Whose Experience is it Anyway?
Examining inter-subject variability in urban beauty and safety judgements

Rohit Priyadarshi Sanatani
Norman B. Leventhal Center for Advanced Urbanism (LCAU),
Massachusetts Institute of Technology (MIT)

Abstract

With rapid developments in urban visual analytics, recent research has focused on crowdsourcing judgements of urban qualities such as beauty and safety, for automated city-wide predictive evaluation. This study examines the extent and nature of inter-subject variability in such judgements and argues that these subjective qualities often have low generalizability across individuals. We conducted an online study involving 94 participants across 19 countries, where subjects arranged streetscape scenes from 4 global cities on visual scales of beauty and safety. A quantitative analysis of the arrangements revealed very low inter-subject consistency, including within demographic groups based on age, sex, race and nationality. K-means clustering of the arrangements also revealed large clusters of contradictory judgements with respect to urban visual features. There was, however, higher *intra*-subject consistency when rating the same scenes twice. Based on these findings, we recommend a cautious approach to the use of "average" crowdsourced judgements of subjective qualities, and encourage the adoption of subject-specific prediction when evaluating such qualities at scale.

Keywords: urban experience, beauty, safety, inter-subject variability

Introduction

Different urban environments have very different social and psychological impacts on citizens, just as different neighborhoods exhibit unique identities, urban-character and perceptual and emotional qualities. Through the 20th century there have been many notable approaches towards empirically analyzing the relationships between urban spaces and lived experience. Such studies have relied on interviews (Lynch, 1960; Nasar, 1990), manual observation/behavioral mapping (Gehl, 1971) and analyses of videography conducted in public spaces (Whyte, 1980).

With the advent of the internet and the development of large-scale data-collection methods, many studies over the past decade have focused on crowdsourcing urban experience judgements. These have included large-scale surveys involving evaluations of thousands of geolocated urban scenes along urban perceptual qualities such as safety, uniqueness, beauty, liveliness, wealth, boringness and the like (Salasses et al. 2013; Dubey et al., 2016). Ranking algorithms have been used to generate scores for each scene for each of these qualities. Such crowdsourced perception-scores have also been used to train machine-learning models for automated large-scale evaluations of cities based on street view imagery. This has allowed for the generation of predictive perception maps for qualities such as beauty and safety in different cities across the world (Naik et al., 2014; Porzi et al., 2015; Dubey et al., 2016; Zhang et al. 2018; Min et al., 2020).

While such lines of inquiry have focused on finding generalizable patterns in urban experience judgements, there have been very few approaches focused on mapping and investigating inter-subject differences for the same. Perceptual qualities such as beauty and safety can differ widely across individuals. The same environments may incite very different lived experiences in inhabitants – ones that are uniquely personal and subjective (Picon and Ratti, 2019). There is thus a gap between "average" perception scores computed from large-scale perception studies, and the rich spectrum of diverse experiences that characterize the urban. As a result, there may also be significant differences between individual experiences and predictive evaluations carried out by models trained on such crowdsourced datasets.

It is difficult to examine inter-subject variability in such datasets, since different subjects have not evaluated the same set of scenes under the same experimental conditions. Moreover, demographic attributes of the subjects themselves are not always collected, making the analysis of the impact of such factors difficult. Also, most past studies have relied on individual image ratings or pairwise comparisons. Qualitative judgements may be very different when made in the context of several other scenes. In this regard, recent studies in cognitive science have used semantic distance estimation tasks (Kriegeskorte & Mur, 2012) to capture subjective mental representations of various stimuli. For example, studies examining politically charged English nouns have revealed that supervised models trained on such representations can be used to accurately identify the political position of individuals (Li et al., 2017). Representations of abstract words - such as 'scenery' - have also been found to have a much greater inter-subject variability as compared to concrete words with direct referents - such as 'washing machine' (Wang & Bi, 2021).

In this work, we conducted an online web-based study involving contextual beauty and safety judgements of 28 streetscape scenes across 4 global cities. Based on the analysis of the data from 94 participants across 19 countries, we found very high-inter-subject variability across participants for both qualities in all cities. Using Principal Component Analysis (PCA) and k-means clustering, we examined the qualitative nature of this variability, and also the rationale of beauty/safety judgements for different clusters with respect to urban features.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Methods: Examining inter-subject variability

Study Design and Data Collection

A study aimed at measuring subjective difference in urban perceptions needed to be designed very differently from one aimed at analyzing common patterns across subjects. While past work focusing on the latter have used randomized images from a larger set as well as random sequences of presentation, our study used a smaller but fixed set of images presented in the exact same order and focused on the differences in subject responses that nevertheless emerged.

Image Selection: We selected four cities across different geographical contexts for this study – New York, Boston, Amsterdam and Mumbai. For each city, we relied on a dataset of Google Street View (GSV) images downloaded at 250m intervals across the street-network of the cities. The dataset comprised 4 images for each location, corresponding to headings of 0, 90, 180 and 270. We used a Dense Prediction Transformer (DPT) (Ranftl et. Al, 2019) trained on the ADE20K dataset (Zhou et al., 2017) for pixel-level segmentation of each image into 150 urban classes such as *tree*, *road*, *building*, *vehicle* etc. We then computed the percentage of each class for each image, and used the resulting 150-dimensional vector as a representation of the high-level visual quality of that image. We then used Principal Component Analysis (PCA) to compute the first principal component for the dataset of visual features for each city, and then extracted 7 images from equally spaced quantiles along this dimension. This allowed us to capture the spectrum of variance in visual qualities for each city. We thus used 7x4=28 images for this study **(Figure 1)**.

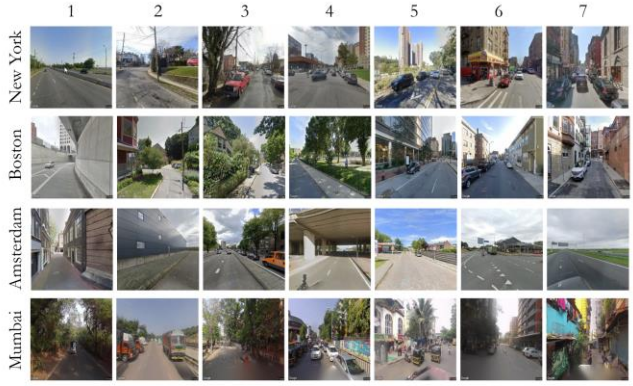

Figure 1: Urban streetscape scenes used for the study

Web Interface: For the online study, we designed a web application using HTML, CSS and JavaScript (front-end) and Flask (back-end). The interface featured an initial instructions page, followed by 10 sequential screens. Each screen presented the 7 scenes for one city at the bottom, and required the subject to drag and arrange the scenes on a visual scale of beauty or safety based on their own subjective notions of these qualities. A color gradient was provided as a reference backdrop, bounded by red (low) to the left and blue (high) to the right **(Figure 2)**. While judging a specific scene, subjects were asked to assume that they were physically standing on the ground and looking into the urban scene. This was to prompt them to judge the scene captured in the image, and not the photographic quality of the image itself. Right clicking on an image would enlarge it for more detailed evaluation.

The first 4 screens focused on beauty judgements of the 4 cities, while the next 4 focused on safety judgements. Finally, as a consistency check, both judgements for Set 1 (New York) were repeated as the final 2 screens. The X and Y screen coordinates of each image for each screen for a given subject were recorded, along with the screen height and width of the subject's device.

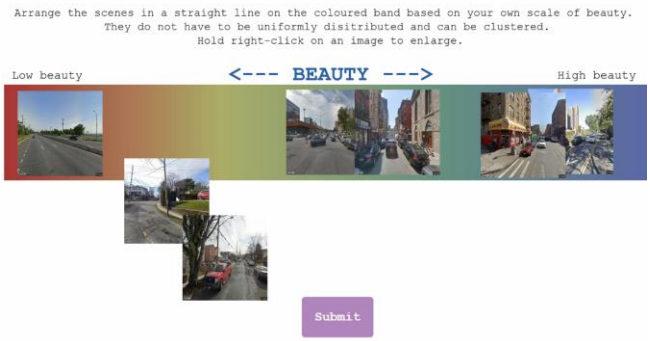

Figure 2: The user interface designed for the study

Subjects: We used the online platform Prolific (Palan and Schitter, 2018) to recruit subjects for this study. The primary qualifying criteria for recruitment was proficiency in English. The platform also ensured that participants completed the study on a laptop or desktop computer and not a mobile phone or tablet. A total of 113 participants began the study, while 10 withdrew without completing all screens. Another 9 participants were rejected by the authors because they either failed attention checks (did not arrange the scenes on the scale), used a device with a screen width lower than 1000 pixels, or took too long to complete the study. The data for 94 participants was used for the final analysis. The median time for study completion was ~12 minutes, which was within the expected range as per instructions provided. Participants were compensated at $9 per hour.

Demographic metadata for each participant was collected by Prolific. 59% of these participants were female, while 41% were male. Their ages ranged from 18 – 76 with a mean age of 34. 54% of participants were Black, while 31% were White, and the remaining were Asian, Mixed or others. They hailed from 19 nationalities, the largest groups being from South Africa followed by the United States.

Data Analysis

Quantification: We quantified a subject's judgement of each scene for a specific quality (beauty/safety) as the distance along the X axis between the origin of the image and the origin of the scale on which the arrangement took place. This value was normalized by screen width to ensure consistency across different devices. The representation of a subject's

overall perception of that quality in the context of a given set (city) was computed as a 7-dimensional vector, with the normalized distance of each image from origin as the value for each dimension. We thus computed 10 separate vectors for each subject, corresponding to the 10 judgement tasks that the subjects had completed. Finally, to quantify inter-subject variability, we computed the Intraclass Correlation Coefficients (ICC) (Bartko, 1966) across subjects for each city and quality.

Visualization: We ran Principal Component Analysis (PCA) on all responses for each city, and extracted the first 2 principal components. We then generated 2D scatter plots to visualize the variability in beauty and safety judgments for each city. We also colored the plots by demographic attributes, to examine possible trends based on factors such as age, sex, nationality or ethnicity. Moreover, we visualized the individual arrangements produced by subjects sampled at equally spaced quantiles along the first principal component for each dataset. This was done to reveal the qualitative nature of the variability in urban experience judgements.

Cluster Analysis: We then used k-means clustering to identify notable clusters of subjects that had similar beauty or safety judgements for a given city. The number of clusters was decided by plotting the Within-Cluster-Sum-of-Squares (WCSS) for various values of k, and using the elbow method to decide on an optimal value. After applying this method to all sets, we found that a value of k=5 was optimal. We also computed Silhouette scores (Rousseeuw, 1987) for each set to estimate the quality of clustering.

To examine the qualitative nature of urban judgements corresponding to each cluster, we computed the cluster centroids and identified 3 subjects from each cluster that were closest to the centroid. We then visualized the actual arrangements for these subjects. We also examined the urban feature distributions (based on pixel segmentation as discussed earlier) corresponding to each image in an arrangement, and examined their correlations with beauty/safety judgements within that cluster.

Inter-context and intra-subject consistency: Finally, we examined the extent to which relative arrangements across subjects were consistent across cities (inter-context), and the extent to which individual subjects produced consistent judgements for the same city presented at different points in the study (intra-subject).

For the former, we generated concatenated vectors that quantified the judgements of all subjects for a given city-quality combination. We then computed the Root Mean Squared Errors (RMSEs) as well as the Pearson's correlation coefficients between these vectors for every city and quality pair. These were used to assess consistency in judgments.

For the latter, we computed the Pearson's correlation coefficients between beauty/safety assessments for New York and the repeat assessments conducted towards the end of the study. We also computed the RMSE values between the two arrangements for each subject and the overall RMSE values for the two qualities.

Results: The Rich Spectrum of Urban Experience Judgements

Judgements revealed high inter-subject variability

Our results revealed very high inter-subject variability in the assessments of beauty and safety for all 4 cities. As we show in **(Figure 3),** despite being presented with the same images in the same order, different subjects made very different and often contradictory judgements for the same perceived quality. Intraclass Consistency Coefficients were very low, ranging from .07 (New York) to 0.26 (Boston) for beauty,

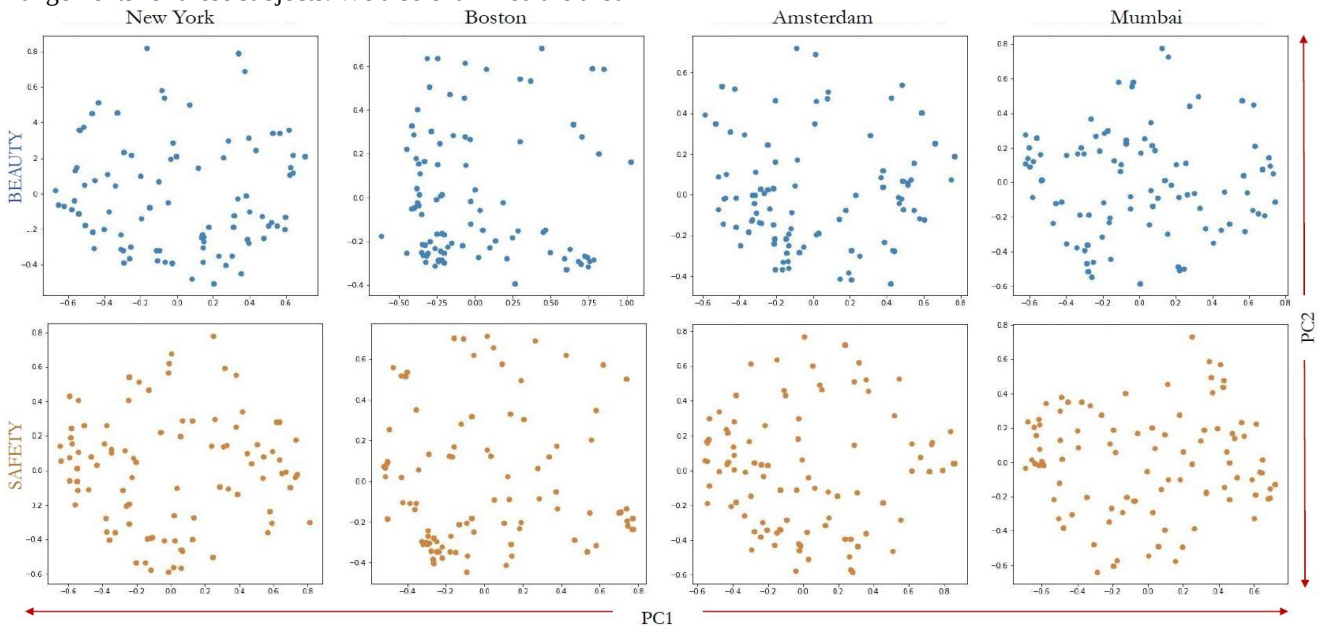

Figure 3: 2D representations of beauty and safety judgements using PCA (N=94)

and .09 (New York) to 0.21 (Boston) for safety. The visualizations also reveal that while there were some dense regions of weak consistency, there was very high dispersion overall. This was consistent with our initial hypothesis that such perceptual qualities are subjective and are influenced by numerous factors beyond the environments themselves. We also found that specific types of scenes like highways (New York 1, Amsterdam 7) presented the highest variance, while others featuring blank facades (such as Amsterdam 2) were more consistently assigned low beauty and safety ratings.

The principal components of variability: We investigated the qualitative nature of the first principal component (PC1) of the judgements for each city and quality. Our main finding was that the variability for both beauty and safety was expressed closely in terms of the variance in urban features for the cities themselves. This meant that in most cases, the scenes which were situated at the quantile extremes in terms of urban features in a given city (scene 1 and 7), were also situated at the extremes in terms of human judgements of beauty and safety. However, their positions were interchanged gradually along the principal component of the variability of human judgements for that city.

For example, in Amsterdam, one extreme end of safety judgements situated scene 1 (narrow dense street with low sky view) as very unsafe, and scene 7 (wide highway) as very safe. The other end of the spectrum of variability presented an opposing view, with scene 1 being rated as the safest, and scene 7 the most unsafe **(Figure 4)**. A very similar pattern of reversal was observed for beauty judgements in Mumbai with scenes 1 (dense narrow lane) and 7 (lush green avenue).

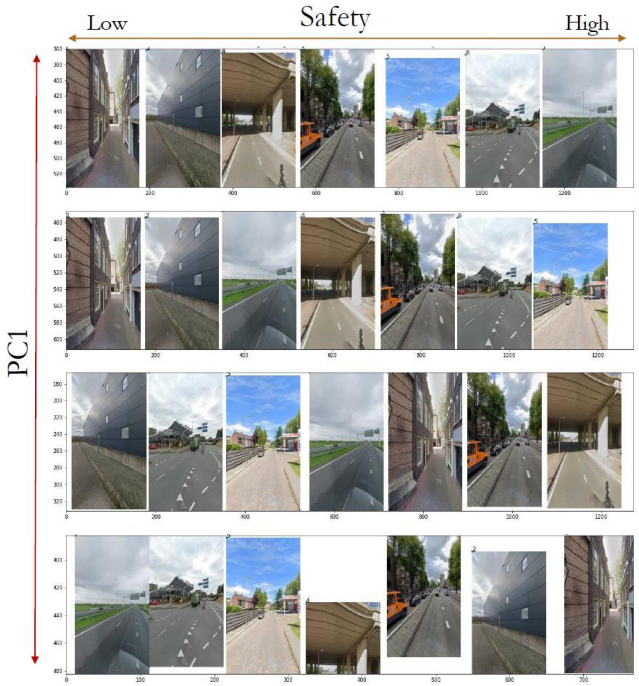

Figure 4: Top to bottom – equal quantiles for variability in safety judgements in Amsterdam.

The (non) demographics of variability: Further analysis based on demographic attributes also revealed that, contrary to common assumptions, the variability observed could not be explained through any of the attributes recorded for this study. While we had expected certain outcomes, such as clustering based on biological sex or age in perceptions of safety, such patterns were not observed in our data **(Figure 5)**. Similarly, we also did not find any significant clustering based on nationality or language – both strong indicators of cultural backgrounds. The principal components of variability analyzed also did not correlate with any of the demographic attributes studied.

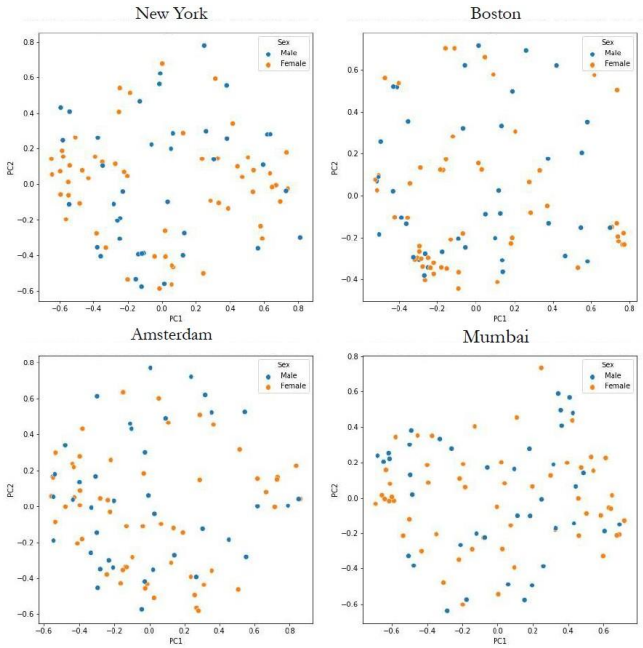

Figure 5: We found negligible clustering of safety judgements based on sex

Overall, while a larger study may be warranted for more confident generalizations, these findings provide strong empirical indications that perceptual qualities such as beauty and safety are extremely subjective, and that randomized crowdsourced judgements across multiple subjects may not be representative of individual lived experiences.

Cluster analysis revealed groups of contradictory judgements

K-means clustering on the data for different cities reveled weak clustering. Silhouette scores ranged from 0.16 (Beauty/Amsterdam) to 0.24 (Beauty/Boston). In line with our analysis of the principal components of variance, the data also revealed clusters of opposing beauty/safety judgements.

For example, for beauty judgements in Boston, we observe two clusters (C and D) at two ends of PC1 **(Figure 6 - top)**. Samples from cluster C reveal judgements where dense built environments are assigned low beauty and green suburban scenes are assigned high beauty. For cluster D, this pattern is reversed. This is trend is confirmed through a quantitative examination of urban features, where buildings and greenery

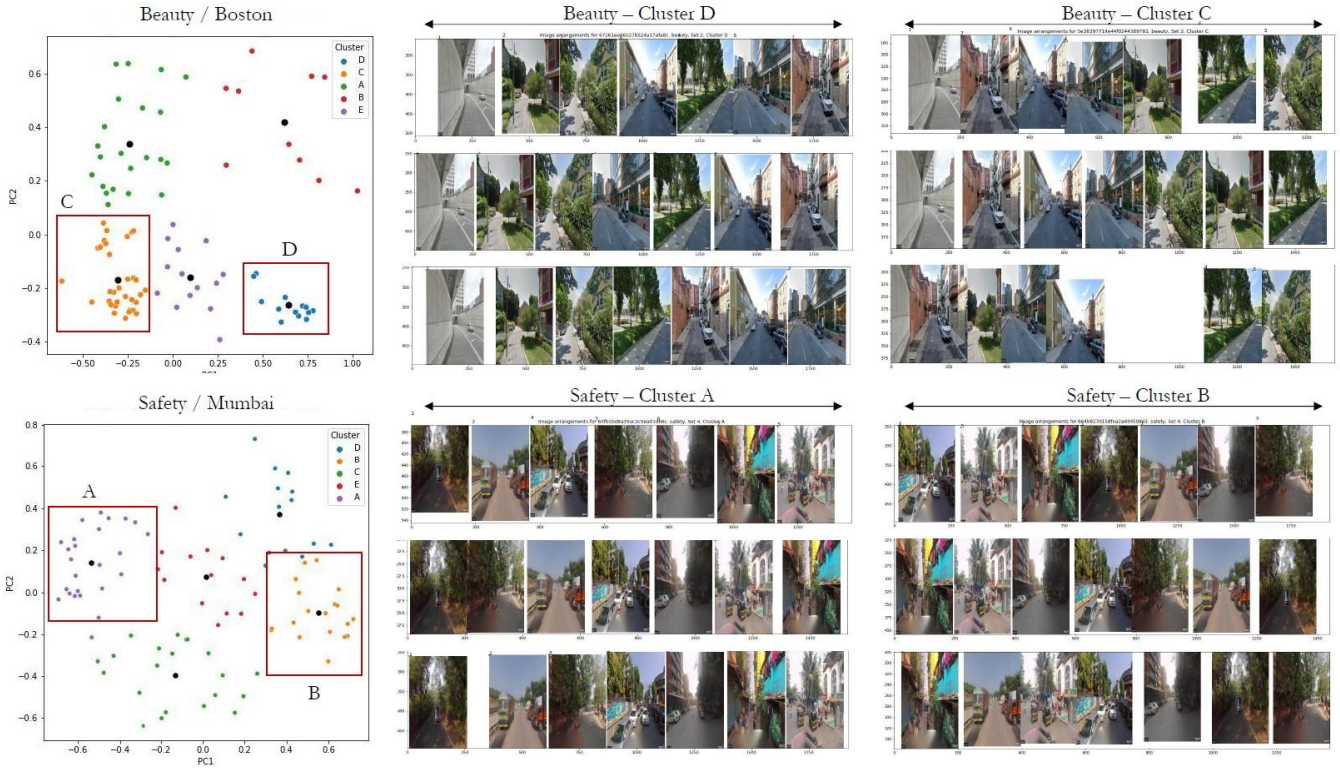

Figure 6: Clusters of opposing judgments for beauty in Boston (top) and safety in Mumbai (bottom)

exhibit correlations coefficients of $R_{building}$ = -0.48 and $R_{greenery}$ = 0.78 respectively for cluster C, but $R_{building}$ = 0.67 and $R_{greenery}$ = -0.22 respectively for cluster D. For both clusters however, a concrete tunnel scene was assigned the lowest beauty. Similarly, for safety in Mumbai **(Figure 6 - bottom)**, samples from cluster A assigned lower safety to open green environments ($R_{greenery}$ = -0.51) and higher safety to denser streetscapes including those with traffic ($R_{building}$ = 0.5). This trend was reversed in cluster B, with $R_{greenery}$ = 0.66 and $R_{building}$ = -0.45.

Similar contradictory judgements were revealed throughout our dataset, and indicated that over and above the high inter-subject variability overall, there were weakly defined clusters of judgements that opposed each other.

Judgements revealed relatively higher *intra*-subject consistency within a set

In contrast to the high inter-subject variability observed for judgements for a specific city, we found a relatively higher *intra-subject* consistency for the experience evaluations. We observed an average RMS error of 0.278 and 0.228**,** and a statistically significant correlation R = 0.38 and 0.55 for beauty and safety judgements respectively between original and repeat evaluations of New York across all subjects. As expected, the correlation was much higher for safety as compared to beauty, indicating that subjects made more consistent safety judgements when evaluating the same set twice, as compared to their beauty judgements for that set. **Figure 7** visualizes the correlations across all subjects between initial and repeat judgements for New York.

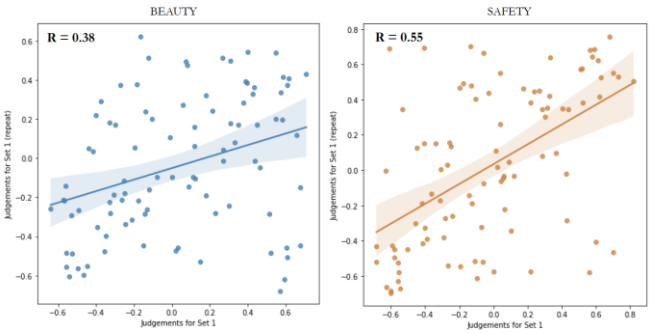

Figure 7: Intra-subject consistency for beauty (left) and safety (right) judgements for New York.

Judgements revealed high inter-quality consistency

Finally, our analysis revealed that there was high consistency between judgements of beauty and safety within a same set and that the arrangements produced by subjects when evaluating the two qualities in a set were very similar. **Figure 8** plots the R values for correlation between every city-quality pair. While the strongest correlations were of course between repeat judgements of the same quality in the same city, correlations between judgements for the same city for the other quality were within the same range (see highlighted diagonals). This meant that a scene assigned high beauty was also likely to be assigned high safety and vice versa. Judgements for the same quality across different cities were however much lower, and in many cases did not exhibit any correlation at all.

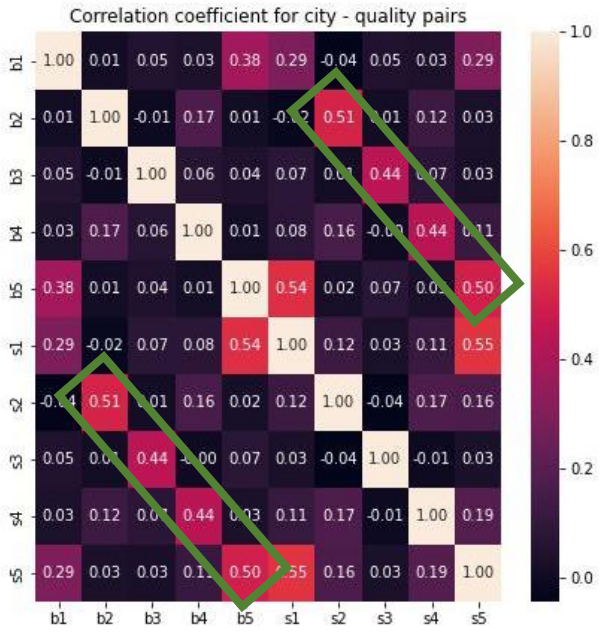

Figure 8: Correlation matrix for judgements across city and quality (b = beauty; s = safety)

Discussion: Whose experience is it anyway?

This study undertook one of the first empirical inquiries into the nature and patterns of inter-subject variability in urban beauty and safety judgements. The key finding that emerged was that while individual subjects made reasonably consistent judgements for specific sets of urban scenes, there was nevertheless very high variability across participants – one that did not correlate with any of the major demographic attributes studied. Moreover, there was a reasonably high correlation between beauty and safety judgements within urban scene sets – significantly higher than the correlation across sets. We also found that the largest principal component of the variability in beauty/safety judgements was expressed in terms of the largest principal component of urban features that characterized the scenes.

These findings have several implications for research into urban cognition. As argued at the outset, these results provide empirical validation for the high degree of subjectivity that characterize urban experience judgements. The same urban scenes may be perceived very differently by different individuals, and qualitative judgements may also be very different and often contradictory. The image regions and feature dimensions that different participants focus on while making such judgements may also be different. As a result, large-scale crowdsourced approaches to understanding subjective qualities may be limited in their scope. It may be important to note that predictive models trained on such crowdsourced datasets in past studies reflect judgement metrics computed across different subjects – each of whom may have had different ways of perceiving the same space.

The high correlation observed between beauty and safety may indicate that the two qualities indeed share some inherent similarities in an urban context, but may also point to a more general underlying mental representation of urban scenes that result in such similarities. The fact that the first principal component of urban features – especially the extremes - influence such judgements indicate that subjects' mental representations may be expressed in terms of this component and also influence general "like–dislike" judgements of scenes which precede the appraisals of beauty or safety.

While these results are interesting, it is worth discussing some limitations of this work from a methodological standpoint. Firstly, our current findings are based on the data of 94 participants. While our primary aim was to provide an exploratory analysis of the nuances of urban subjectivity, a larger study may be required for more confident generalizations, and for stronger validation of many of our key findings. Secondly, there is always a gap between judgements of images of scenes, and judgements of the scenes themselves. While we had explicitly asked subjects to evaluate the scenes and not the photographs, our results may nevertheless not generalize beyond images. Moreover, such judgements are also often influenced by senses beyond vision – such as sound and smell – which lie beyond the scope of this current work. Thirdly, for the repeat evaluations of New York, it was difficult to differentiate between arrangements made based on conscious re-evaluations and those made from memory. A part of the intra-subject consistency may possibly be owed to this factor. Fourthly, while the structure of this remote study was identical for all subjects, it was not possible to control the precise experimental conditions at the subjects' end. Extraneous factors may thus have contributed to some of the inter-subject variability observed. Finally, the difference in consistency for scenes within and across sets indicate that the results are sensitive to scene-selection. However, this also indicates that certain kinds of scenes may be better suited than others to examine inter-subject variability. A more detailed investigation into the nature of such scenes may be valuable for future studies.

In conclusion, one can speculate on the implications of these findings for future methods of empirical urban experience evaluation. Given the high variability observed in our data, one may want to train subject-specific models that predict urban qualities for specific individuals or groups, based on their own unique patterns of judgement. Reinforcement learning may be used to continuously learn the nuances of a single subject's urban experience, or few shot learning methods may be employed to infer an individuals' judgement based on relatively few samples. Such models may be used for 'subjective' experience evaluation in urban analytics, where the results of such analyses change depending on the target user and also at different points in time for a given user.

Urban environments are complex, and subjective lived experiences characterize the urban condition in fundamental ways. This work is positioned as a step towards empirical approaches to the subjective realm. Because at the end of the day, whose experience is it anyway?

References

Bartko, J. J. (1966). The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep Learning the City: Quantifying Urban Perception at a Global Scale. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (Vol. 9905, pp. 196–212). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_12

Gehl, J. (2011). *Life between buildings.*

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00245

Li, P., Schloss, B., & Follmer, D. J. (2017). Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior Research Methods*, *49*(5), 1668–1685. https://doi.org/10.3758/s13428-017-0931-5

Lynch, K. (2008). *The image of the city* (33. print). M.I.T. Press.

Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). Streetscore—Predicting the Perceived Safety of One Million Streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 793–799. https://doi.org/10.1109/CVPRW.2014.121

Nasar, J. L. (1990). The Evaluative Image of the City. *Journal of the American Planning Association*, *56*(1), 41–53. https://doi.org/10.1080/01944369008975742

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Picon, A., & Ratti, C. (2019). Mapping the future of cities: Cartography, urban experience, and subjectivity. *New Geographies*, *9*, 62–65.

Porzi, L., Rota Bulò, S., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. *Proceedings of the 23rd ACM International Conference on Multimedia*, 139–148.

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The Collaborative Image of The City: Mapping the Inequality of Urban Perception. *PLoS ONE*, *8*(7), e68400. https://doi.org/10.1371/journal.pone.0068400

Wang, X., & Bi, Y. (2021). Idiosyncratic Tower of Babel: Individual Differences in Word-Meaning Representation Increase as Word Abstractness Increases. *Psychological Science*, *32*(10), 1617–1635. https://doi.org/10.1177/09567976211003877

Whyte, W. H. (1980). *The social life of small urban spaces*. Conservation Foundation.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., & Torralba, A. (2017). Scene parsing through ade20k dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 633–641.