Perceptual Acuity in the Face Space: Organization in Machine Face Recognition is Perceptible by Humans

Blake Moya

The University of Texas at Dallas

Abstract

The face space model of human face perception describes a perceptual organization of faces in which individual faces can be represented as a point in a space, and similarity judgements among different faces are made based on the distance among those faces' representations in that space. Computer algorithms designed to emulate human face recognition and individuation utilize a face space model for their computation. Some faults in human face perception, such as the Other Race Effect, also manifest in these computer algorithms and it warrants asking whether faults in computer face perception may also manifest, yet unnoticed, in humans. To begin to answer this, the present study asked participants to sort faces along an order determined by a computer face recognition algorithm and measured how much these participants erred relative to participants asked to sort faces along a randomly ordered sequence. The study found a significant effect of sequence condition on participant's error in sorting faces. Participants erred less when sorting into the sequences ordered by the face recognition algorithm. This showed a link between the organizational method acquired by machine learning algorithms through training and the organizational method of the brains they were designed to emulate. This implies that flaws and features of computer face recognition can be used to pilot questions regarding human face recognition.

Perceptual Acuity in the Face Space

The ability to recognize and individuate faces is central to the human experience. From the moment children enter the world, their eyes will follow and fixate on any face-like stimulus that enters their field of view (Johnson, Dziurawiec, Ellis, & Morton, 1991). Beyond humans being innately able to recognize a face as a face, the importance of faces in the mind is further indicated by the special activation of the Fusiform Face Area (FFA) of the brain (Rhodes, Byatt, Michie, & Puce, 2006). This area activates upon exposure to upright face-like stimuli and helps explain the uniqueness of faces in a person's perceptual experience. The special attention granted to faces in human biology is extremely relevant to humans as social creatures, for many obvious reasons. Being able to recognize one's friends and families quickly and with confidence is a core foundation of the most basic social structures. Face preference, however, has more subtle impacts on social interactions. Irrelevant facial similarity to an underperforming employee can significantly damage an applicant's chances of being hired, despite relevant qualifications (Helverson, Herzog, & Rieskamp, 2014). Competency judgements based only on candidates' faces can predict the outcome of federal elections (Todorov, Mandisodza, Goren, & Hall, 2005). Whatever the brain's method for analyzing faces may be, the verdicts handed out by it carry serious weight in decision making.

In addition to the sometimes exaggerated influence of face recognition in flawed decision making, the method itself is also not without its flaws. Key among these flaws, as they relate particularly to face individuation, is the Other Race Effect (ORE). The ORE is a phenomenon in which a person of one race generally has more difficulty individuating faces of some other race than faces of his or her own race. Children under one year old do not exhibit the ORE and the effect manifests following underexposure to other race faces during development (Anzures et al.,

2013). This flaw is especially threatening when one recalls the effect of apparent similarity in employment decisions. An increased likelihood of finding other race faces similar implies that a hiring officer may be more likely to judge a candidate as similar to a previous employee of the same race and allow his or her opinion of that previous employee's performance to affect his or her opinions of the candidate. The implications of the ORE bring forward a great need for understanding the brain's method of face recognition and individuation. Although hardware limitations might prevent a full analysis of the workings of the FFA, there may yet be a viable alternative for the isolation and investigation of flaws and biases in human face perception: machine face recognition model analysis.

The ORE also exists in machine face recognition models and can be caused by underexposing the model to other race faces during training (Caldara & Abdi, 2006). This happens in machine models that utilize a face space for face recognition.

A face space is an abstract mathematical space in which any individual face can be represented as a single coordinate. Most face spaces are designed such that each axis, or dimension, measures some attribute of the faces it describes (i.e., low values on some axis indicate low cheekbones and high values on that axis indicate high cheekbones). No matter what attributes the axes of a face space represent, a machine model utilizing such a space will train in such a way that bends and reconfigures the space until it becomes easy for the model to individuate the faces it was trained on. Faces belonging to a race that the model has been underexposed to will be represented by a smaller region of the model's face space and are thus closer together in the space and less easily discriminable (Caldara & Abdi, 2006). Gao and Wilson (2013) showed that the face space model has a neural analogue by creating synthetic faces within a particular face space and showing that the neural response of participants to these

faces could predict where in the face space they resided. This gives some credence to the model on a mechanical level, but it should also be noted that a face space model can, beyond grouping faces into a single identity, predict attributes about faces that people find meaningful, like the presence of eye bags or a mustache (Zhong, Sullivan, & Li, 2016). This should give credence to the model on a semantic level as well. Machine models not only accurately recognize human faces but may also accurately represent the human brain's method for doing the same task. If there is a chance that flaws and features in machine models can be used to extrapolate some understanding of the brain's method—in the way that the face space model can explain the ORE for both processes—then these flaws and features should be catalogued, especially given the ease with which machine models can be analyzed compared to the brain.

To do this, I selected a specific face recognition model distributed in OpenFace, a python library from Amos, Ludwiczuk, and Satyanarayanan (2016). This model was based on an architecture developed by Schroff, Kalenichenko, and Philbin in 2015. The model takes an image of a face as input and outputs 128 attribute values that have a square sum of one. This effectively embeds the faces as points onto a 128-dimensional unit sphere. Every image containing a face can be mapped to a single point on this 128-sphere. Distance between points in this space, as is typical of a face space, are proportional to the similarity of the faces represented by the points (Schroff et al., 2015). Because of this, any line radiating in any direction from a point representation in this face space can be understood as being one direction or attribute upon which faces can differ in that space. If these attributes are meaningful to humans, then it would imply that the organization of the face space generated by the model could be analogous to the organization of faces in the brain, given that both humans and the models they have designed will have selected the same attributes upon which to individuate faces through evolution and

training. This would open new possibilities for detecting flaws and phenomena in the human method by first discovering them from machine analysis and then testing for their presence in humans. The present study sought to test the assertion that the organization learned by the selected model is meaningful or perceptible to humans as a first step toward a paradigm for extrapolating mental phenomena from easier to analyze machine models.

Methods

Participants

Participation in the survey included 46 adults from The University of Texas at Dallas between the ages of 18 and 29. Sixteen (35 %) of participants identified as male, 28 (61%) identified as female, one (2%) identified with an unlisted gender identity, and one (2%) did not respond. Twenty-five of the participants identified as White, one identified as American Indian of Alaskan Native, 18 identified as Asian, 3 identified as Latino or Hispanic, one identified as an unlisted race, and one did not respond. Percentages for race are unlisted because participants were able to identify with more than one race.

Participants were recruited via an email containing the survey link sent to their associated university email address, and by a flyer with a scannable Quick Response (QR) code containing the survey link posted around campus. No compensation was given for completing the survey. Participants gave informed consent before beginning the survey and all participation was voluntary.

Materials

Perceptual acuity was measured with a face sorting task that showed images generated using the model being investigated and the VGGFace2 data set (Cao, Shen, Xie, Parkhi, & Zisserman, 2018). The model takes as input an image containing a face and outputs a vector with

128 entries, each representing an abstract and nameless attribute (Amos et al., 2016). The data set contains over 9000 folders each representing a unique individual, and each containing images of that individual's face (Cao et al., 2018). Using images from the data set and those images' representations from the model, I generated sequences and unsorted faces to be used in the face sorting task.

Sequences. Sequences appear as an ordered list of six faces with integers one through five denoting the spaces between. To create these sequences, vector representations of 100 randomly selected images from 600 randomly selected folders (each representing one unique individual) in the data set were calculated, and an average representation for each of the 600 individuals was recorded. These representations were then sorted by one of the 128 attributes. Before sorting the representations, six were randomly selected and tagged to become the unsorted faces for the task. The list of the 594 remaining representations was ordered into three different sequences. The sequence sorted by the 49th unnamed attribute of the vectors was named Sequence A. The sequence sorted by the 45th attribute was named Sequence B. A third sequence was shuffled randomly instead or sorted by any attribute (to serve as a control condition) and was named Sequence Z. The sorted lists were divided into six subgroups with each subgroup representing one sixth of the range of that sequence's attribute values. For Sequence Z, these subgroups represented an even split of the shuffled list, with 99 representations in each.

Average face images were then generated to represent each subgroup of each sequence. To do this, 25 representations from each subgroup were randomly selected. For each of these representations, four images were randomly selected from the folder in the data set from which that representation was computed. All these images (100 in total) for each subgroup were then combined into an "average face" image. These images (six per sequence) were then arranged in

the order they appear in the sequence and the spaces between them were labeled with integers one through five. Each of the sequences generate can be seen in Figure 1.

Unsorted Faces. The six representations tagged before creating the sequences were used to select 100 images from the folders in the data set from which each was computed. These 100 image sets were each combined into one average face of the individual represented by the vector. Each of the unsorted faces generated can be seen in Figure 2. Because these representations were calculated, their values on the 49^{th} and 45^{th} attribute of the model are known. Because these values were known, a correct position for each unsorted face exists in sequences A and B. For Sequence Z, "correct" placements were chosen randomly.

Procedure

Participants who opened the survey were presented with information about the study and were asked to give their informed consent. Participants who consented were then prompted to respond to three demographic inquiries regarding their age group, racial identity, and gender identity. After submitting this information, participants were randomly directed to one of three survey conditions: a condition in which the participant sorted faces along Sequence A (ordered by attribute 49), a condition in which the participant sorted faces along Sequence B (ordered by attribute 45), or a condition in which the participant sorted faces along Sequence Z (shuffled). Within a single branch, the participant was presented with the corresponding sequence and each of the six unsorted faces in random order. To sort a face, participants are asked to select which spot (labeled 1 through 5) the unsorted face best fit in the sequence. After doing this for each unsorted face, the survey ended.

Measures and Analysis

Participants responses were evaluated based on their average error. Sequence groups were divided at breakpoints defined by sixths of the range of scores on the attribute on which the sequence was sorted. Because of this, there was a breakpoint in each sorted sequence in which the unsorted face best fit: the breakpoint that represented a score nearest to the score of the unsorted face. Error was measured as a distance in intervals from the best fit breakpoint as determined by the unsorted faces attribute score (e.g., if an unsorted face was assigned by the model to fit best at breakpoint 3, and the participant selected breakpoint 4, that would be measured as an error of 1. If that participant had instead selected breakpoint 1, that would be measured as an error of 2). For the shuffled sequence (Z), the best fitting breakpoint for each of the unsorted faces was determined randomly, as if the unsorted faces had been shuffled in with the rest originally. Error was chosen to be an absolute distance so that averaging a participant's error scores would generate averages closer to zero for more accurate sorting. Data describing the amount of time participants spent answering each question was also recorded.

These data were analyzed using a one factor between-subjects analysis of variance (ANOVA) wherein sorting condition (sequence A, sequence B, sequence Z) was the between-subjects variable and average error was the dependent variable.

Results

A one factor between-subjects analysis of variance indicated a significant effect of sequence condition on participant's error in sorting the unsorted faces, $F(2, 43) = 71.763$, $MS_e = 0.007$, $p < .001$, partial $\eta^2 = .769$. Post hoc comparisons using Bonferroni correction showed that participant's shown sequence Z (the randomly sorted sequence) erred significantly more than those shown sequence A ($p < .001$) and those shown sequence B ($p < .001$). Participants shown sequence A (the sequence sorted by the predictor on which identities most varied) also erred

significantly less than participants shown sequence B (p = .002). Figure 3 visualizes these differences.

Discussion

A clear association was found between sorting condition and participant error, with participants erring less when sorting faces into a series aligned on an axis of the face space generated by the model. This indicates that the model learned an organizational method meaningful to humans and supports the face space model of human face perception, similar to that described in Valentine (2016) and observed by Gao (2013).

This is consistent with O'Toole, Castillo, Parde, Hill, and Chellappa (2018), who found meaningful organization in a reduced version of a face space generated by a similar neural network. O'Toole et al. found that more ambiguous images (in which faces were blurred or pointed away from the camera) were represented as points nearer to one another in the face space they were examining, and that less ambiguous images (in which faces were clear or facing directly toward the camera) were represented as points distant from representations of any other individual. This is an intuitive organization. It increases prediction confidence for less ambiguous images because representations of other individuals are so far away in the face space while it also decreases confidence for more ambiguous images because representations of one individual can be so close to another. Recall that in this face space, distance between point representations is proportional to visual similarity of the faces being represented (Schroff et al., 2018). The present study further confirms this assertion, in that the sequences of faces generated by Schroff's architecture seems to show visually consistent differences across an axis (Sequence A appears to show a consistent change in jaw height across its levels, and Sequence B appears to show a change in wrinkle depth, visible in Figure 1). This architecture was designed to be pose

and lighting invariant (meaning that no matter the angle of the face or the light source in an image, the model would classify it as the same identity), and the fact that its organization was found to be meaningful to human participants is consistent with the previous findings of pose and lighting tolerance in human face identification (Blank & Yovel, 2011). All these findings continue to fall in line with the face space model of human face perception, and the present study contributes by showing that computer implementations of this model can be understood by a human participant.

This study was certainly not without its flaws or without potential. One of the original hypotheses during the design of this studies was that the positions participants chose to sort faces into would follow a normal distribution for the sequences based on the model and a uniform distribution for the random sequence. This would support the idea that the randomly sorted sequence has no one spot that is better fitting than any other for the unsorted face, and that the sequences based on the model do (meaning that these sequences were sorted in some logical way). A Chi-square test for goodness of fit with these predicted distributions could have been used to test this. Foremost among the issues preventing me from using this design was a confound that ultimately became principal to the design that was used. When participants were asked to sort faces into the random sequence, they seemed to agree on particular breakpoints just as much as participants who were asked to sort faces into the sorted conditions. This seems to be because participants in the random condition would sort a face by putting it next to the randomly generated face that looks most like the unsorted face, without regard for the sequence as some sort of logical progression. This led to the distribution of responses seeming normally distributed but having extremely high error because the typical responses were not related with the

predetermined “correct” breakpoint. The design I chose to use, measuring participant error, was able to account for this fact, but it does distance the experiment from the original hypothesis.

Beyond this, the sequences generated could have been improved in a few ways. First, having more breakpoints would allow for a higher resolution distribution of responses, better able to pinpoint where exactly participants in general thought an unsorted face should fit. I was stopped from doing this because, unfortunately, the sequence needed to fit on a mobile screen to be accessible to the number of participants I wanted to recruit. An interesting way to overcome this pitfall of the digital age, if this mode of face space exploration is found to be useful, would be to replace sequences with an animation of a rolling average of faces across the sequence wherein participants could select a time at which the face presented looks most like the unsorted face. This would also allow for a much more continuous distribution of possible responses which would increase the validity of the results found.

I also considered dividing sequences by frequency rather than by interval (i.e., to assign an equal number of faces to each of the six groups rather than having each group consist of faces represented by one sixth of the range of that measurement). I ultimately rejected this on the grounds that it would cancel out the effects of density of faces in certain portions of the face space, which was found to be a good representation of similarity between faces (O’Toole et al., 2018; Schroff et al., 2015); however, it did lead me to consider using *k*-means clustering to divide sequences instead. Dividing sequences by cluster would preserve the effects of density but also draw clear lines between groups of faces that are distant from each other in the face space. This would have made each image in the sequences look more distinct from one another, perhaps strengthening the already highly significant results.

The final alteration I considered was to avoid average faces all together. Averaging images of faces together creates an average face that is more attractive than, as one might say, the sum of its parts (Halberstadt & Rhodes, 2003). This could unnecessarily alter the experiment by filtering out unattractiveness from sequences. To avoid this, I considered sacrificing some of the orderliness of the sequences in favor of bins or folders of raw images of faces belonging in each portion of the sequence. Considering this, an attractive prospect for future study would be to run average face images through the model and examine whether the point representation of an average face image is near the average of the point representations of its constituent images. This could offer further insight into the use of average face images as a means to interpolate latent regions of the face space. Another prospect of future study would be to manually alter certain continuous features in images of faces (such as increasing or decreasing jaw height or brow width) and repeatedly run those images through the model and record their changing measurements. One could then calculate a regression line in the face space that best represents the feature that was altered. This would show a valid method of mapping out the face space generated by this model for certain known features.

Despite its flaws, this study added to the accumulating evidence supporting the face space model of human face perception, and showed that mathematically efficient machine learning models can be used to deduce some of the functional properties of the brains they were designed to emulate.

References

Amos, B., Ludwiczuk, B., & Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. CMU School of Computer Science, 6.

Anzures, G., Quinn, P. C., Pascalis, O., Slater, A. M., Tanaka, J. W., & Lee, K. (2013). Developmental origins of the other-race effect. *Current Directions in Psychological Science, 22(3).* 173-178. doi: 10.1177/0963721412474459

Blank, I., & Yovel, G. (2011). The structure of face-space is tolerant to lighting and viewpoint transformations. *Journal of Vision, 11(8),* 1-13. doi: 10.1167/11.8.15

Caldara, R., & Abdi, H. (2006). Simulating the 'other-race' effect with autoassociative neural networks: further evidence in favor of face-space model. *Perception, 35(5).* 659-670. doi: 10.1068/p5360

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., Zisserman, A. (2018). *VGGFace2: A dataset for recognising face across pose and age.* Paper presented at the IEEE Conference on Automatic Face and Gesture Recognition. doi: 10.1109/FG.2018.00020

Gao, X., & Wilson, H. R. (2013). The neural representation of face space dimensions. *Neuropsychologia, 51,* 1787-1793. doi: 10.1016/j.neuropsychologia.2013.07.001

Halberstadt, J., Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review, 10(1),* 149-156. doi: 10.3758/BF03196479

Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelganger: Irrelevant facial similarity affects rule-based judgement. *Experimental Psychology, 61(1),* 12-22. doi: 10.1016/j.neuropsychologia.2013.07.001

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40(1-2).* 1-19. doi: 10.1016/0010-0277(91)90045-6

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Science, 22(9).* 794-809. doi: 10.1016/j.tics.2018.06.006

Rhodes, G., Byatt, G., Michie, P. T., & Puce, A. (2006). Is the fusiform face area specialized for faces, individuation, or expert individuation? *Journal of Cognitive Neuroscience, 16(2).* 189-203. doi: 10.1162/089892904322984508

Schroff, F., Kalenichenko, D., & Philbin, J. (2015, June 7-12). *Facenet: A unified embedding for face recognition and clustering*. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR.2015.7298682

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308(5728).* 1623-1626. doi: 10.1126/science.1110589

Valentine T., Lewis, M. B., Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology, 69(10).* 1996-2019. doi: 10.1080/17470218.2014.990392

Zhong, Y., Sullivan, J., & Li, H. (2016, June 13-16). *Face attribute prediction using off-the-shelf CNN features.* Paper presented at the International Conference on Biometrics (ICB). doi: 10.1109/ICB.2016.7550092

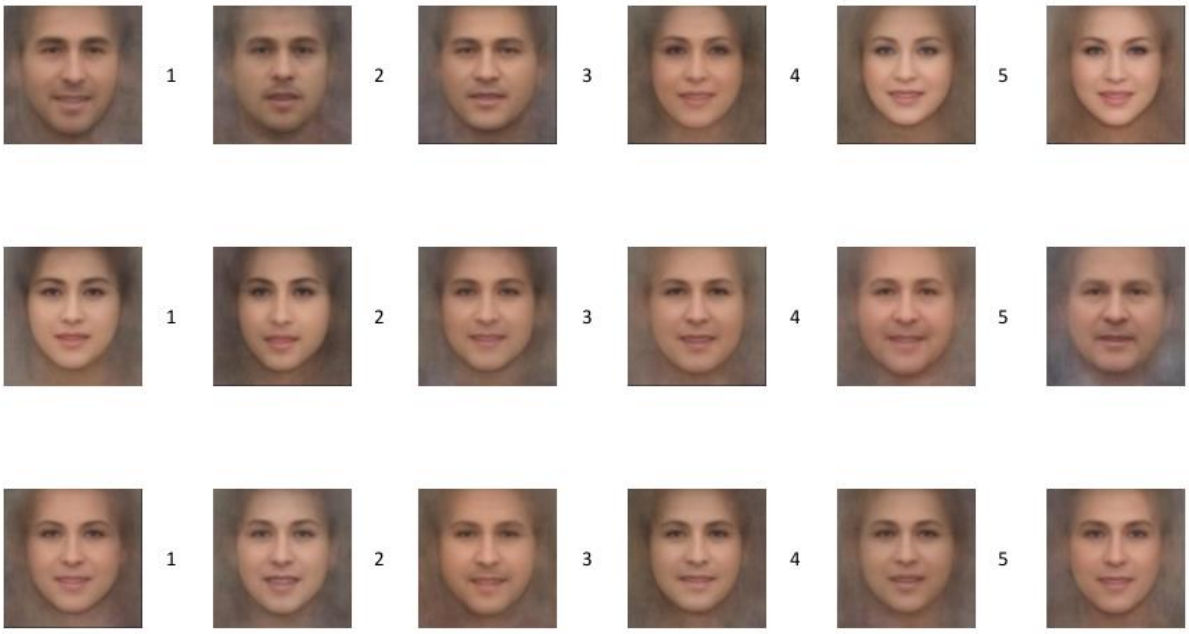

Figure 1. Sequences generated by sorting faces along the model's 49th attribute (Sequence A; top), the model's 45th attribute (Sequence B; middle), and shuffled randomly (Sequence Z; bottom).

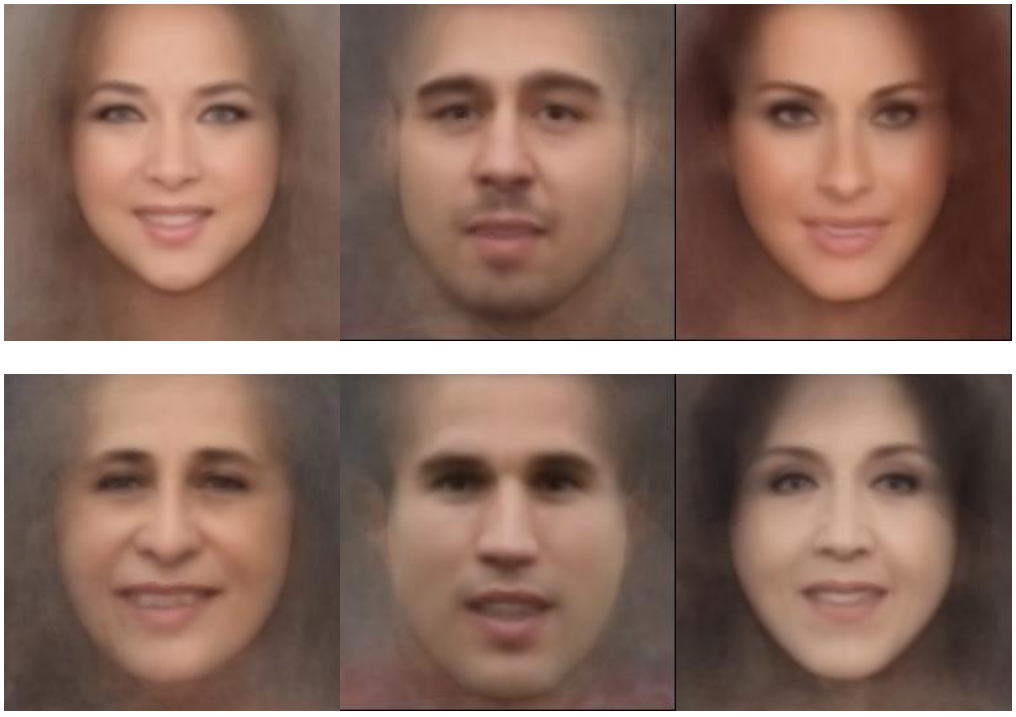

Figure 2. Unsorted faces generated by averaging 100 images of the same individual.

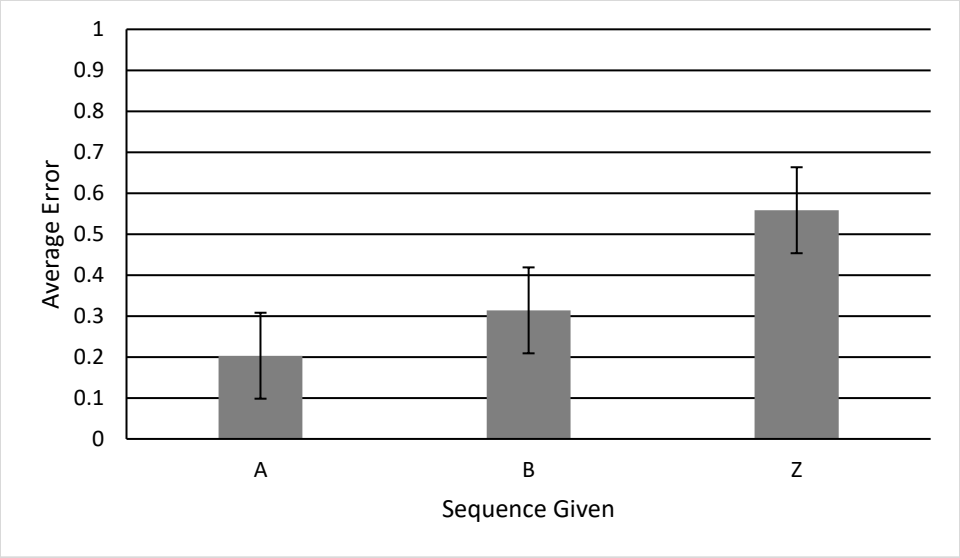

Figure 3. Participants who sorted faces into the sequences generated by the model (A and B) erred significantly less than participant's sorting faces into the sequence generated randomly (Z). Participants who sorted faces into the sequence sorted by the model's best predictor (A) erred significantly less than participants who sorted faces into the sequence sorted by one of the model's less effective predictors (B).