Study on how social efficacy effects integration of tertiary level local and international students in Hong Kong

Hypothesis

The hypothesis tested in this analysis was that local students, compared to non-local students, tend to be more introverted, which could create social barriers for international students, making it harder for them to integrate into the local student community at CUHK.

Objective

The purpose of this survey is to assess the social activeness and efficacy of participants through a set of 19 personality-related questions. The questions are designed to capture a variety of social behaviors and traits, ranging from comfort in social settings to ease in making friends. The survey was administered to both local and international students, and we aim to explore patterns in the responses, as well as identify socially active and inactive groups.

Data analysis

Data Cleaning and Preprocessing

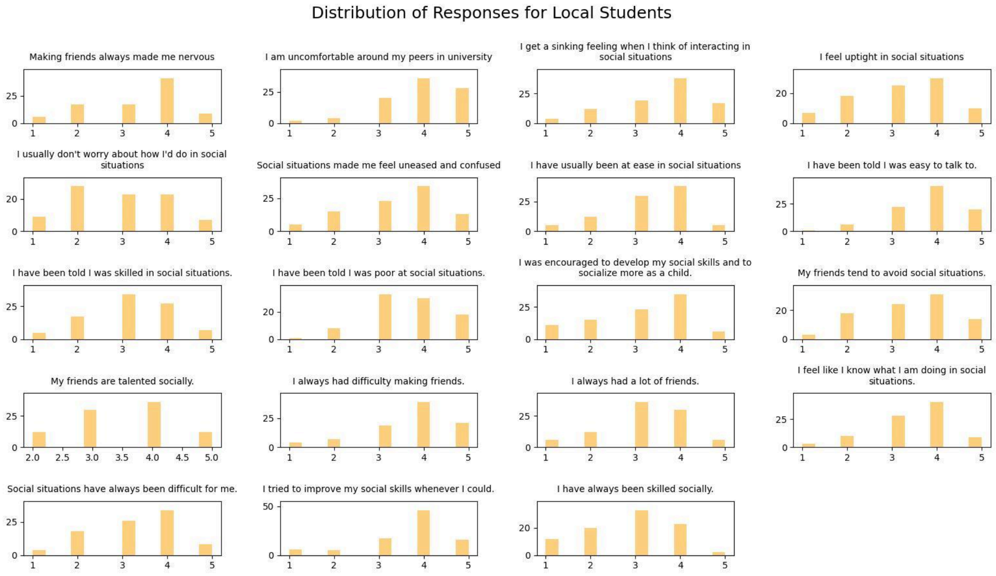
Before doing data analysis of the data, the first step was to do data cleaning to filter out the outliers and to check for missing data as outliers and missing values can distort the results of the analysis, but in this case, there were no missing entries, allowing us to proceed directly with the available data.

To create a consistent scale across all 19 questions in the survey, we applied normalization of the data. The survey questions are phrased in both positive and negative ways. For example, one question may ask about being comfortable in social situations (where a high score of 5 means strong social efficacy), while another might ask about avoiding social situations (where a high score of 5 would indicate a lack of social efficacy). Specifically, for questions framed negatively, we reversed the scoring. For instance, if a respondent selected a score of 5 for a negatively phrased question (indicating they avoid social situations), we transformed that into a score of 1, aligning it with the overall goal of making a score of 5 consistently representing the most positive (or socially active) response. This normalization process allowed us to standardize the results and ensure that all questions were comparable.

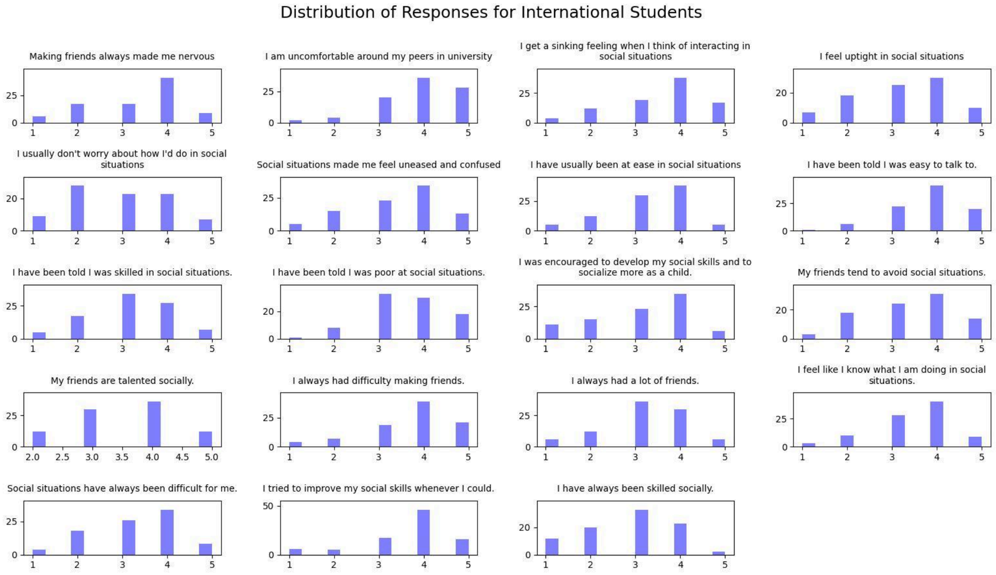
Visualizing Questions and Answer Distribution

Visualizing the distribution of responses helps to understand how participants answered each question. By generating histograms for each of the 19 questions, we were able to see the overall trends in responses for both local and international students.

For each group, we created a set of histograms that show how frequently each response (from 1 to 5) was chosen. This revealed how participants viewed themselves in terms of social skills, social comfort, and social engagement. This is a key step in identifying patterns in behavior before proceeding to deeper analysis.



X-axis: 1 being least positive and 5 being most positive. Y-axis: Frequency



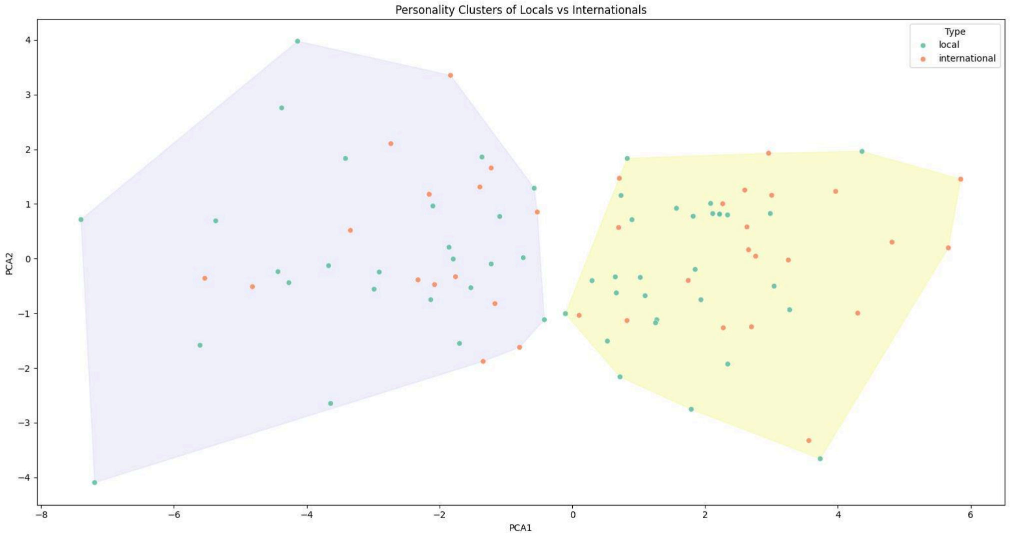
X-axis: 1 being least positive and 5 being most positive. Y-axis: Frequency

With 19 different questions, it’s difficult to draw meaningful conclusions just by examining the distributions alone. Therefore, we applied clustering techniques to group the participants based on their overall social behavior.

To simplify the analysis, we decided to use K-Means clustering, a common unsupervised learning technique that organizes data into distinct groups. Since we are particularly interested in identifying socially active versus socially inactive participants, we set the number of clusters to two. This way, the first cluster would represent participants who tend to be more socially active, while the second cluster would capture those who are less active socially.

With 19 dimensions (one for each question), it’s challenging to visualize the clustering results in a meaningful way. Principal Component Analysis (PCA) is a technique that reduces the complexity of high-dimensional data while retaining most of the variation in the data. This allows us to project the results onto a 2D plot, making it easier to interpret.

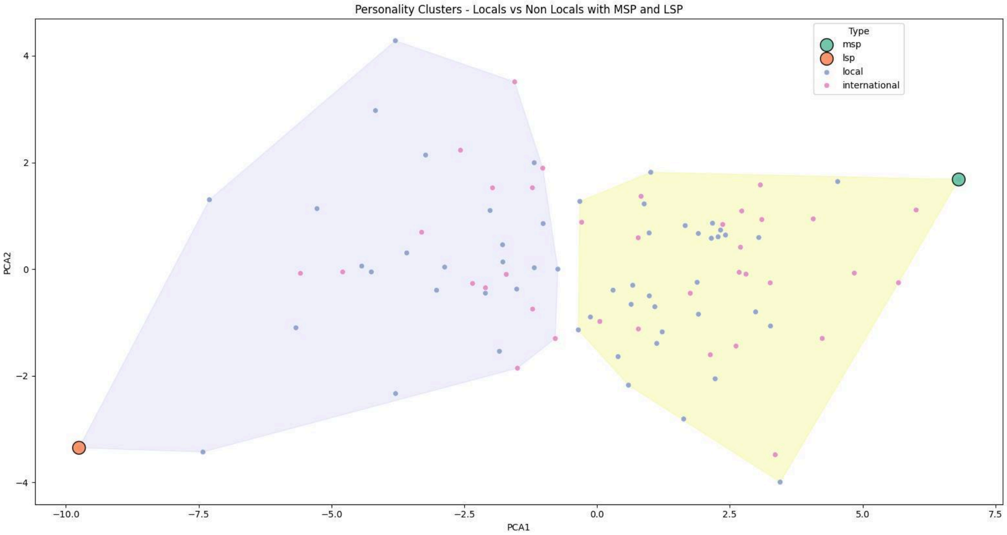
PCA works by identifying the directions (or principal components) along which the data varies the most. These components are used as the new axes, with the first principal component explaining the most variance, and the second component explaining the next highest amount of variance. When we plot the data using these two components, we effectively capture the most important aspects of the variation in social behavior.



To give more meaning to the clusters and to interpret the results in a clearer way, we introduced two control figures:

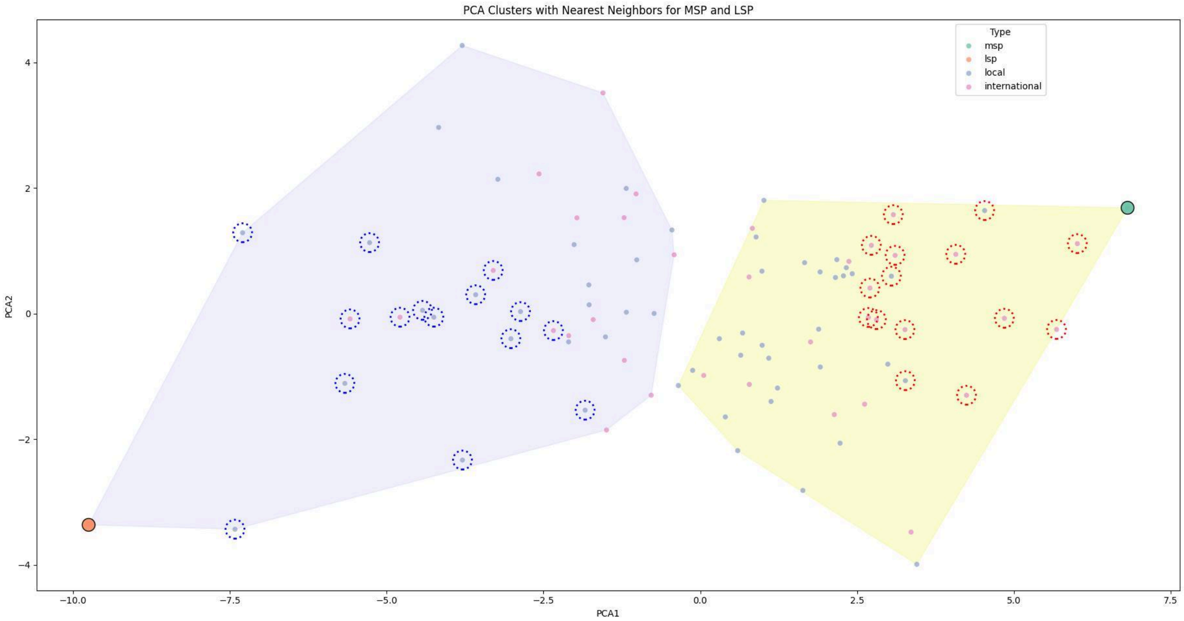
- MSP (Most Social Person): This hypothetical participant was constructed by giving them a score of 5 on all 19 questions. MSP represents the ideal case of an individual with high social self-efficacy.
- LSP (Least Social Person): In contrast, LSP was designed with a score of 1 on all 19 questions, representing the case of an individual with low social self-efficacy.

This 2D plot helps us visualize how participants are grouped in relation to MSP and LSP. The distance between participants on this plot reflects the degree of similarity or difference in their responses.

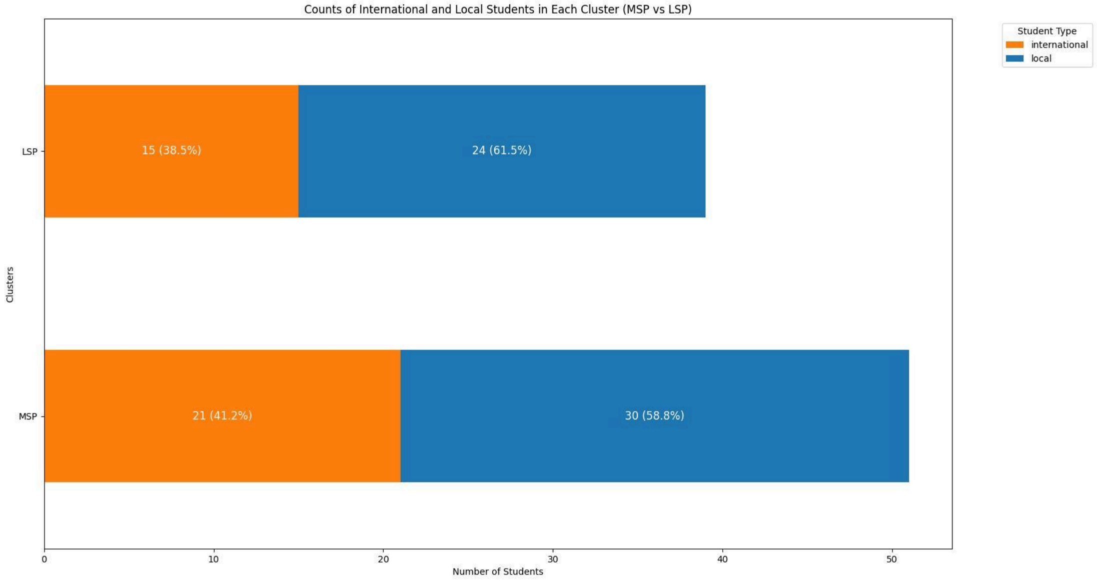


Closer Examination of clustering results of Social Self-Efficacy

To further refine our understanding of the participants, we identified the 15 participants closest to MSP and LSP by the euclidean distance. We circle those nearest 15 to MSP and LSP by blue color. This allowed us to explore the most extreme cases in the dataset: Participants closest to MSP can be classified as having high social self-efficacy. They responded similarly to the MSP control figure, indicating high levels of social comfort, engagement, and efficacy. Conversely, Participants closest to LSP indicate lower social engagement and possible discomfort in social situations.



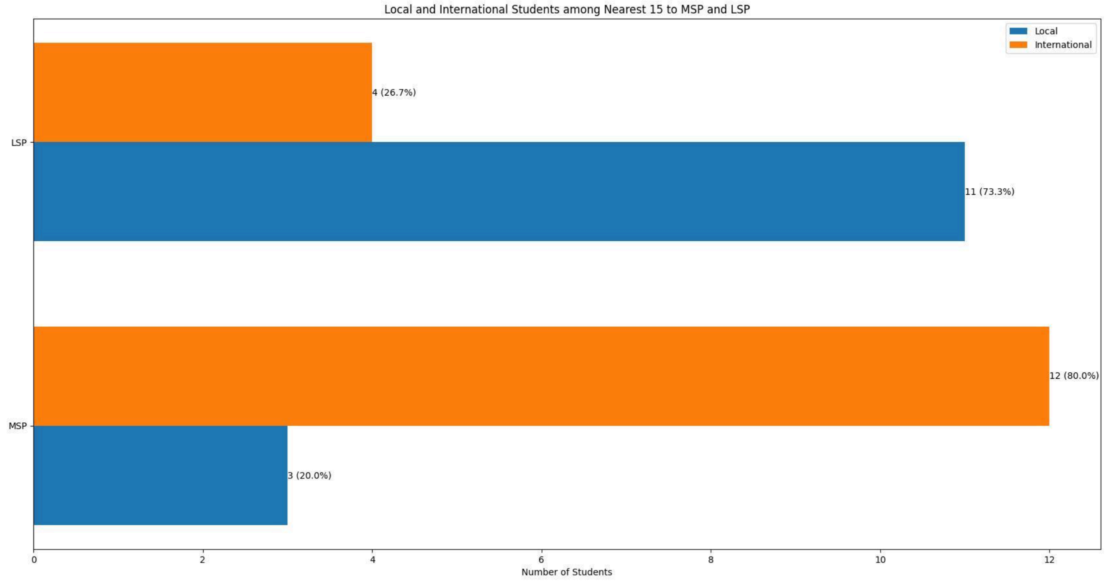
Counts of local and international in each cluster image



the overall distribution of local and international students in each cluster showed:

- MSP Cluster: There were 22 international students and 31 local students, meaning the MSP cluster, which represents the more socially efficacious group, had a larger proportion of local students, though international students still formed a significant portion of this cluster.

- LSP Cluster: There were 14 international students and 23 local students, indicating that the less socially efficacious group had a higher concentration of local students compared to international ones.



Findings

As expected, the participants clustered around the intersection of the two groups form the majority of the sample. This suggests that most respondents are socially competent but not extreme in their behavior. They may be comfortable in social situations but are not highly extroverted or socially dominant. This middle group represents a balanced approach to social engagement. The clustering analysis provided insightful data about the proximity of local and international students to two control participants: the most social person (MSP) and the least social person (LSP). The results revealed that 11 of the 15 nearest neighbors to MSP—a control person who scored the highest in social engagement—were international students, while 12 of the 15 nearest neighbors to LSP—who scored the lowest on social engagement—were local students.

Additionally, the overall distribution of local and international students in each cluster showed:

- MSP Cluster: There were 22 international students and 31 local students, meaning the MSP cluster, which represents the more social or extroverted group, had a larger proportion of local students, though international students still formed a significant portion of this cluster.

- LSP Cluster: There were 14 international students and 23 local students, indicating that the less social or introverted group had a higher concentration of local students compared to international ones.

It is important to note that this study does not suggest that all local students are socially inactive. In fact, there are more local students than non-locals in the socially active MSP group, showing that many locals are indeed socially engaged. However at the same time, there is also a high number of local students in the less social LSP group. What makes this study particularly interesting is the finding that, although local students are well-represented in both clusters, the 15 closest to MSP are predominantly international students, while the 15 closest to LSP are mostly local students. This is a very intriguing result that warrants further research to better understand the factors contributing to this phenomenon.

Limitations

The reliability of these findings is impacted by several limitations. One of the key challenges is the uneven ratio of local and international students in the survey sample, which could skew the clustering results. Additionally, the analysis only included a small portion of the CUHK student population, limiting its generalizability. A more balanced and comprehensive dataset that includes a broader range of students would yield more robust conclusions.

Conclusion

Despite these limitations, the analysis offers valuable insights into the social dynamics at CUHK. Along with the language barrier, which has already been identified as a major challenge for international students, this study suggests that personality traits, particularly introversion and extroversion, may also play a key role in shaping social interactions and integration. The MSP cluster, with its 58% share of the study population, shows that both local and international students are socially active. However, the concentration of international students in MSP's nearest 15 and local students in LSP's nearest 15 is a particularly intriguing finding that deserves further study to explain this pattern.

Further research is needed to explore these findings with a more representative sample, but the current analysis highlights the potential influence of personality in the integration process. The next steps would involve further analysis to see if certain demographic factors (such as age, gender, or cultural background) might explain some of the differences in social behavior. Understanding how personality traits interact with cultural and language barriers could lead to more effective strategies for fostering inclusivity and improving integration at CUHK.