Similarity and set size impact the efficiency of hypothesis testing among faces

Andrew T. Hendrickson (a.hendrickson@tilburguniversity.edu)
Department of Cognitive Science & Artificial Intelligence
Tilburg University

Carolyn Semmler (carolyn.semmler@adelaide.edu.au)
School of Psychology
University of Adelaide

Abstract

Face processing is an important cognitive process that has been extensively studied in the domains of recognition memory and visual search. However, with the advent of biometric question-answering systems to assist experts searching for specific individuals, it is increasingly important to understand if the known biases and strengths of face processing extend to more deliberate natural language question formation and search. In this work we present a novel experimental task where people ask natural language questions to test hypotheses about faces in which the efficiency of the questions can be directly observed. The results indicate some aspects of existing paradigms do transfer to hypothesis testing, including less efficient questions when the number of items increases or as the similarity between items increases. However, unlike recognition memory specific demographic features (gender and ethnicity) do not seem to have a strong impact on efficiency.

Keywords: Identification; face processing; similarity; hypothesis testing; Visual Question Answering

Introduction

Faces have become a key biometric for determining identity across a wide variety of settings (Bone & Blackburn, 2003). However, a challenge in using the face for identity is the sheer volume of imagery available that could assist in identifying unknown individuals. With the advancement of biometric technology, systems have been developed that take natural language queries from a human operator and use these to narrow the list of possible candidates during search (Toor & Wechsler, 2018). These are often referred to as ‘soft biometrics,’ and include gender, race and age of the individual (Denman, Halstead, Fookes, & Sridharan, 2015).

Although these systems are still in development, their success depends upon supporting the full variety and scope of dimensions of identity that can be expressed by humans via natural language. There is very little known about how humans express the visual characteristics of other humans to test hypotheses about identity. Furthermore, there is little known about how efficiently people identify features while searching an array of faces. Instead, identity and face processing have been extensively studied using visual search and recognition tasks, while being largely ignored in the field of hypothesis testing. This work aims to evaluate the efficiency of hypothesis testing for facial identity across a range of factors known to impact face processing.

Face Processing in Search and Recognition

The majority of the evidence on how people process faces comes from search tasks. These tasks include search among faces stored in memory (e.g. recognition memory, Yin, 1969; Shapiro & Penrod, 1986) or within a visual array (Wolfe, 1994; Treisman & Gelade, 1980).

Despite the inherent differences between searching in memory and searching in space, the recognition memory and visual search literature show surprising convergence on the impact of similarity and specific features on face processing. Increasing the similarity between the item that is the target of the search and the non-target item decreases search efficiency for faces in recognition memory (Tanaka, Kantner, & Bartlett, 2012) and visual search (Kuehn & Jolicoeur, 1994). Furthermore, changes in similarity by increasing attention to unique features have been theorized produce the advantage people have for processing faces similar to their own. This has been demonstrated for more efficient processing of faces of the same race in visual search (Levin, 2000) and recognition memory (Valentine, 1991; Valentine & Endo, 1992) as well as the same gender in recognition studies (Herlitz & Lovén, 2013; De Frias, Nilsson, & Herlitz, 2006).

Hypothesis Testing

In contrast to the focus on processing images of faces in search and recognition, the literature on hypothesis testing has primarily focused on inferring properties of individuals based on written descriptions (Snyder & Swann, 1978; Snyder & Campbell, 1980). Perhaps unsurprisingly, the existing results are consistent with the large literature documenting the systematic, sub-optimal hypothesis testing due to confirmation bias (Wason, 1960; Nickerson, 1998). However, sub-optimal queries are not universally found; a strong match between the surface features of a task and the structure of mental hypotheses (Hendrickson, Navarro, & Perfors, 2016), a match between the distribution of mental and true hypotheses (Navarro & Perfors, 2011; Oaksford & Chater, 1994), or extensive background domain knowledge (Cosmides, 1989) can result in highly efficient queries in hypothesis testing.

Faces and hypothesis testing

In this paper we present empirical evidence suggesting that the efficiency of hypothesis testing to identify an individual person via natural language questions shows similar patterns to visual search and recognition memory in key areas, but disagreement in other respects. The agreement between tasks highlights that question-based hypothesis testing of faces is not a unique search task and that traditional effects in search, including increasing the number of items (Treisman & Gelade, 1980) and increasing item similarity

(Barras & Kerzel, 2017; Tanaka et al., 2012) decrease question efficiency. However, we do see evidence of a novel interaction effect on efficiency between the the number of faces and the similarity between faces. Despite the similarities between the question-based hypothesis testing and traditional search tasks, not all effects of face processing appear to transfer from the recognition memory and visual search. Specifically, we find no evidence of the own-race (Tanaka, Kiefer, & Bukach, 2004; Levin, 2000) and own-gender (Herlitz & Lovén, 2013) effects in hypothesis testing.

Pilot rating task

Prior to the hypothesis testing experiment, a pilot study was conducted to identify the features and properties of the face images (Hendrickson, Wang, & Atzmueller, 2018). Specifically, participants rated and evaluated the face images along a number of feature dimensions. Accurate evaluation of these features in a similar presentation context to the actual experiment was necessary to manipulating the similarity of the array as well as calculating own-race and own-gender effects.

Participants

Two hundred people were recruited from Amazon's Mechanical Turk and paid US$2 for approximately 12 minutes of work. Self-reported demographic information indicated the ages ranged from 18 to 75 years (mean: 34.6), 92.1% were from the United States, 40.8% were female, 58.4% were male, and <1% were neither male or female.

Stimuli

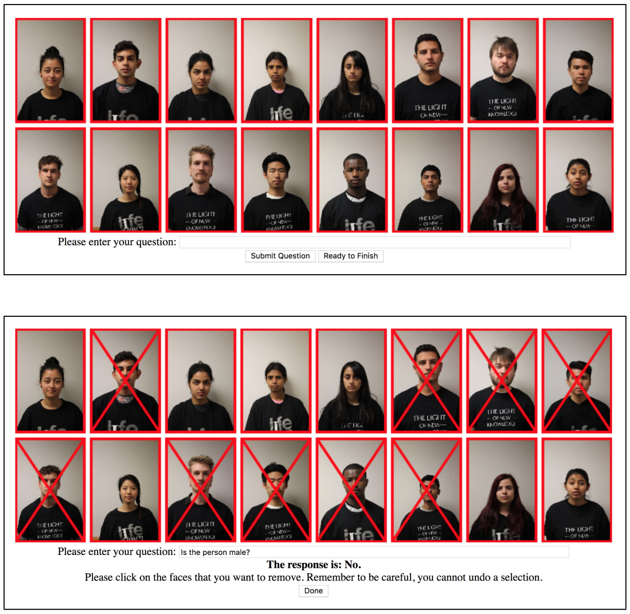
The stimuli consisted of 193 color images of the head and shoulders of people standing in front of an off-white background. These individuals were recruited from the University of Adelaide campus and compensated AUD$10 for their participation. Individuals were instructed to look directly into the camera lens and maintain a neutral facial expression. Multiple images were taken of each individual but only one image was used for each person.

Procedure

Participants were shown five randomly selected images in a random order. Participants were asked to estimate the age, hair color, eye color, race (selected from a list of African, Asian, Latino, or White), and gender (Male, Female, or Unsure) of each face. In addition, the participants were asked to make judgments about the person in the image, including rate their typicality and attractiveness on a five point scale, guess their occupation, and write a brief description of the person in the image. Each face image was evaluated by at least three participants. Judgments of gender and race for a face, the two properties manipulated in the subsequent experiment, were highly consistent across raters.

Experiment

Using the features of the face images extracted from the pilot study, the experiment evaluates the role of similarity between faces on the efficiency of a question-based search for a face. Participants were instructed to ask yes-or-no questions to eliminate candidate images, akin to the game Guess Who. Unlike the traditional two-person game, this task was not competitive and no true candidate face was selected. Instead, responses to the participants questions were randomly generated, resulting in participants producing a "successful" final guess regardless of which face they selected.



Please enter your question:

Please enter your question: Is the person male?

The response is: No.

Please click on the faces that you want to remove. Remember to be careful, you cannot undo a selection.

Figure 1: **Top panel: *Question Phase* of the experiment**. Participants are shown a set of faces and ask a yes/no question about the face they are trying to identify. **Bottom panel: *Elimination Phase* of the experiment**. Participants identify faces that are inconsistent with the response to their question and eliminate them by clicking on them.

Across multiple rounds, the similarity between the set of candidate images was manipulated by systematically selecting a set of possible images based on their rated gender and racial characteristics. The analysis of question efficiency focuses on a number of factors, including the number of available faces, the similarity between candidate faces, as well as the relationship between the demographic information of the participant and the candidate faces.

Participants

Twelve hundred people were recruited from Amazon's Mechanical Turk and paid US$3.75 for approximately 16 minutes of work. Complete data was collected from 1196 participants. Self-reported demographic information indicated the ages ranged from 18 to 71 years (mean: 34.5), 95.3% were from the United States, 46.2% were female, 53.4% were male, and <1% were neither male or female.

Method

Design Participants completed four rounds in which the similarity between the visible faces was manipulated between rounds by filtering the set of possible faces by gender and race. The High Similarity condition contained faces that all matched on gender and race. The set of 193 face images contained enough variation to create a group of Asian females (N = 25), Asian males (N = 16), White females (N = 60), and White males (N = 74). The Medium Similarity condition contained faces that either all matched in gender while race was evenly split between Asian and White faces, or matched in race while gender was evenly split among the faces. The Low Similarity condition contained faces that maximized racial and gender diversity (2 Asian female faces, 2 Asian male faces, 2 White female faces, 2 White male faces, and from non-White and non-Asian races: 4 female faces and 4 male faces). Each participant completed two High Similarity rounds, one Medium Similarity round, and one Low Similarity round in a random order. Except for preventing exact duplication of features in the High Similarity condition, the gender and race constraints for each round were randomly selected given the similarity condition and the set of faces to display was randomly selected given these constraints. The selected faces were displayed in a random order.

Procedure Each round consisted of alternating cycles of the *Question Phase* and the *Elimination Phase*. In each *Question Phase* participants asked a yes-or-no question about the face, and in the *Elimination Phase* participants eliminated faces that were not consistent with the response to their question from the *Question Phase*.

Question Phase. Participants were instructed to ask binary questions that could be answered with a yes or no response about the face they were searching for (top panel of Figure 1). Questions were required to contain only letters and numbers and contain at least four words (measured by a sequence of letters and numbers separated by a space) and at least 10 total characters. After a valid question was submitted, a random Yes or No response was displayed and participants transitioned to the *Elimination Phase*. Critically, though the response was randomly generated, at no time were participants explicitly told this was true. In fact, the instructions implied one face had been selected as the 'correct answer' face and the answers to the questions were about that face.

If only one face remained that had not been eliminated, participants could select the option to make a final guess and select the remaining face. In all rounds participants were given feedback that their selection was correct because there was no actual 'correct answer' face. Following this feedback, participants transitioned into another round with new faces or a debriefing page if all rounds were complete.

Elimination Phase. After asking a question and receiving a response, participants were instructed to click on faces that were not consistent with the response to their question and thus not the true identity face. A red X appeared over each face after it was clicked (see bottom of Figure 1) and clicking on an already eliminated face had no effect. Participants returned to the *Question Phase* when they indicated they were done eliminating candidate faces in the *Elimination Phase*.

Results

Defining question quality The quality of a specific question (q) was evaluated based on the expected reduction in the proportion of remaining faces after the *Elimination Phase* and not the actual reduction. This expected utility of a question, $Eu(q)$, was defined to be the sum across all possible answers (a), of the utility of that answer, $u(q,a)$, weighted by the probability of that answer occurring given the question, $p(a|q)$:

$$Eu(q) = u(q, `yes')p(`yes'|q) + u(q, `no')p(`no'|q) \tag{1}$$

This formulation results in a trade-off between utility and probability. For example, the utility of a question and answer pair ('Does the person have red hair?', 'yes') which eliminates 7 of 8 faces is $u(q, `yes') = 7/8 = 0.875$. However, from the perspective of participants who believe questions are answered about a specific true identity face and not randomly, this answer is unlikely given this question because the answer 'yes' will only be given for 1 out of 8 faces ($p('yes'|q) = 1/8 = 0.125$).[1] The reverse trade-off occurs for the 'no' response ($u(q, `no') = 1/8 = 0.125$ and $p(`no'|q) = 7/8 = 0.875$), resulting in a relatively low expected utility for this question ($Eu(q) = 0.875 * 0.125 + 0.125 * 0.875 = 0.22$). As in this example, all necessary values to compute expected utility can be directly observed in the data from each turn of the *Elimination Phase*.

The maximum expected utility will always occur for a question that produces answers that are equally likely, for example a different question q' that eliminates four faces with both a 'yes' or 'no' answer: $Eu(q') = (4/8)(4/8) + (4/8)(4/8) = 0.5$. Therefore, the expected utility of any question will be between 0 and 0.5. To facilitate comparisons across set sizes (which can produce rounding issues) and provide a more intuitive range of values (0 to 1), the expected efficiency of a question is defined to be the expected utility normalized by the maximum possible expected utility:

$$Ee(q) = Eu(q) / \max_{q'} Eu(q') \tag{2}$$

In practice the maximum possible expected utility is calculated based on an unknown optimal question which eliminates exactly half the faces with each response.

Available faces First, we evaluate the relationship between the number of available faces and question efficiency. A linear mixed-effects model with a random intercept for each participant shows a significant decrease in the expected efficiency of a question as the number of available faces increases

[1] This assumption underlies the computation of expected utility and expected efficiency for all analyses. However, it appears to be justified as no participants explicitly noted in open-ended comments that the responses were random. Furthermore, search behavior and expected question efficiency did not show major differences as a function of the number of games played, suggesting no qualitative shift in strategy based on learning about the question answering mechanism or the positive feedback about which face was the true identity.

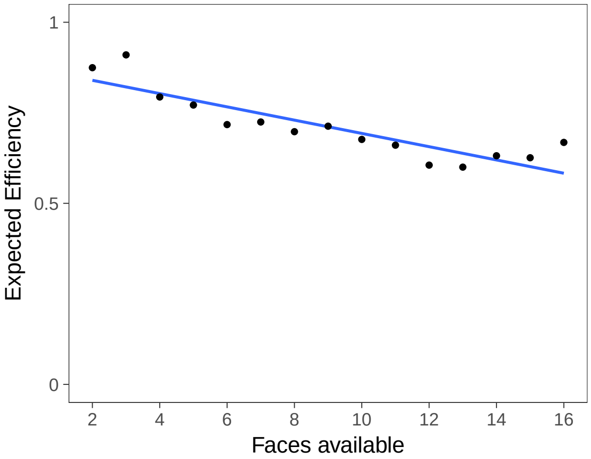


Figure 2: The expected efficiency of questions decreases as the number of available faces increases. The blue line indicates the best fitting linear model of the data and indicates a decrease in efficiency of 1.3% for each additional face. Points indicate the mean expected efficiency across all questions, aggregated across conditions and participants.

($b = -0.013, \chi(1) = 1137.6, p < 0.001$).[2] This suggests that questions asked with fewer available faces were more efficient and each additional face present when a question was asked decreased the expected efficiency of the question by approximately 1.3% (indicated by the blue line in Figure 2).

Face similarity Second, we consider the effect of face similarity on question efficiency. The similarity condition, coded as a three-level categorical variable,[3] affected efficiency when compared to a baseline model with only the random intercept for each participant ($\chi(2) = 174.85, p < 0.001$). Evaluation of the model parameters shows that the LOW SIMILARITY condition ($M = 0.78, SD = 0.31$) resulted in a reliably higher efficiency than the MEDIUM SIMILARITY condition ($M = 0.71, SD = 0.32, t(17654.7) = 2.00, p = 0.046$). Similarly, the HIGH SIMILARITY condition ($M = 0.71, SD = 0.32$) resulted in a reliably lower efficiency than the MEDIUM SIMILARITY condition ($t(17675.7) = 9.52, p < 0.001$).[4]

However, Figure 3 shows that the expected efficiency varies both as a function of the similarity between faces as well as the number of available faces. The negative relationship (seen in Figure 2) between the number of available faces and the efficiency of search is mirrored across all three similarity conditions, though the relationship between availability and efficiency appears to vary as a function of the similarity.

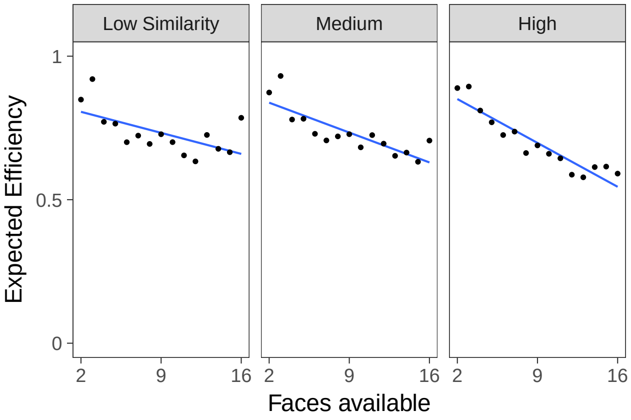


Figure 3: The expected efficiency of questions across the number of available faces and the similarity of the faces. Though overall expected efficiency was lowest in the high similarity condition, the results indicate an interaction between face similarity and the number of available faces. Specifically, the slope of the best fitting lines (shown in blue) for each similarity condition become progressively more negative as the faces become more similar. However, this is partially offset by an increase in the intercepts as similarity increases. Points indicate the mean expected efficiency across all questions.

The model with both similarity condition and availability provides a better account of efficiency than only similarity ($\chi(1) = 1097.80, p < 0.001$) or only availability ($\chi(2) = 134.99, p < 0.001$). However, including the interaction term between similarity and availability significantly improves the model ($\chi(2) = 344.09, p < 0.001$).[5]

Similarity Condition	Intercept	Slope
Low	0.81	-0.0026
Medium	0.86	-0.010
High	0.85	-0.020

Table 1: **Face similarity regression weights.** The beta values for each condition reflect the slope and intercept of expected efficiency as a function of the number of available faces within each similarity condition. The parameter estimates for the intercept and slope for both the LOW SIMILARITY and HIGH SIMILARITY conditions were reliably different than the values of the MEDIUM SIMILARITY condition (see text for details).

The parameter estimates (Table 1) show a trade-off between the intercept for each similarity condition and the slope across availability. The intercept for LOW SIMILARITY was reliably lower than for MEDIUM SIMILARITY ($t(17784.3) = 5.10, p < 0.001$) while the slope across availability for LOW SIMILARITY is reliably higher than the slope for MEDIUM SIMILARITY ($t(17652.5) = 9.95, p < 0.001$). Similarly, the intercept for HIGH SIMILARITY is reliably higher than

[2] Model: `efficiency ~ available + (1 | subject.ID)`). All linear mixed effects models were fit in R (3.3.1) using the lme4 package (1.1-15) and all comparisons of these models utilize the likelihood ratio test for nested model comparisons.

[3] Model: `efficiency ~ similarity + (1 | subject.ID)`

[4] All t-tests for the effect of individual parameters were done with the Satterthwaite correction to the degrees of freedom for fixed effects in linear mixed-effects models using the lmerTest package (2.0-36) in R (Luke, 2017).

[5] Full model: `efficiency ~ similarity * availability + (1 | subject.ID)`

MEDIUM SIMILARITY ($t(17782.6) = 5.1, p < 0.001$) while the slope across availability for HIGH SIMILARITY is reliably lower than the slope for MEDIUM SIMILARITY ($t(17660.4) = 7.11, p < 0.001$).

Demographic effects Third, we consider the impact of race on the expected efficiency of questions. Specifically, we restrict the conditions to those in which all faces were from the same race, either White or Asian, and to participants who self-reported as White (N = 870) or Asian (N = 218).

The race of the participant affected expected efficiency relative to a baseline model with a random intercept for each participant ($\chi(1) = 12.48, p < 0.001$), with less efficient questions asked by self-reported Asian participants ($M = 0.70, SD = 0.34$) than White participants ($M = 0.74, SD = 0.31$). The race of the faces did not affect efficiency ($\chi(1) = 0.13, p = 0.72$). Neither the model containing both race factors ($\chi(1) = 0.17, p = 0.68$) nor a model with those factors and an interaction term ($\chi(2) = 5.62, p = 0.060$) was preferred when compared to to a model with participant race as the sole predictor.[6]

Finally, we consider the impact of gender on the expected efficiency of questions. Specifically, we restrict the conditions to those in which all faces were of the same gender. The gender of the faces affected expected efficiency relative to a baseline model with a random intercept for each participant ($\chi(1) = 19.33, p < 0.001$), with less efficient questions for female faces ($M = 0.71, SD = 0.33$) than male faces ($M = 0.73, SD = 0.31$). However, the gender of the participant did not affect efficiency ($\chi(1) = 1.99, p = 0.16$). Neither the model containing both gender factors ($\chi(1) = 2.13, p = 0.14$) nor a model with those factors and an interaction term ($\chi(2) = 2.25, p = 0.32$) was preferred when compared to a model with the gender of faces as the sole predictor.[7]

Discussion

The current work focuses on the effect of similarity and set size on the efficiency of testing hypotheses about identity from faces. Most notably, the similarity between faces has a impact on the efficiency in hypothesis testing that is consistent to those seen in recognition memory and visual search. Increasing similarity has been shown to decrease recognition accuracy (Tanaka et al., 2012) and visual search speed (Barras & Kerzel, 2017), and here in the questions were increasingly less efficient in conditions with highly similar faces. This agreement may be the result of consistent use of discriminative features between the fast search tasks and the deliberative question asking task. Dissimilar faces result in many features that uniquely identify the target face, improving processing in the speeded search tasks, as well as features that can be used to construct more optimal questions.

The impact of the number of faces also seems to follow a similar pattern in hypothesis testing as in visual search and recognition memory, where increases in set size leads to decreases in accuracy. In the hypothesis testing task, increasing the number of available faces led to less efficient questions, which may be due to larger sets of images requiring more time to evaluate possible questions and participants settling on questions that are ‘good enough’ (Simon, 1956).

Despite the consistent effect of the number of faces, our results do show an unexpected interaction the number of faces and the similarity between faces: the decrease in efficiency due to set size *becomes larger* if faces are more similar. To our knowledge, this pattern has not been demonstrated in visual search or recognition memory of faces, though there is evidence visual search efficiency may vary based on the familiarity of faces (Tong & Nakayama, 1999). One possible explanation for this pattern is that the number of available faces is not explicitly manipulated in the task but the variance in this feature emerges through the sequence of participants eliminating some faces before asking further questions. The sequential nature of this process opens the possibility that the distribution of similarity or other features impacting efficiency might subtly shift throughout the sequence of questions. For example, if highly salient features are more often asked about in the earliest questions, this may change the distribution of faces in later questions to be more or less similar. Evaluating this hypothesis will require careful explicit manipulation of the number and similarity of available faces.

The own-race and own-gender effects seen in recognition memory and visual search were not found in the question-based hypothesis testing task. Question efficiency was not impacted by the agreement between the race of the participant and the race of the faces or the gender of the participant and the gender of the faces. The lack of evidence for either effect, despite their consistency in meta-analyses of visual search (Levin, 2000) and recognition (Herlitz & Lovén, 2013), suggest that the advantage for faces similar to the participant may not extend to more deliberate search tasks.

Conclusion Overall we see evidence that the efficiency of the questions people ask to search for an individual is more strongly influenced by the number of faces and the similarity of the faces than specific demographic characteristics of the people in the images, including gender and race. This only partially matches the patterns seen in recognition memory and visual search for faces and highlights the deliberate nature of the question asking task. These results suggest the most helpful support from biometric systems that assist human operators in processing identity may depend more on the number and similarity of the people displayed in the search task rather than the specific properties of the people (Heyer, MacLeod, Carter, Semmler, & Ma-Wyatt, 2017). The degree to which this pattern is unique for questions about processing images of faces or identity remains an open research question.

[6] Full model: `efficiency ~ subject.race * face.race + (1 | subject.ID)`. Including the number of available faces produced a similar result. Including any interactions between the number of available faces and race factors produced a model with singularities that precluded accurate model estimation.

[7] Full model: `efficiency ~ subject.gender * face.gender + (1 | subject.ID)` Including the number of available faces produced a similar result.

Acknowledgments

This research was supported by an University of Adelaide Interdisciplinary Research Grant to C. Semmler, A. Hendrickson, R Heyer, A. Dick and A. van den Hengel. Stimulus materials were collected with support from the Australian Research Council (DP16010148) to C. Semmler.

References

Barras, C., & Kerzel, D. (2017). Target-nontarget similarity decreases search efficiency and increases stimulus-driven control in visual search. *Attention, Perception, & Psychophysics*, *79*(7), 2037–2043.

Bone, J., & Blackburn, D. (2003). *Biometrics for narcoterrorist watch list applications* (Tech. Rep.). Technical report, Crane Division, Naval Surface Warfare Center and DoD .

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, *31*(3), 187–276.

De Frias, C. M., Nilsson, L.-G., & Herlitz, A. (2006). Sex differences in cognition are stable over a 10-year period in adulthood and old age. *Aging, Neuropsychology, and Cognition*, *13*(3-4), 574–587.

Denman, S., Halstead, M., Fookes, C. B., & Sridharan, S. (2015). Searching for people using semantic soft biometric descriptions. *Pattern Recognition Letters*, *68*(Part 2), 306–315. (Special Issue on "Soft Biometrics")

Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2016). Sensitivity to hypothesis size during information search. *Decision*, *3*(1), 62–80.

Hendrickson, A. T., Wang, J., & Atzmueller, M. (2018). Identifying exceptional descriptions of people using topic modeling and subgroup discovery. In *International symposium on methodologies for intelligent systems* (pp. 454–462).

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: a meta-analytic review. *Visual Cognition*, *21*(9-10), 1306–1336.

Heyer, R., MacLeod, V., Carter, L., Semmler, C., & Ma-Wyatt, A. (2017). Profiling the facial comparison practitioner in australia.

Kuehn, S. M., & Jolicoeur, P. (1994). Impact of quality of the image, orientation, and similarity of the stimuli on visual search for faces. *Perception*, *23*(1), 95–122.

Levin, D. T. (2000). Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*(4), 559–574hypo.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502.

Navarro, D. J., & Perfors, A. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*(2), 139–156.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129–138.

Snyder, M., & Campbell, B. (1980). Testing hypotheses about other people: The role of the hypothesis. *Personality and Social Psychology Bulletin*, *6*(3), 421–426.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202–1212.

Tanaka, J. W., Kantner, J., & Bartlett, M. S. (2012). How category structure influences the perception of object similarity: The atypicality bias. *Frontiers in Psychology*, *3*, 147.

Tanaka, J. W., Kiefer, M., & Bukach, C. M. (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. *Cognition*, *93*(1), B1–B9.

Tong, F., & Nakayama, K. (1999). Robust representations for faces: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(4), 1016.

Toor, A. S., & Wechsler, H. (2018). Biometrics and forensics integration using deep multi-modal semantic alignment and joint embedding. *Pattern Recognition Letters*, *113*, 29–37.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *43*(2), 161–204.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *The Quarterly Journal of Experimental Psychology*, *44*(4), 671–703.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145.