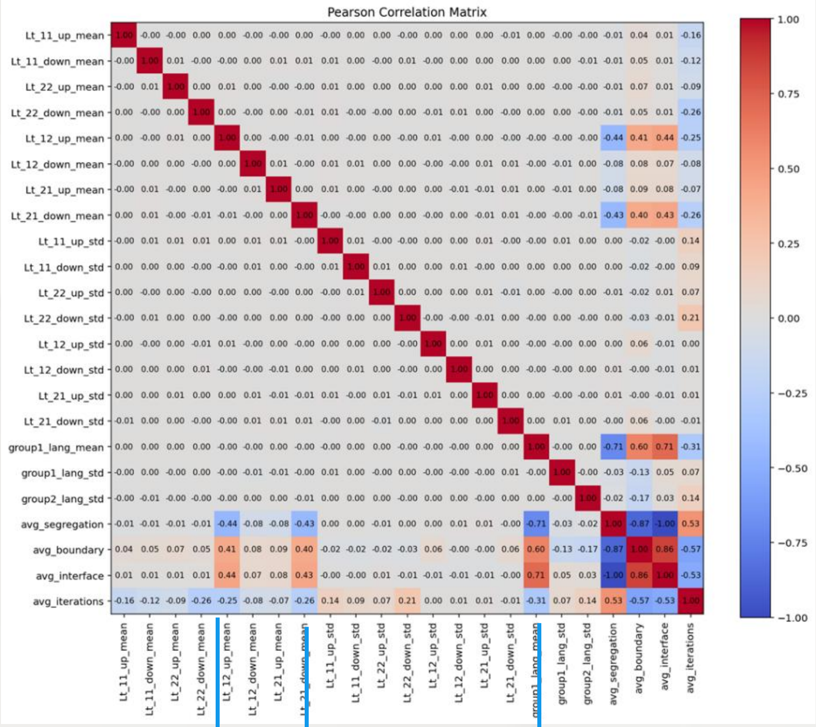
Scenario GL



Scenario L



Scenario L^{B}

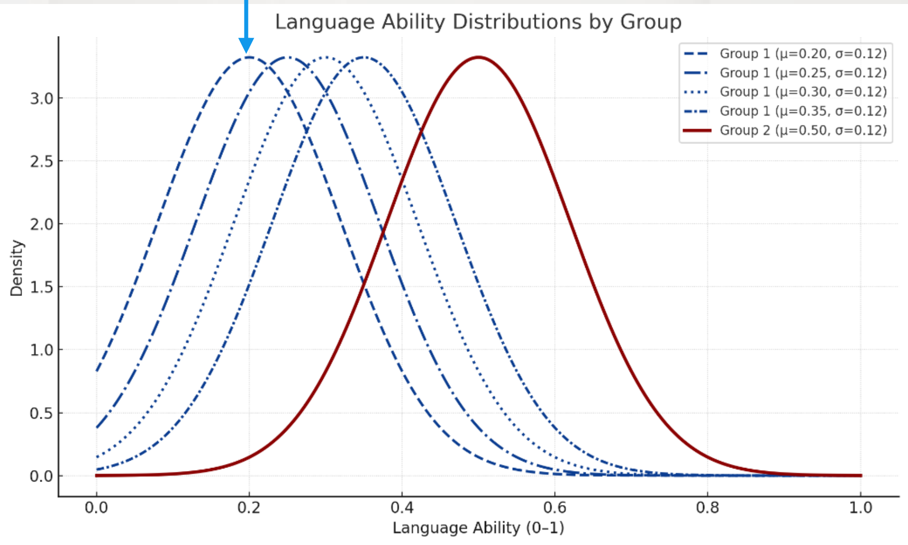


μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are key inputs across all outcomes

Top three inputs control the typical interaction

$$\mu_1^{La} \quad \alpha \quad \frac{1}{segregation}$$

μ_1^{La} determines the language gap, directly shaping the results



$$\mu_{1<2}^{Lt} \quad \alpha \quad \frac{1}{segregation}$$

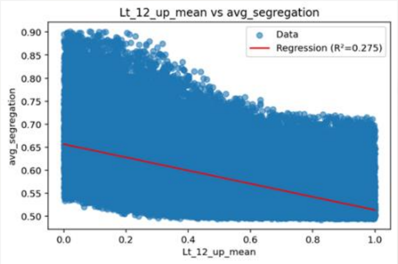
$$\mu_{2>1}^{Lt} \quad \alpha \quad \frac{1}{segregation}$$

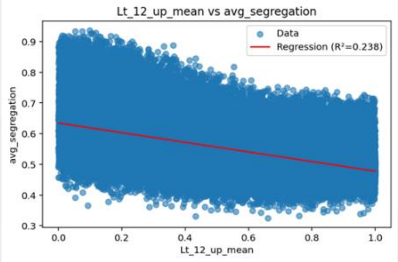
$\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ also affect the typical interaction

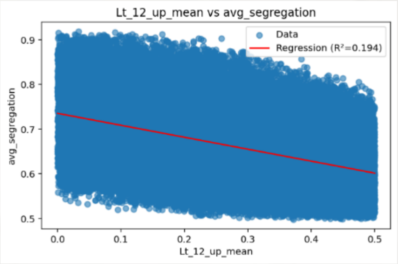
Used to assess the language gap

Top three inputs control the typical interaction

Scenario GL | **Scenario L** | **Scenario L^B**

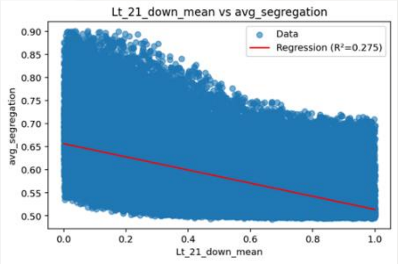


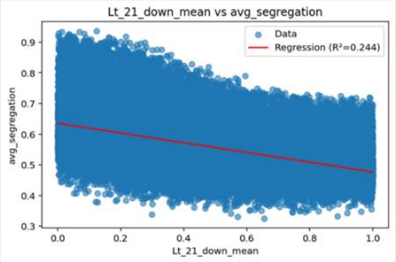


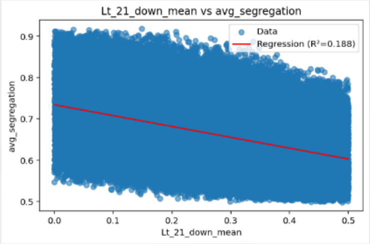


$\mu_{1<2}^{Lt}$

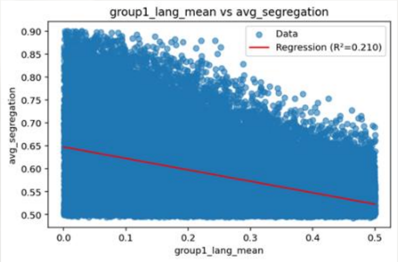
Segregation

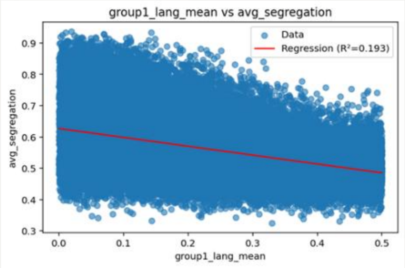


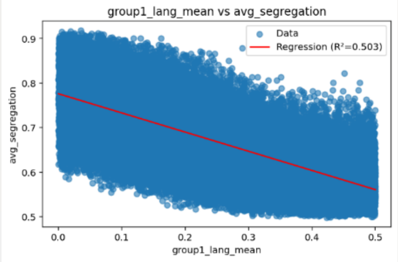




$\mu_{2>1}^{Lt}$







μ_1^{La}

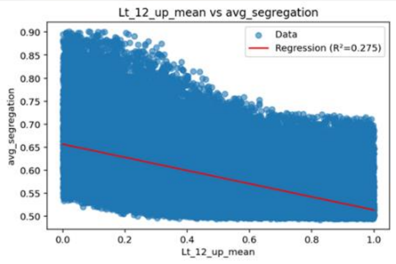
Specifically, μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are negatively correlated with segregation.

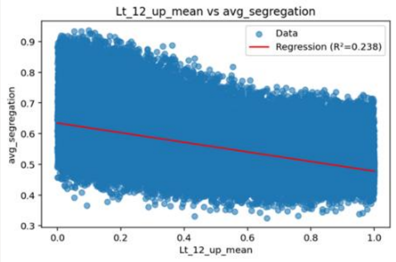
Our third finding!

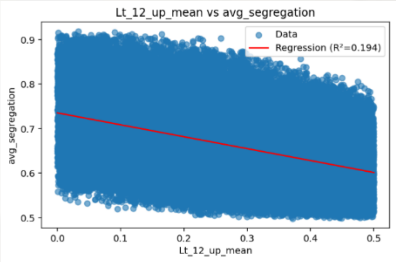
Top three inputs control the typical interaction

Scenario GL | **Scenario L** | **Scenario L^B**

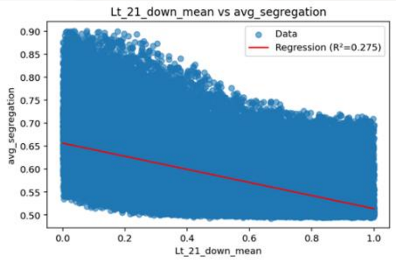
Segregation

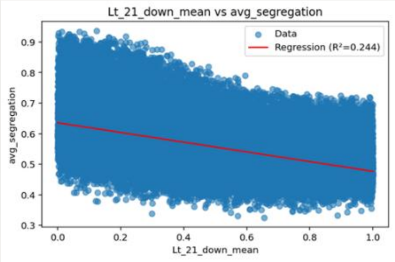


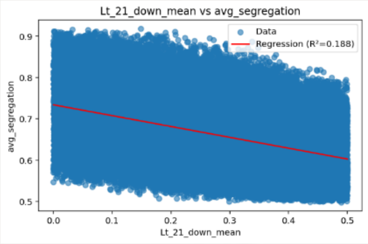




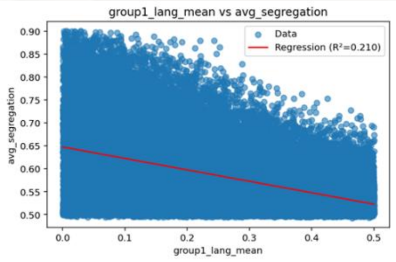
$\mu_{1<2}^{Lt}$

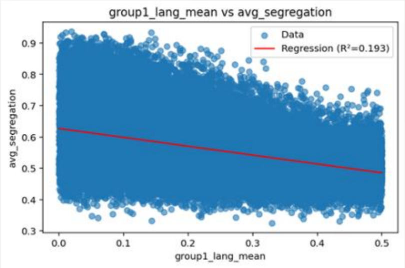


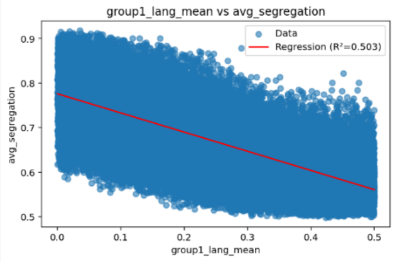




$\mu_{2>1}^{Lt}$







μ_1^{La}

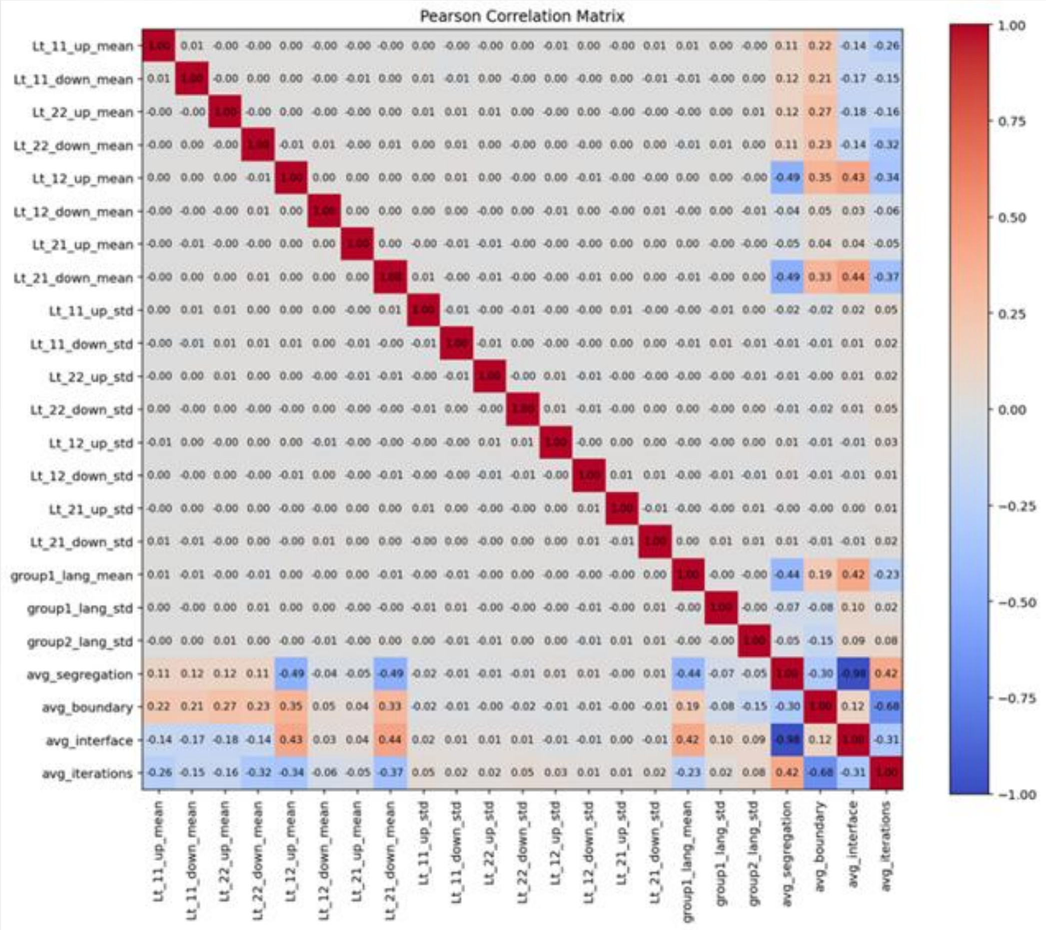
Specifically, μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are negatively correlated with segregation.

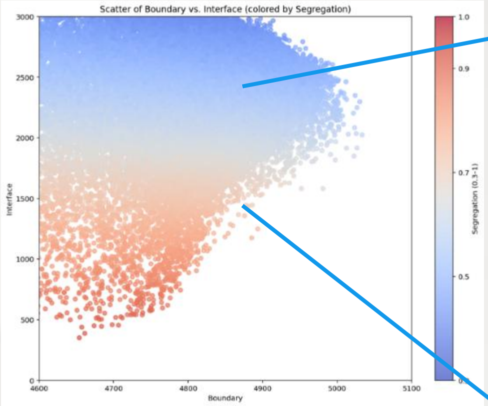
Implication:

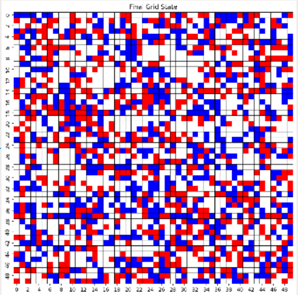
policymakers should target the typical interaction and encourage mutual tolerance.

Without explicit in-group preferences, segregation dynamics depend on more parameters

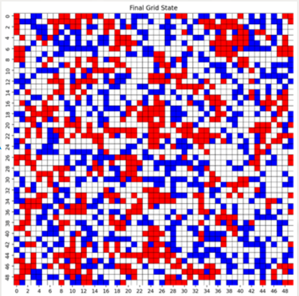
Properties specific to Scenario L







Seg=0.48, bou=4805



Seg=0.72, bou= 4834

Intragroup tolerance promotes segregation in many small clusters. Also, fast transient dynamics.

Our fourth finding!

$\mu_{1>1}^{Lt}\ \alpha\ segregation$

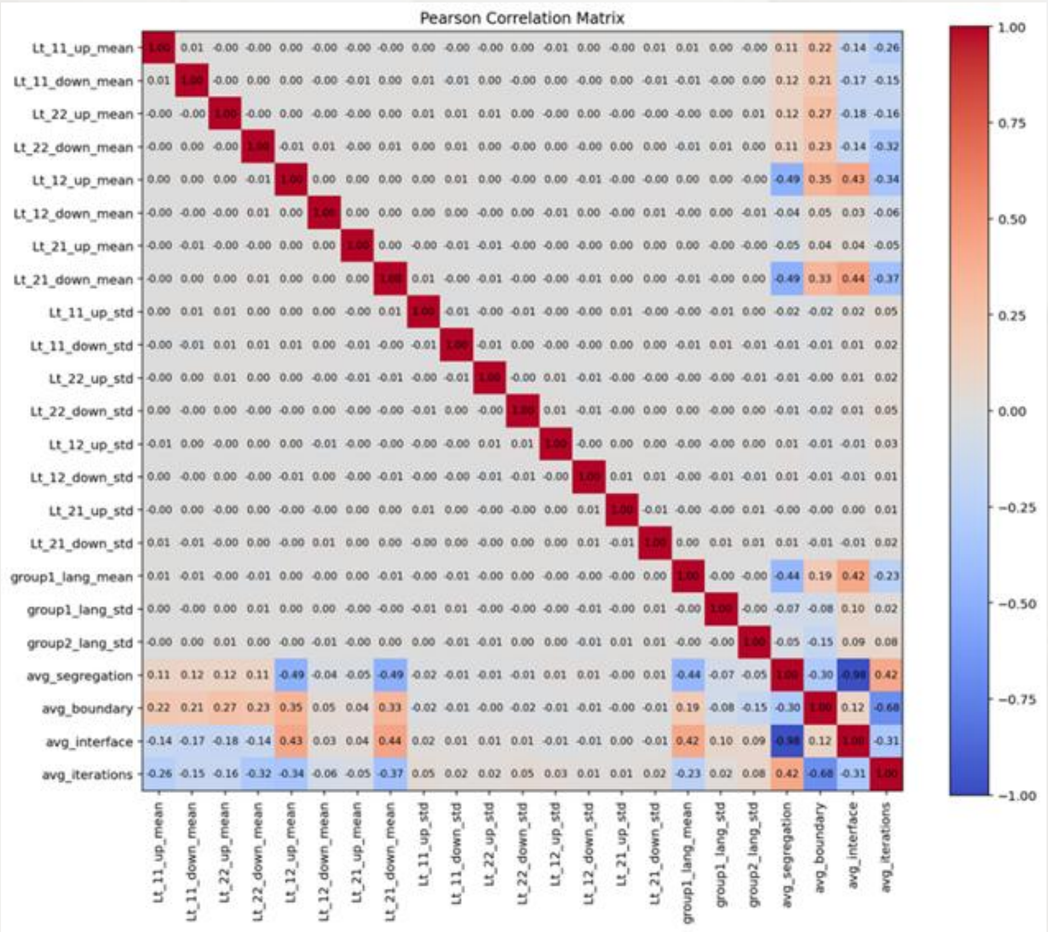
$\mu_{1>1}^{Lt}\ \alpha\ \frac{1}{interface}$

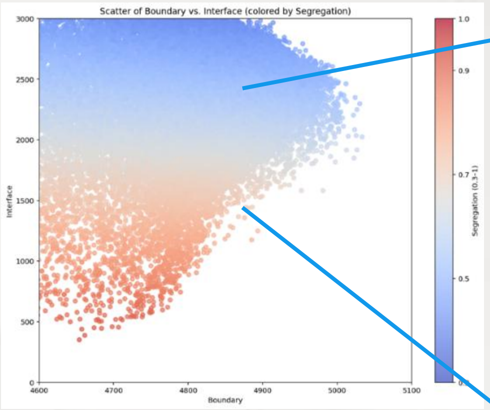
$\mu_{1>1}^{Lt}\ \alpha\ boundary$

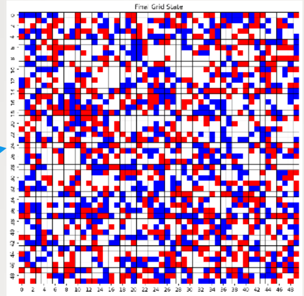
$\mu_{1>1}^{Lt}\ \alpha\ \frac{1}{iterations}$

$\mu_{1<1}^{Lt},\ \mu_{2<2}^{Lt},\ \mu_{2>2}^{Lt}$ have the same properties

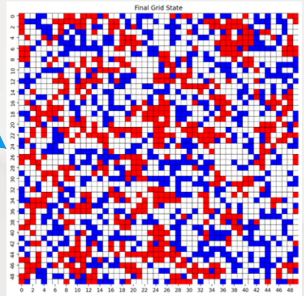
Properties specific to Scenario L







Seg=0.48, bou=4805



Seg=0.72, bou= 4834

Intragroup tolerance promotes segregation in many small clusters. Also, fast transient dynamics.

$\mu^{Lt}_{1>1}\ \alpha$ *segregation*

$\mu^{Lt}_{1>1}\ \alpha\ \frac{1}{interface}$

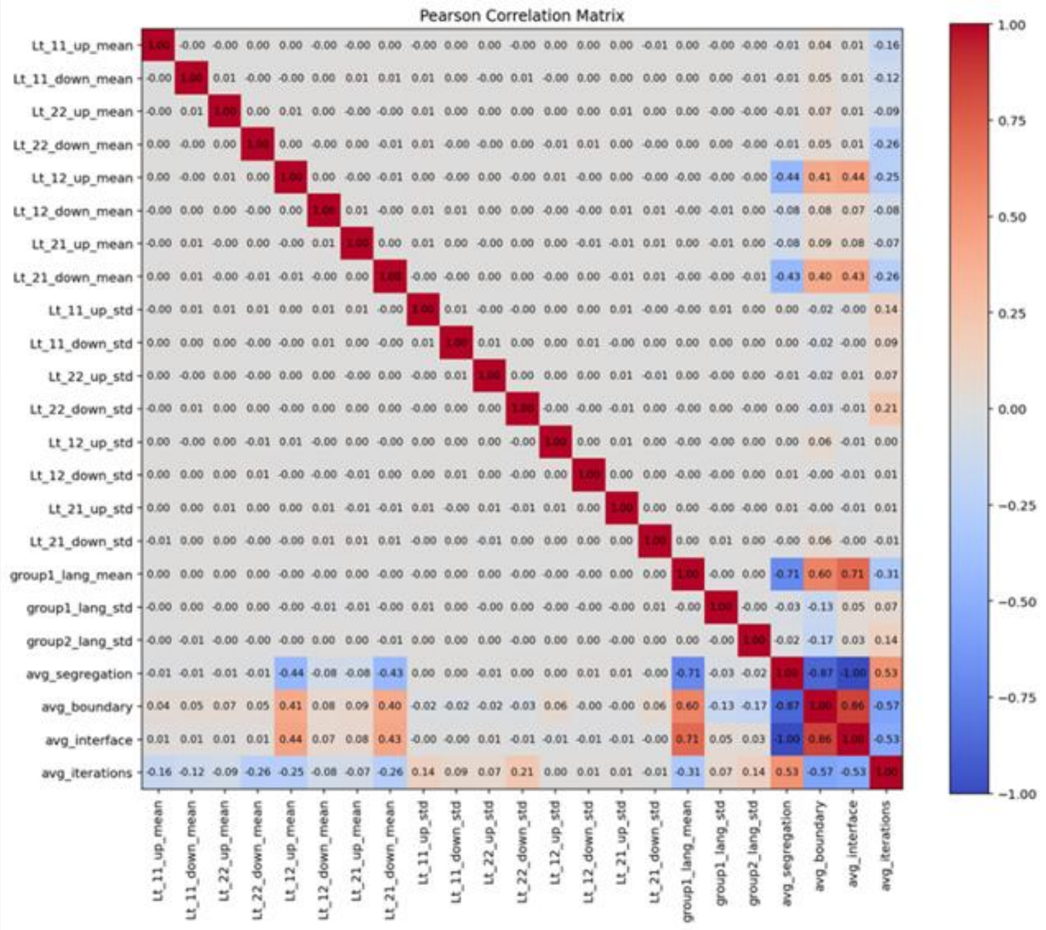
$\mu^{Lt}_{1>1}\ \alpha$ *boundary*

$\mu^{Lt}_{1>1}\ \alpha\ \frac{1}{iterations}$

$\mu^{Lt}_{1<1},\ \ \mu^{Lt}_{2<2},\ \ \mu^{Lt}_{2>2}$
have the same properties

Implication:
Intergroup mixing comes at the cost of intragroup tension.

Properties specific to Scenario L^B



$$\mu_1^{La} \;\; \alpha \;\; \frac{1}{segregation}$$

Compare to GL and L become stronger
0.46 $\rightarrow$ 0.71

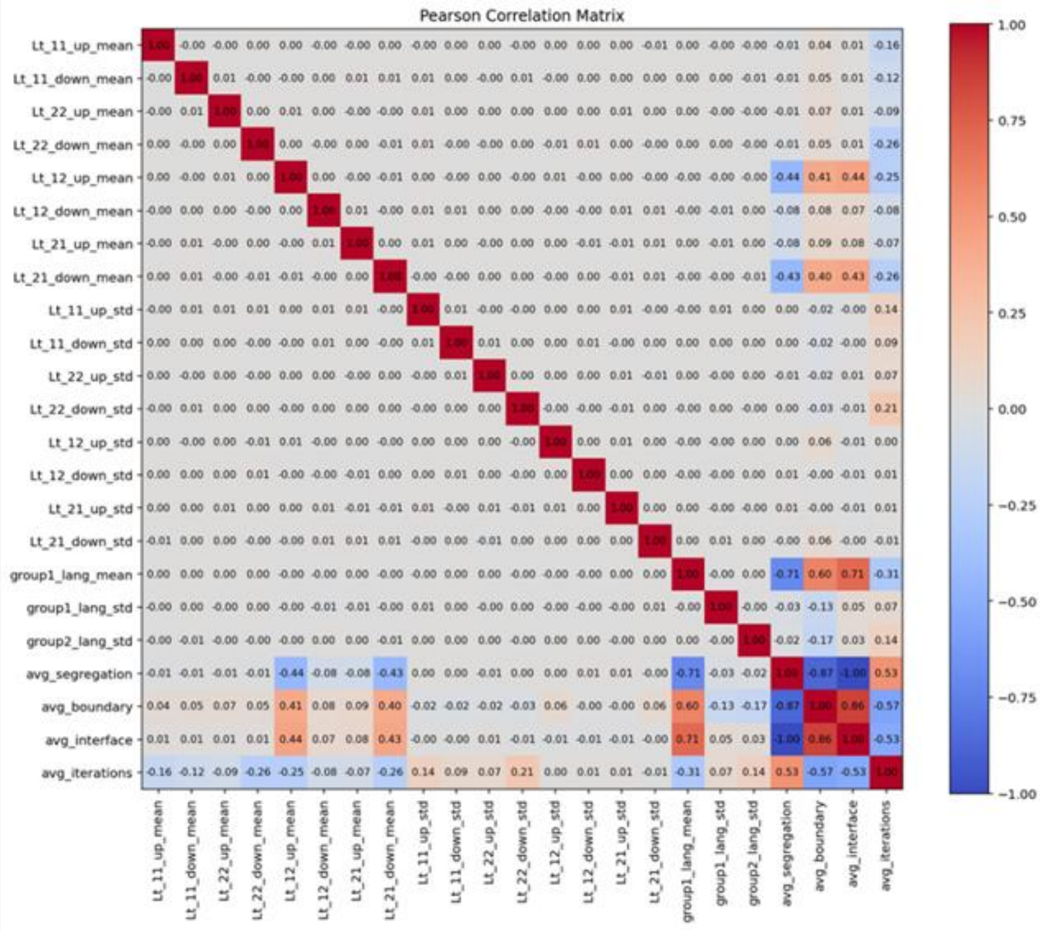
$$\mu_1^{La} \;\; \alpha \;\; \frac{1}{iterations}$$

become weaker
0.45 $\rightarrow$ 0.31

In the linguistically biased scenario, so there's an extra linguistic burden on the immigrant group.

Our fifth finding!

Properties specific to Scenario L^B

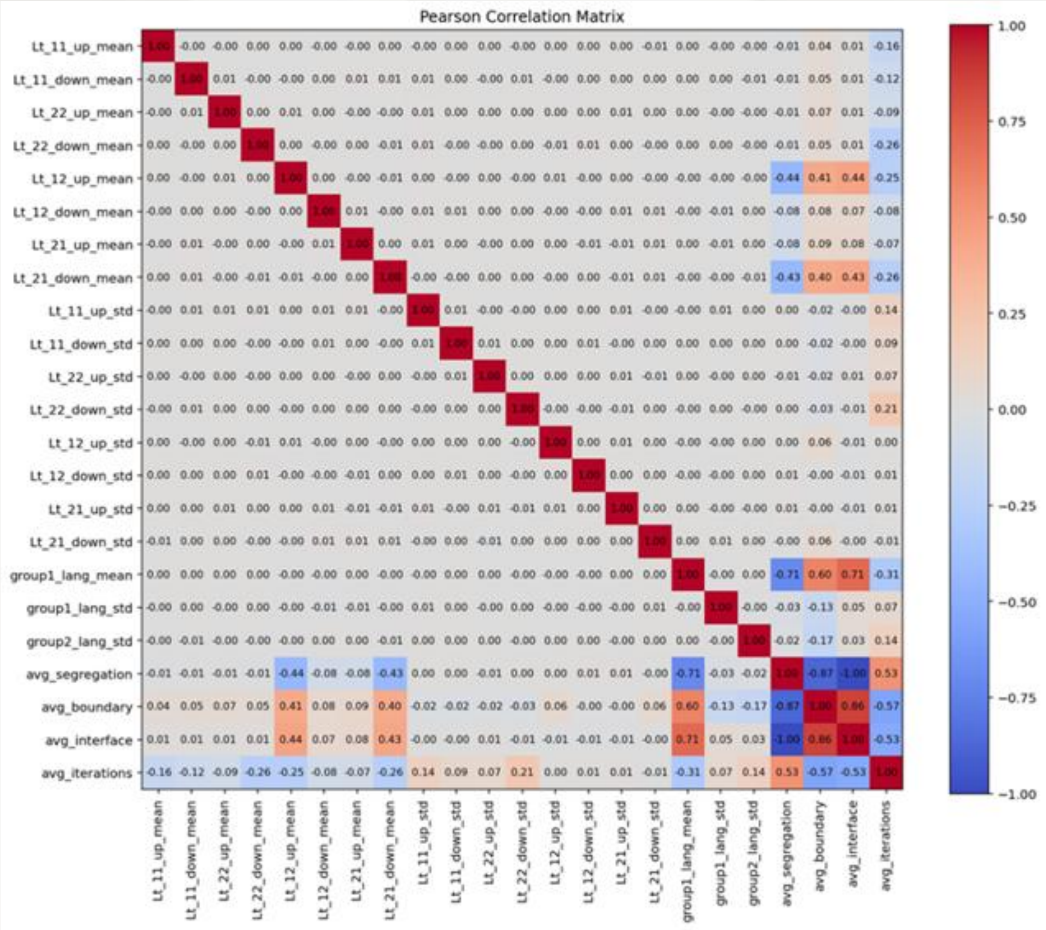


$$\mu^{Lt}_{all} \; \alpha \; \frac{1}{iterations}$$

Ironically, in the linguistically based scenario, tolerance in general discourages agents from meeting new people.

Our sixth finding!

Properties specific to Scenario L^B



$\sigma^{Lt}_{same\ group}\ \alpha$ *iterations*

$\sigma^{La}_{1\&2}\ \alpha$ *iterations*

Intragroup diversity encourages agents to meet new people.

Our seventh finding!

Validation by random forest classification and principal component analysis.

Random Forest

Split dataset into 3 regimes according to the value of **segregation, boundary, interface and iteration** separately. (k-means)

We used Random Forest classification to rank the importance of input variables in predicting low, moderate, and high outcome levels.

Segregation

Bin	Data	Interval
Low	27892	[0.49, 0.57]
Mid	15903	[0.57, 0.68]
High	6205	[0.68, 0.90]

boundary

Bin	Data	Interval
Low	10814	[4712, 4852]
Mid	17901	[4852, 4912]
High	21285	[4912, 5008]

Interface

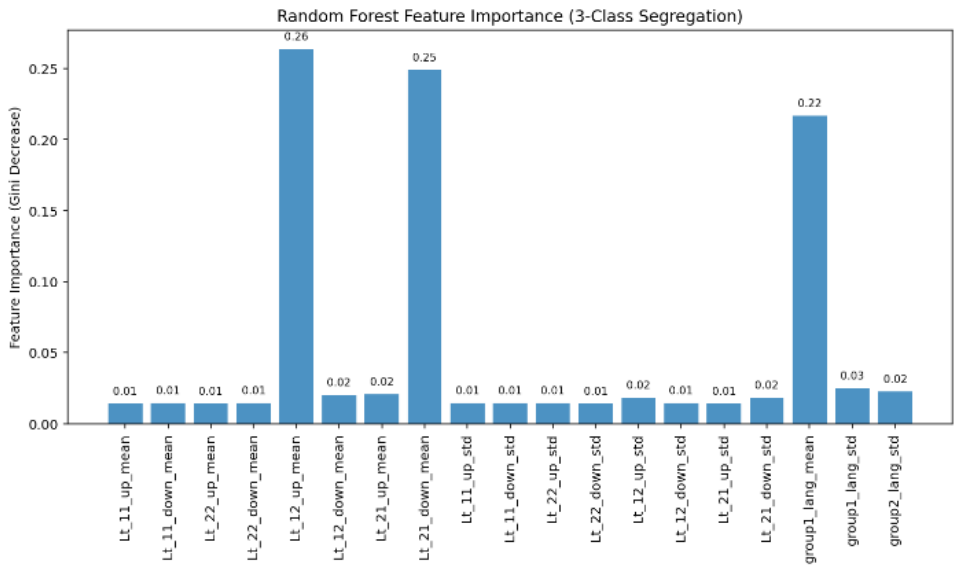
Bin	Data	Interval
Low	4606	[531, 1616]
Mid	15460	[1616, 2162]
High	29934	[2162, 2606]

Iteration

Bin	Data	Interval
Low	18098	[0.9, 4.9]
Mid	18590	[4.9, 8.7]
High	13312	[8.7, 14.8]

Random Forest

Feature importance for predicting Segregation level in Scenario GL



X-axis: features (19 variables)

Y-axis: Gini Decrease

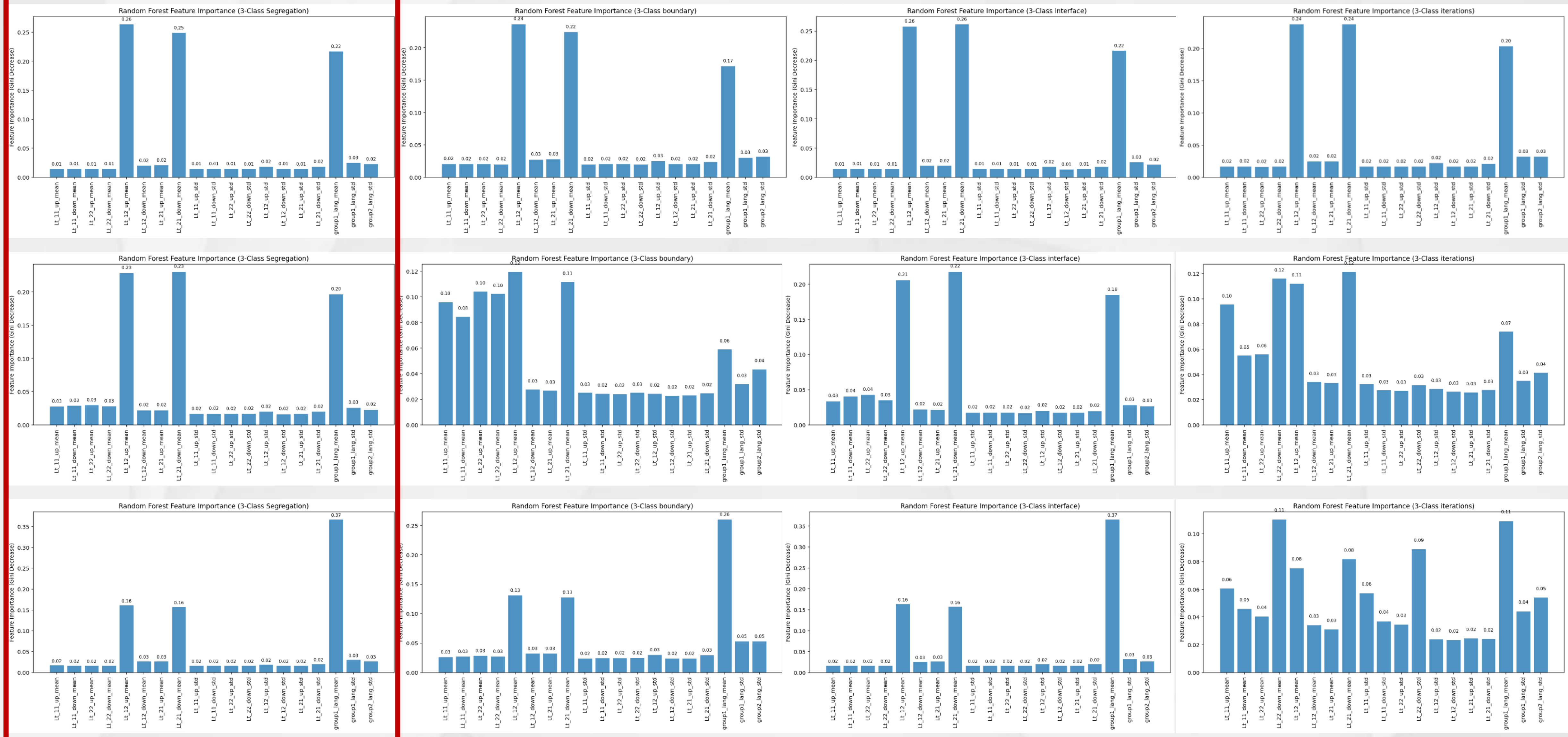
μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are the top three features in all three scenarios. Third finding confirmed.

	Segregation	boundary	Interface	Iteration

GL

L

L^B



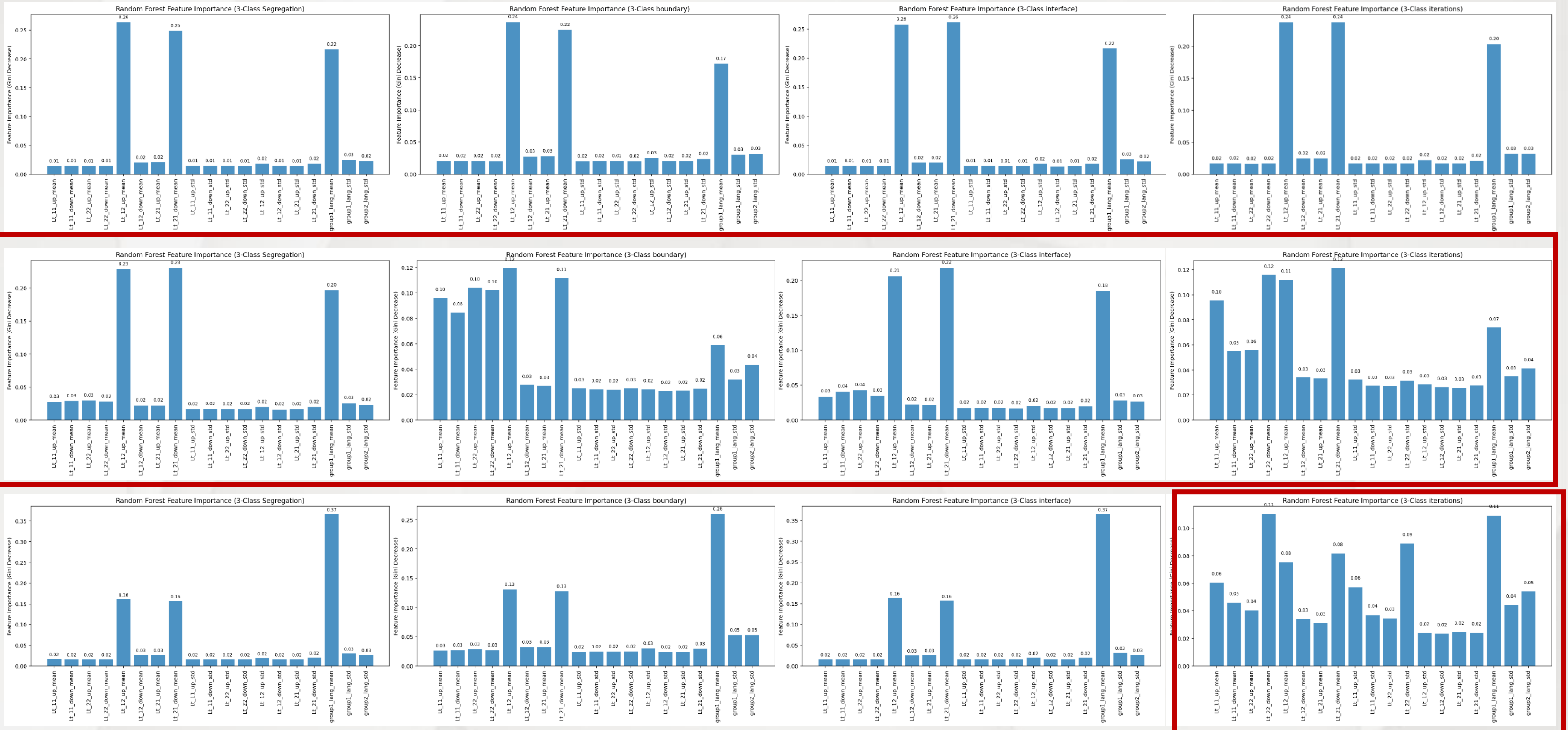
In scenarios L and L^B, boundary and iterations are more complicated. First and second findings confirmed.

Segregation | boundary | Interface | Iteration

GL

L

L^B



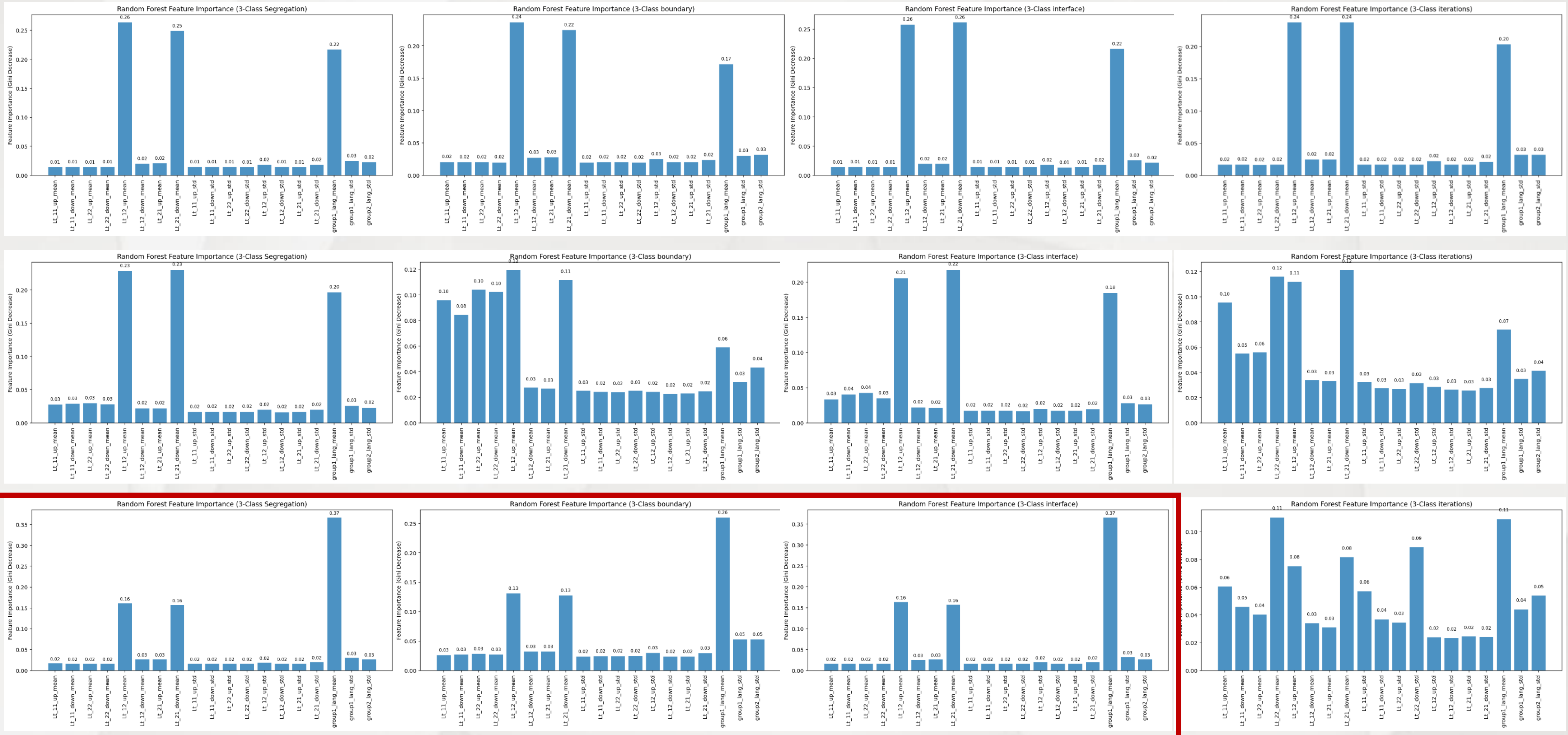
In scenarios L^B, immigrant group's language ability matters more. Fifth finding confirmed.

Segregation **boundary** **Interface** **Iteration**

GL

L

L^B



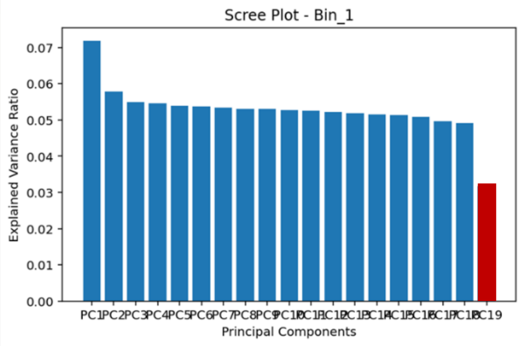
Principal Component Analysis

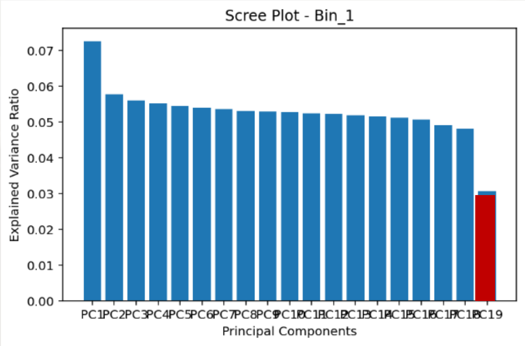
- Split dataset into three segregation regimes (low, moderate, high)
- Within each regime, apply PCA on 19 inputs
- Compare explained variance (scree plots) and variable loadings (PCs)

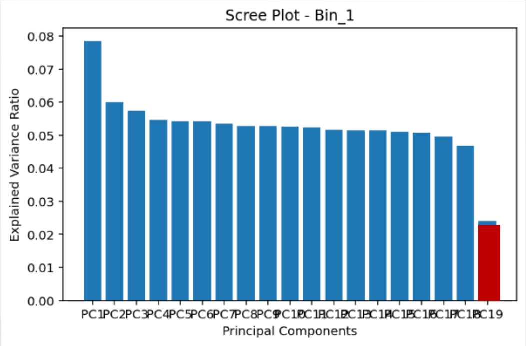
Scree Plots

Scenario - GL | Scenario - L | Scenario - L^B

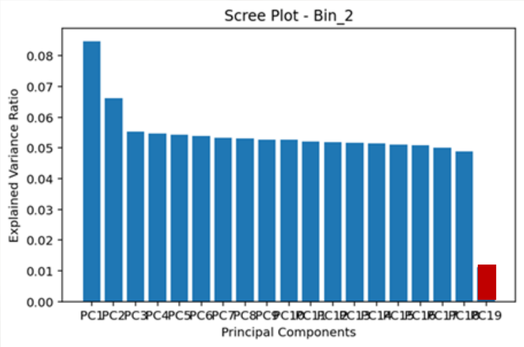
Bin 1

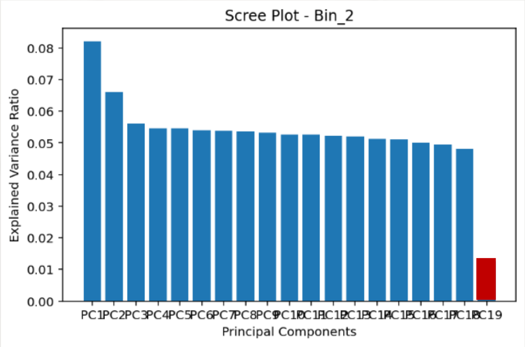


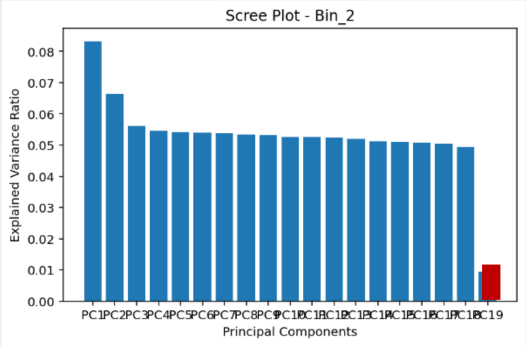




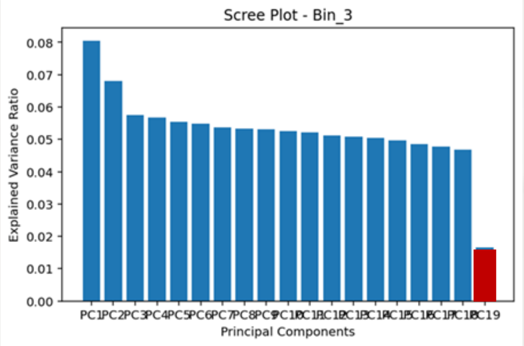
Bin 2

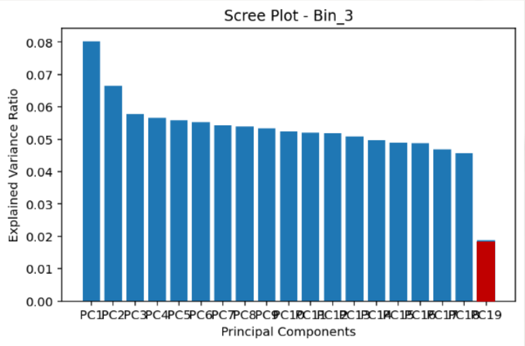


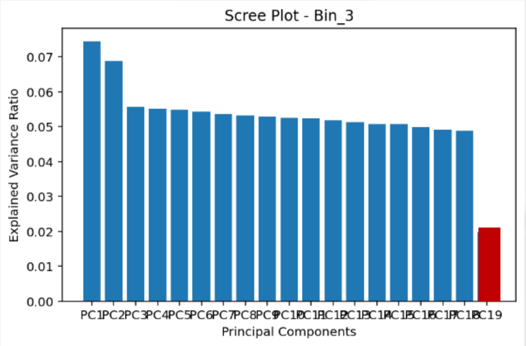




Bin 3



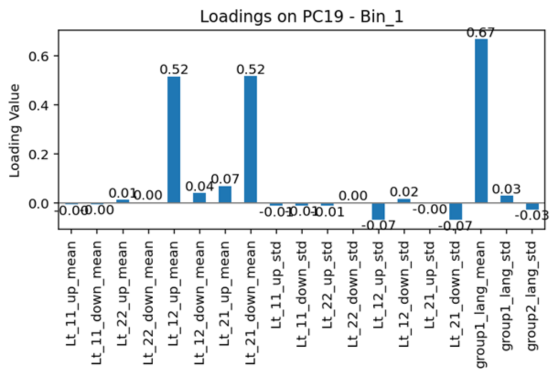


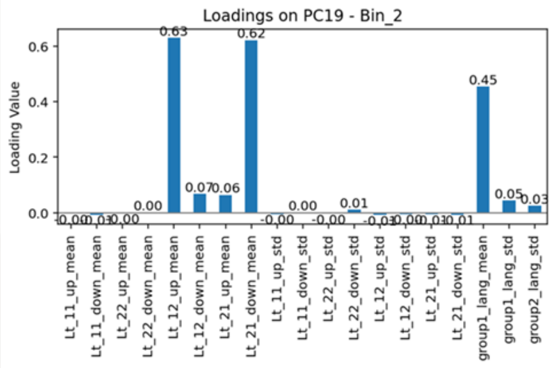


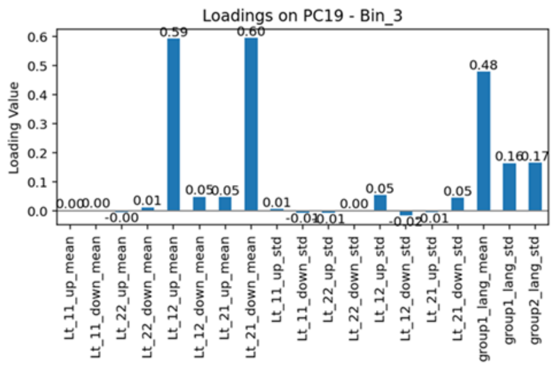
Variable Loadings (PC19)

	Bin 1	Bin 2	Bin 3

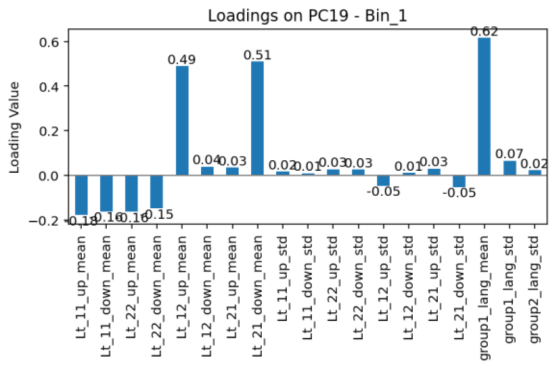
Scenario - GL

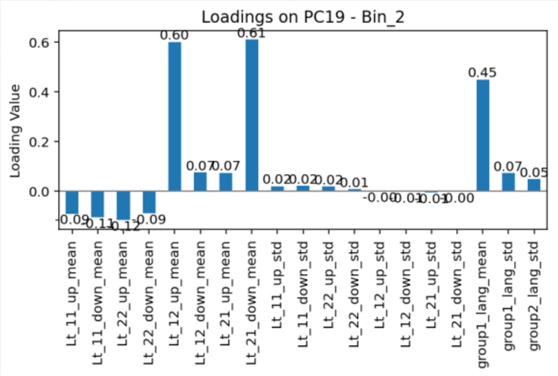


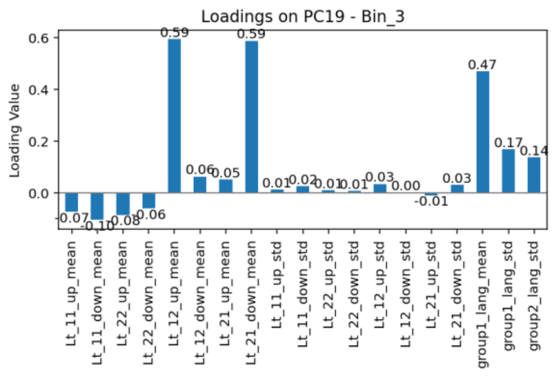




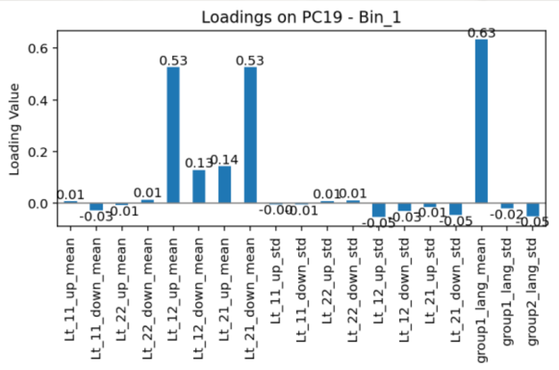
Scenario - L

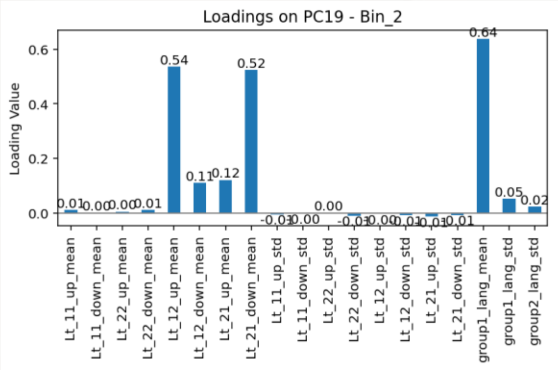


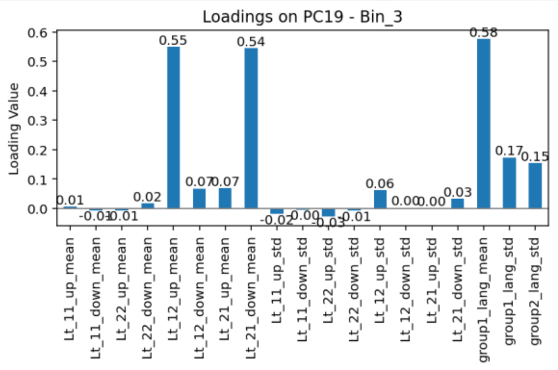




Scenario - L^B

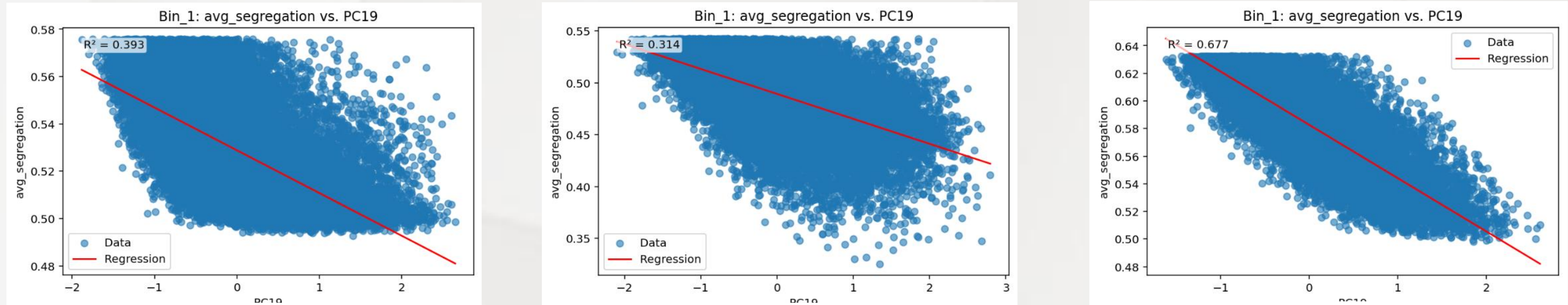




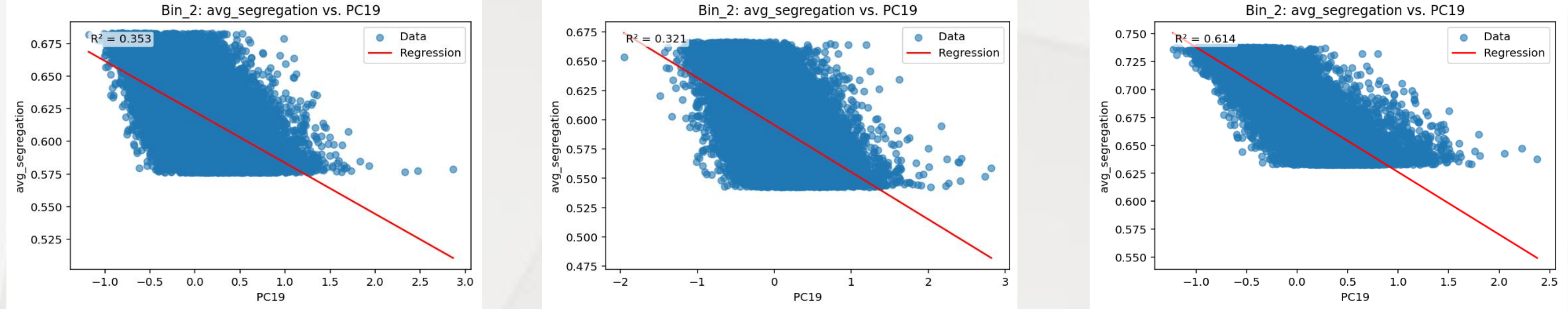


Increasing PC19 takes a society out of a regime

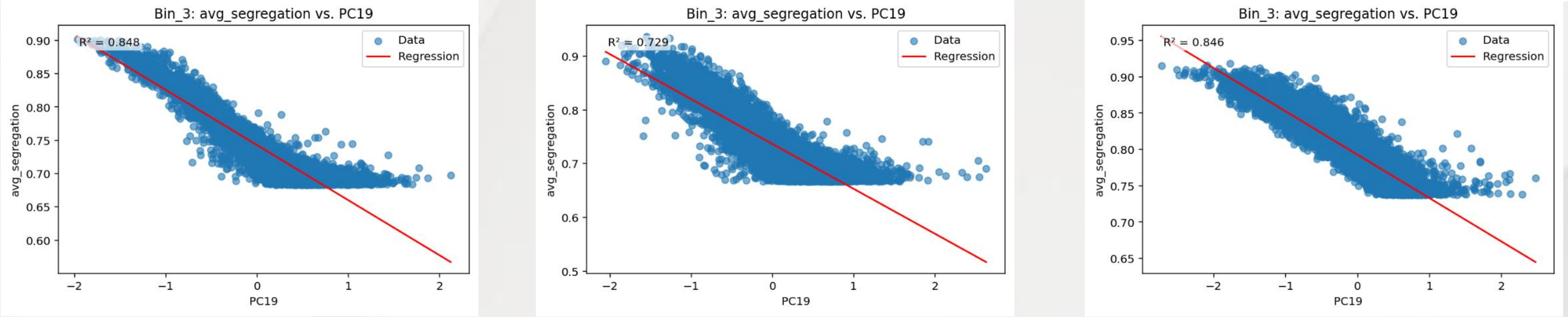
bin1



bin2



bin3



Principal Component Analysis

- Split dataset into three segregation regimes (low, moderate, high)
- Within each regime, apply PCA on 19 inputs
- Compare explained variance (scree plots) and variable loadings (PCs)

PCA results across all regimes:

- Scree plots are similar across all regimes
- The last component (PC19) consistently shows this trend:

μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are heavily loaded in the same direction

To turn a segregated society into an integrated one, both the immigrant group and the local group must be involved.

Specifically, μ_1^{La} , $\mu_{2>1}^{Lt}$ and $\mu_{1<2}^{Lt}$ are negatively correlated with segregation.

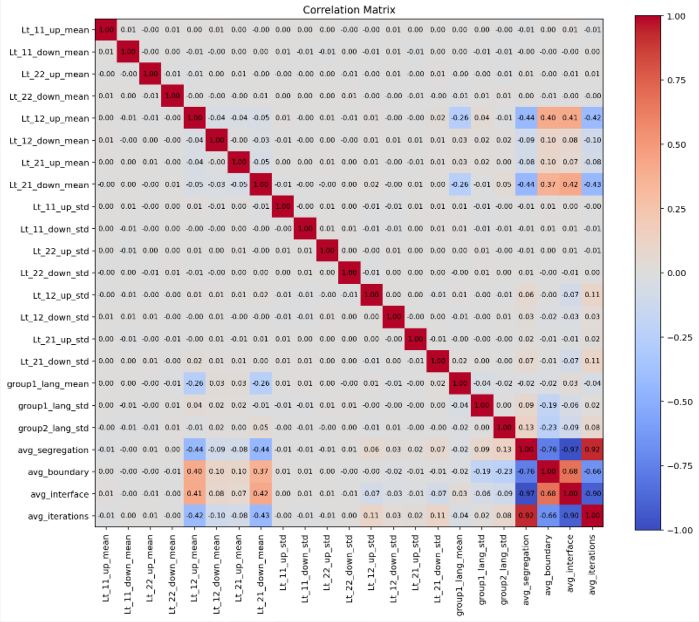
Third finding confirmed.

Results and Discussion

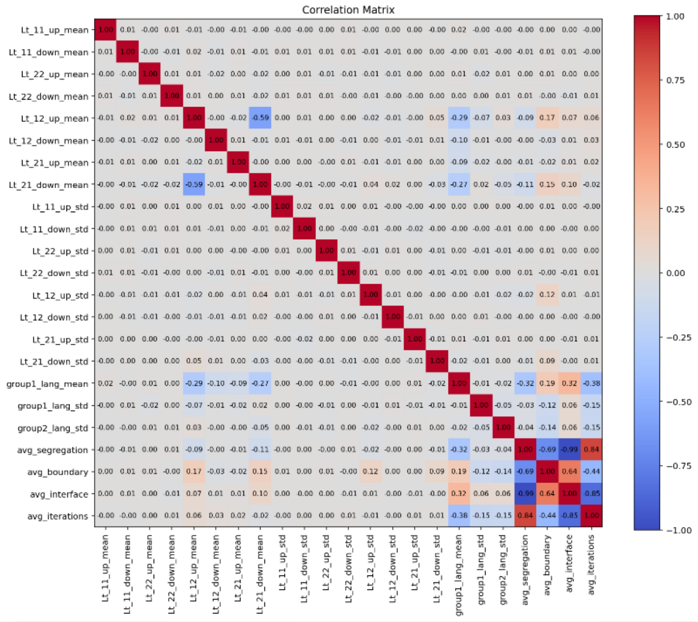
- Relationship between four outputs
- Relationship between four outputs and 19 inputs
- Regime-specific properties

Three Regimes in Scenario GL

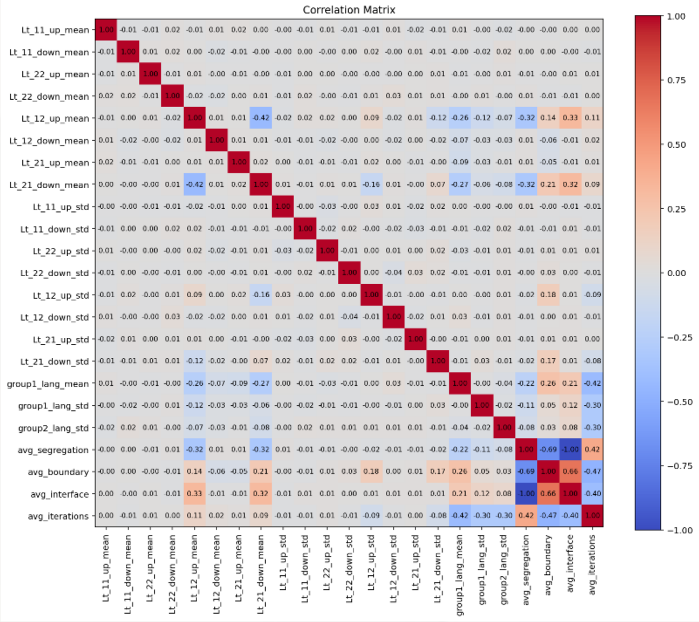
Different patterns show up in different regimes



Low segregation condition



Moderate segregation condition



High segregation condition

Regime-specific Finding 1

In each regime (low, moderate, high segregation) there is a negative correlation which is not in full dataset .

$$\mu_1^{La}\ \alpha\ \frac{1}{\mu_{1<2}^{Lt}}$$

Immigrant group's average language ability

Both average tolerance levels in the typical interaction

$$\mu_1^{La}\ \alpha\ \frac{1}{\mu_{2>1}^{Lt}}$$

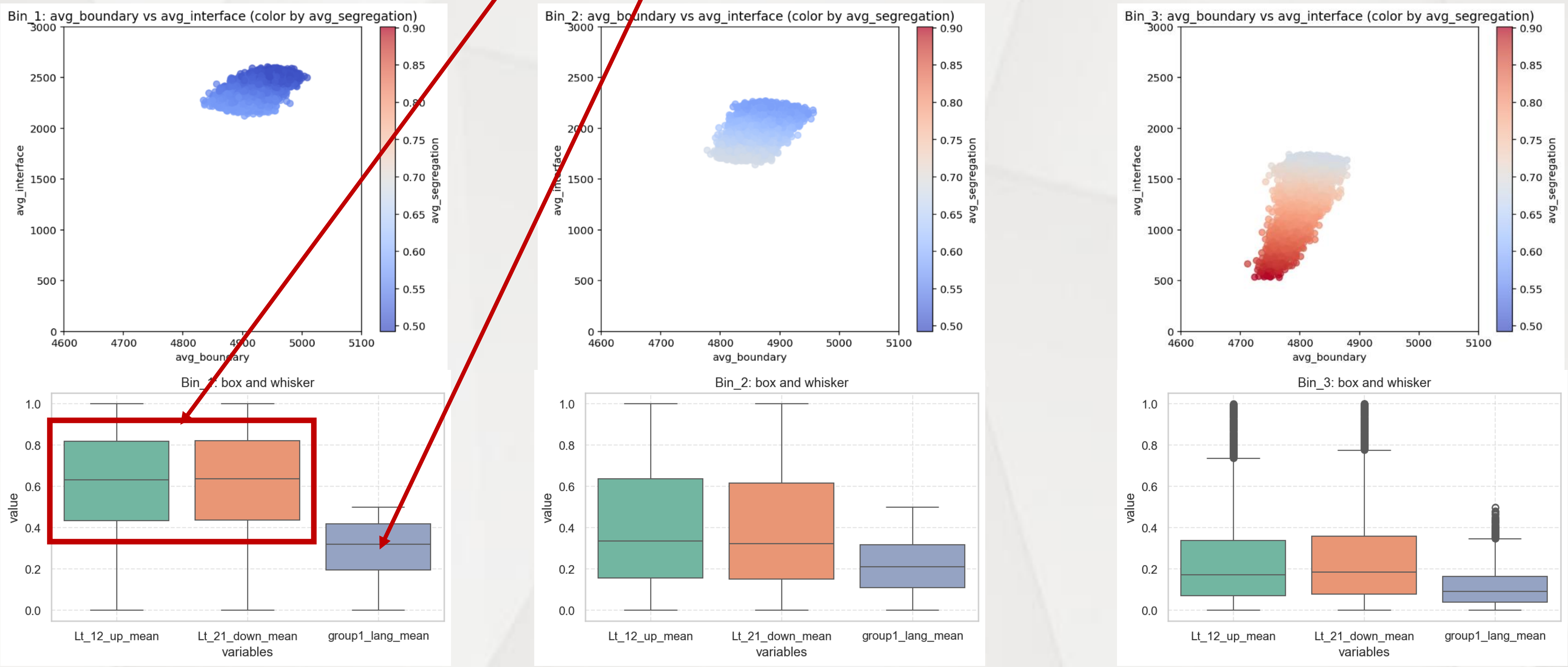
In each regime, when language gap shrinks, people become less tolerant to stay in the regime.

Implication:

Intergroup mixing depends on both language resources and ideologies.

Regime-specific Finding 2

Immigrant group's language ability matters less in low segregation regime because people are **tolerant**



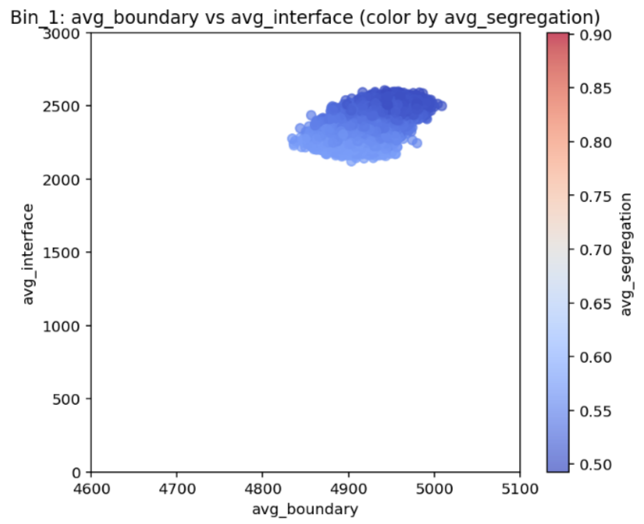
Low segregation condition

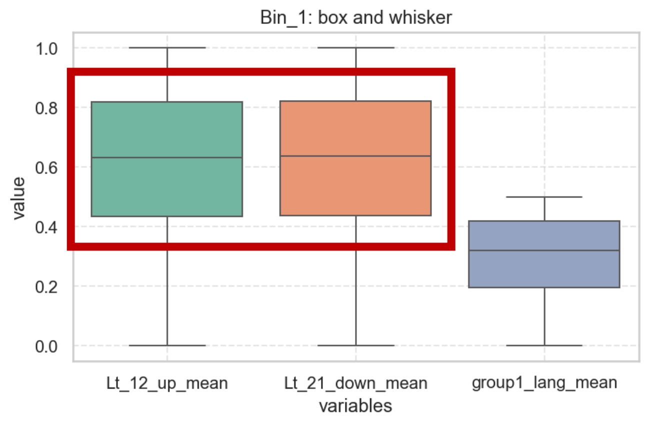
Moderate segregation condition

High segregation condition

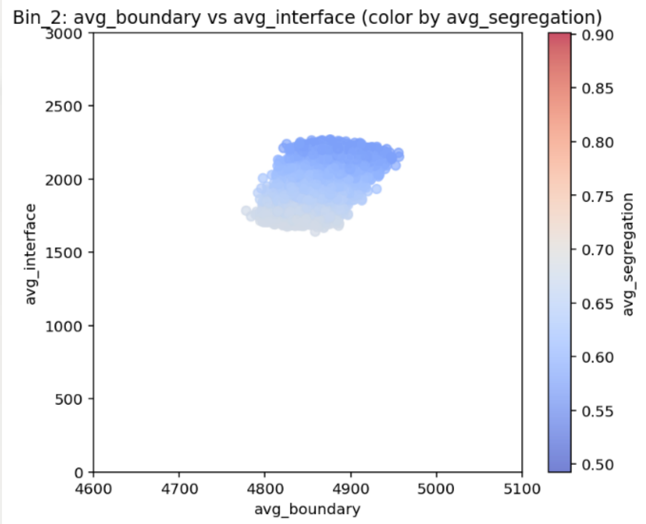
Regime-specific Finding 2

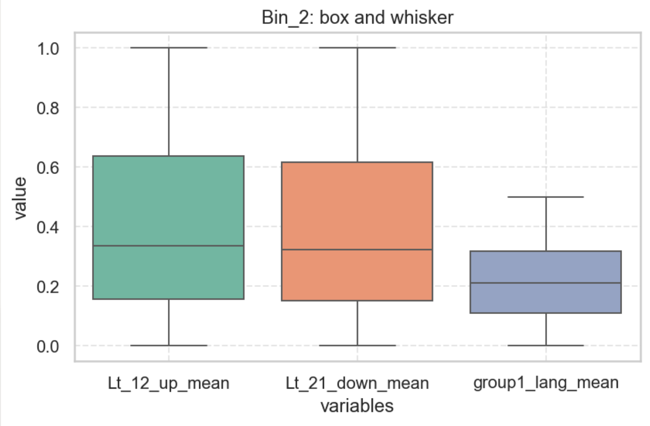
Implication:
A well-mixed society is marked by high levels of tolerance. It doesn't require immigrant group to improve their language ability further.



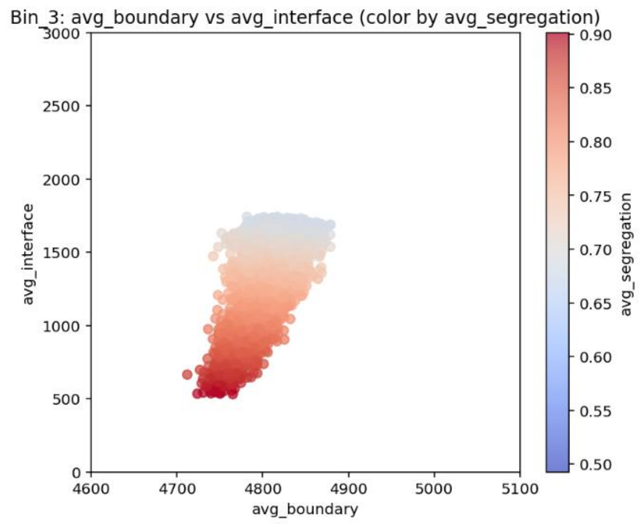


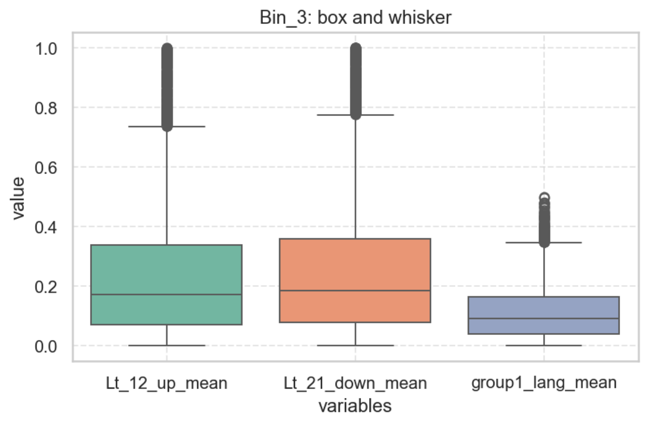
Low segregation condition





Moderate segregation condition





High segregation condition

Regime-specific Finding 3

In the moderate and high segregation regimes:

Immigrant group's average language tolerance in the typical interaction

$$\mu_{2>1}^{Lt} \; \alpha \; \frac{1}{\mu_{1<2}^{Lt}}$$

Local group's average language tolerance in the typical interaction

A segregated society is marked by a lack of mutual tolerance.

Implication:

Consistent with finding 2, policies targeting ideologies are better than policies targeting resources.

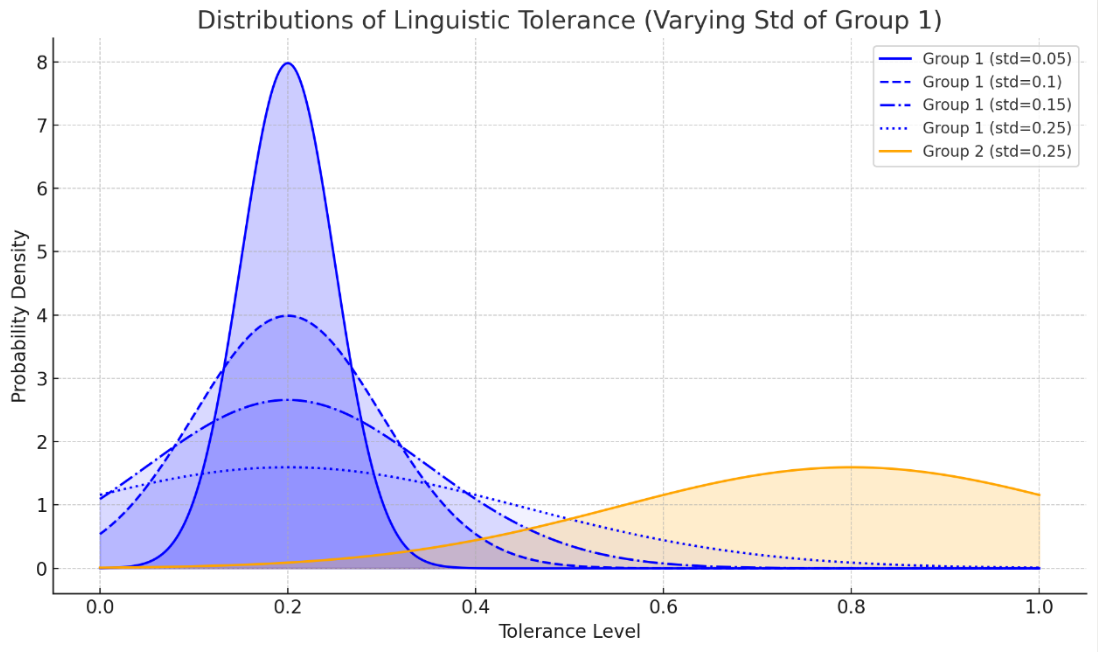
Encourage both groups to be patient with each other.

Regime-specific Finding 4

In the high segregation regime:

$$\sigma^{Lt}_{2>1} \ \alpha \ \ boundary$$

$$\sigma^{Lt}_{1<2} \ \alpha \ \ boundary$$



In a segregated society, intra-group diversity in out-group attitudes (tolerance) results in fragmentation.

Implication:

In a segregated society, intergroup mixing is less correlated with fragmentation.

Could reducing fragmentation be a stepping stone to intergroup mixing?

Regime-specific Finding 5

Low segregation condition

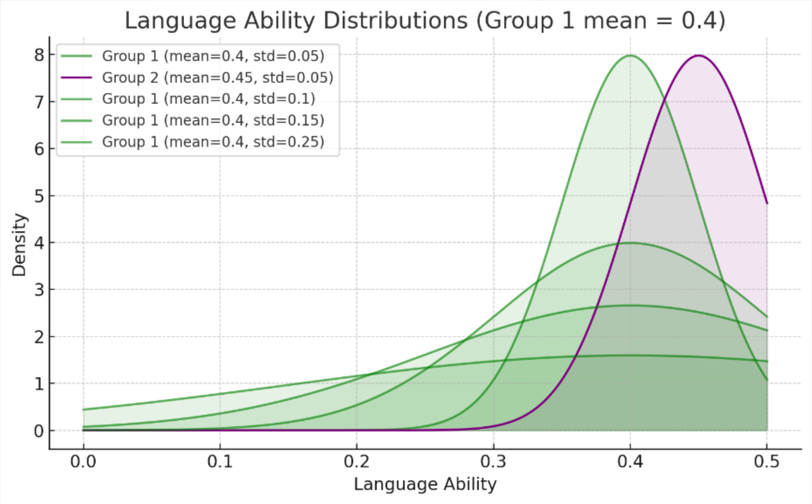
$\sigma^{La}_{1\&2}\ \alpha$ segregation : 0.1

Moderate segregation condition

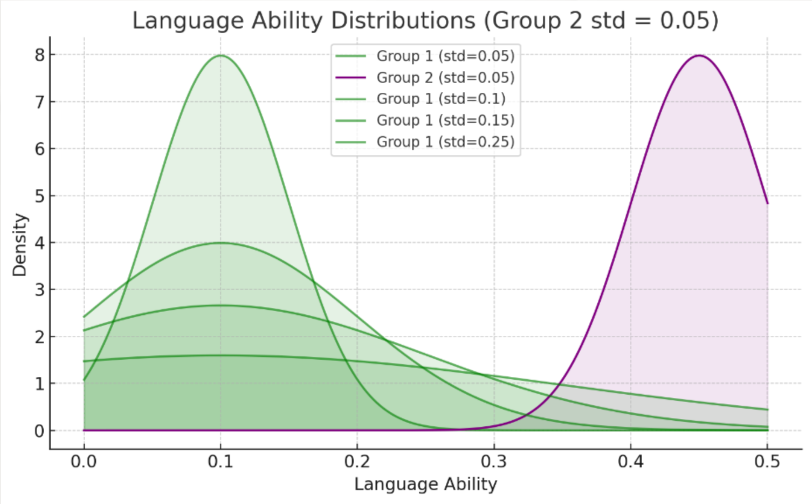
$\sigma^{La}_{1\&2}\ \alpha$ segregation : ≈ 0

High segregation condition

$\sigma^{La}_{1\&2}\ \alpha$ segregation : ≈ - 0.1



In a well-mixed society, intra-group diversity in language resources creates segregation.



In a segregated society, intra-group diversity in language resources mitigates segregation.

Regime-specific Finding 5

Low segregation condition

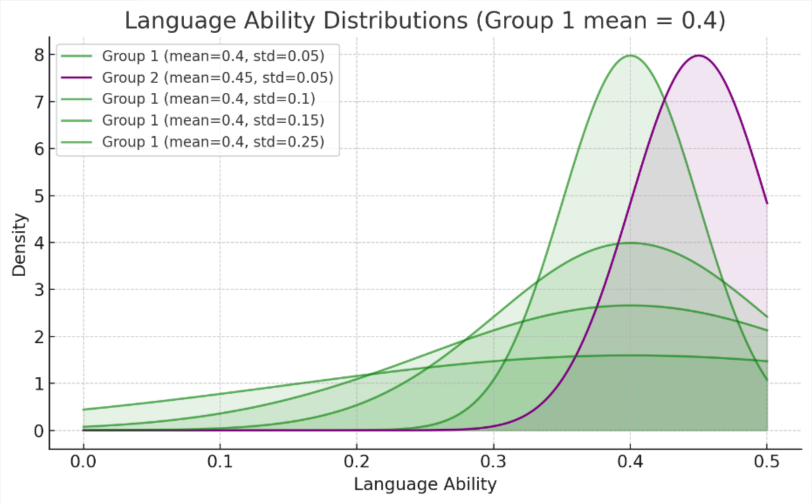
$\sigma^{La}_{1\&2}\ \alpha$ segregation : 0.1

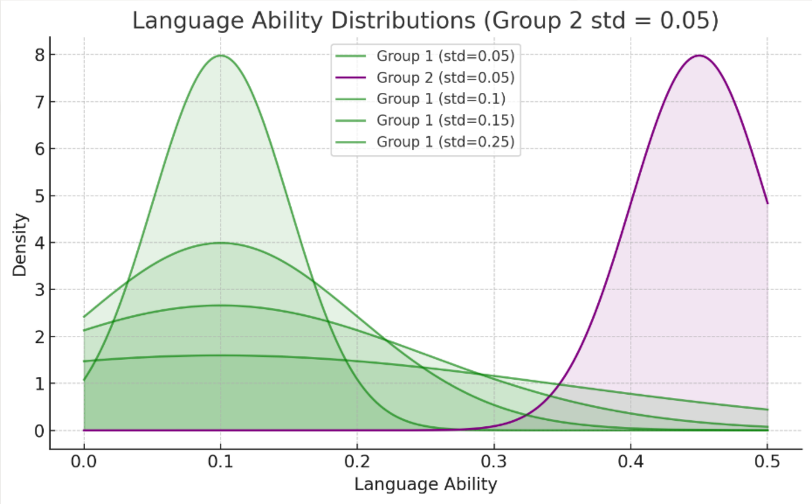
Moderate segregation condition

$\sigma^{La}_{1\&2}\ \alpha$ segregation : ≈ 0

High segregation condition

$\sigma^{La}_{1\&2}\ \alpha$ segregation : ≈ - 0.1





Implication: In a well-mixed society, rigid language norms maintain integration.

Implication: In a segregated society, policies promoting linguistic diversity reduce segregation.

Next steps

- Regime-specific findings in scenarios L and L^B
- Make language ability and tolerance level multidimensional
- Connect our findings to ethnographic studies (potential collaboration with Prof. Sender Dovchin)
- Use LLM agents and model agent interactions as conversations (potential collaboration with Dr Katie Cunnah)
- Incorporate EEG measurements representing physiological responses to agent interactions (potential collaboration with Dr Kate Stone)

Conclusion

1. Linguistic factors result in more complex segregation dynamics than in-group preferences.

- In GL, intergroup mixing occurs in many small clusters.
- Without explicit in-group preferences (L and L^B), agents meet more people during the transient period.

2. When language matters (GL, L, and L^B), typical interaction should be the top target for policymakers. Language resources and ideologies both matter.

Conclusion

3. Without explicit in-group preferences, segregation dynamics depend on more parameters.

- In L, intergroup mixing comes at the cost of intragroup tension.
- In L^B, tolerance in general ironically discourages agents from meeting new people.
- In L^B, intragroup diversity in language resources and ideologies encourages agents to meet new people.

4. When in-group preferences are hidden in language ideologies only (L^B), the immigrant group bears an extra linguistic burden.

Conclusion

5. In GL, intergroup mixing depends on both language resources and ideologies in all three levels of segregation. Policies targeting language ideologies rather than resources are consistently more effective.

6. In GL, societies with different levels of segregation display minor differences.

- In a segregated society, the precise level of segregation is less correlated with fragmentation.
- In a well-mixed society, rigid language norms maintain integration. In a segregated society, policies promoting linguistic diversity reduce segregation.

THANK