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

The ability to process faces is a critical cognitive process that supports interaction across many social and interpersonal domains. The factors that influence face processing, including the number of faces, the similarity between faces, and the similarity between the faces and the perceiver, have been extensively studied using paradigms where people search visually or in memory. What has received less attention are their influence on face processing in slower, deliberative search tasks such as hypothesis testing. Here we present a novel experimental task where people were asked to test hypotheses about faces in an iterated question asking task in which the efficiency of the test questions can be directly observed. The results indicate that, unlike visual search or recognition memory, increasing similarity between faces *improves* the efficiency of hypothesis testing and the relationship between the race and gender of the participant and the faces has no impact on performance.

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

Introduction

Imagine three every-day scenarios you might encounter in a crowded room: remembering if you had previously met someone you are being introduced to, searching for a specific friend, and asking questions of a friend on the phone about someone they want you to meet. Each of these scenarios involves processing information about a person's identity, yet each is likely to rely on different representational features or cognitive mechanisms. The first scenario involves searching in memory for individuals you remember meeting and comparing the current person to those memories, the second scenario involves visually searching for a particular known individual amongst a set of distractors, and the third scenario involves verbally communicating properties of an individual and asking questions about specific features to aide identification.

Identity and face processing has been extensively studied using visual search and recognition tasks that formalize the first two scenarios, yet have been largely ignored in the field of hypothesis testing, which comprises the third. 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 behavioral evidence on how people processes faces comes from search tasks. This search can either be amongst faces stored in memory using a recognition memory task (Yin, 1969; Shapiro & Penrod, 1986) or in the physical world using a visual search task (Wolfe, 1994; Treisman & Gelade, 1980). Despite the inherent differences between searching in memory and searching in physical space, these literatures show surprising convergence on the impact of particular factors 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, increased similarity has been theorized to be involved in the advantage people have for processing faces similar to their own by increasing attention to unique features. 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 intense study of visual face processing in search and recognition, the literature on hypothesis testing of identity has focused on written descriptions of people (Snyder & Swann, 1978; Snyder & Campbell, 1980). Physical and personality features do not seem to be unique in the domain of hypothesis testing and match the overall finding that people seem to show systematic, sub-optimal biases in the types of queries they make (Wason, 1960, 1968; Nickerson, 1998). However, these sub-optimal queries may be due to a mismatch between the actual distribution of hypotheses and the perceived distribution of hypotheses (Navarro & Perfors, 2011; Oaksford & Chater, 1994; Klayman, 1987). Furthermore, a better match between the surface features of a task and the true structure of the hypotheses (Hendrickson, Navarro, & Perfors, 2016) or background domain knowledge (Cosmides, 1989) can have a dramatic improvement on the efficiency of queries in hypothesis testing.

Faces and hypothesis testing

In this paper we present empirical evidence suggesting that hypothesis testing to identify an individual person via questions shows qualitatively different patterns than visual search or recognition memory in key areas but agreement in other respects. First, increasing the similarity of faces improves the efficiency of questions in hypothesis testing, despite harming visual search efficiency (Barras & Kerzel, 2017) and recog-

nition accuracy for faces (Tanaka et al., 2012). Second, although there is a large literature demonstrating own-race (Tanaka, Kiefer, & Bukach, 2004; Levin, 2000) and own-gender (Herlitz & Lovén, 2013) effects in visual search and recognition memory, we find no evidence of these effects in hypothesis testing. Despite these differences, hypothesis testing for faces does share the property with recognition memory that increasing the set size of possible faces decreases performance.

Pilot rating task

Prior to developing the hypothesis testing experiment, a pilot study was conducted to identify the features and properties of the face images. Identifying these features is necessary to construct manipulations of similarity as well as evaluate own-race and own-gender effects. In this pilot study a set of participants rated and evaluated the face images along a number of feature dimensions.

Participants

200 people were recruited from Amazon's Mechanical Turk and paid US$2 for approximately 12 minutes of work. 40.8% were female, 58.4% were male, and less than 1% reported 'neither.' Ages ranged from 18 to 75 years (mean: 34.6). 92.1% were from the United States and representation from every other countries was less than 1%.

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. They were also asked to rate their typicality and attractiveness of the person 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 raters and judgments of gender and race for a face, the two properties manipulated in the subsequent experiment, were highly consistent across raters.

Experiment

The experiment evaluates the role of face similarity in hypothesis testing by systematically selecting faces based on the rated gender and racial characteristics. The hypotheses in this case correspond to which of a set of possible faces is the target face and the task is to repeatedly reduce the set of possible faces by asking yes or no questions, akin to the game Guess Who. Unlike the traditional two-person game, this task was not competitive and no true target 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. By recording the set of available faces, the questions people ask, and the faces they eliminate as a result of the feedback, the efficiency of their questions can be evaluated. Specifically we focus our analysis on two aspects that may impact question efficiency: the number of available faces and similarity both between faces as well as the similarity between the participant and the faces with regard to gender and race.

Participants

1200 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. 46.2% were female, 53.4% were male, and less than 1% reported 'neither.' Ages ranged from 18 to 71 years (mean: 34.5). 95.3% were from the United States and representation from every other country was less than 1%.

Method

Figure 1: **Top panel: *Question Phase* of the experiment**. Participants are shown a set of faces and ask a yes/no question about the target 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.

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 and race was evenly split between Asian and White faces, or matched in race and 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 the *Question Phase* participants asked yes or no questions about the target 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 target face they were searching for (top panel of Figure 1):

> You will see a set of faces and you will need to determine which face is the "Correct" face [by asking yes-no questions] ... An example of a yes-no question might be: "Is the person a boy?" An unacceptable question is "What color eyes does the person have?" because it cannot be answered with a yes or no response. Please also refrain from asking questions about the clothes the person is wearing. You can ask as many yes-no questions as you like but you should try to use fewer questions.

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*. 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 instead of submitting a question. In all rounds participants were given feedback that their selection was correct because no true target face was actually selected. 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 target face.

> After each question, you will be told the answer and then given a chance to click on as many faces as you like to eliminate them from the screen. If you eliminate a face, it cannot be your final guess. When you are done eliminating faces and ready to ask another question, press the Done button.

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

In all analyses, we evaluated the expected efficiency of a question ($EE(q)$). Expected efficiency was calculated as the ratio of the expected utility of a question ($EU(q)$) relative to the maximum utility ($MU(N_a)$) given the number of available faces (N_a):

$$EE(q) = EU(q)/MU(N_a) \tag{1}$$

where utilities were positive and $MU \geq EU$, thus $0 \geq EE(q) \geq 1$. The expected utility of a question is calculated following the general formula for expected utility in which the probability of an outcome was multiplied by the utility of an outcome:

$$EU(q) = \sum_{o \in outcomes} U(o)P(o) \tag{2}$$

The utility of an outcome in this context was the ratio of the number of eliminated faces (N_e) to the number of available faces (N_a). The probability of getting the outcome that eliminates N_e/N_a faces was the proportion of faces that were not eliminated ($(N_a - N_e)/N_a$). The two outcomes are symmetrical, thus expected utility was defined as:

$$EU(q) = 2 * (N_e/N_a) * ((N_a - N_e)/N_a) \tag{3}$$

Utility was maximized when half of the available faces were eliminated:[1]

$$MU(N_a) = 2 * ((N_a/2)/N_a) * ((N_a - N_a/2)/N_a) \tag{4}$$

Available faces First, we evaluated the relationship between the number of available faces and question efficiency. A linear mixed-effects model with a random intercept for each participant (`efficiency ~ subject.race * face.race + (1 | subject.ID)`) found a significant decrease in the expected efficiency of a question as the number of available faces increased ($b = -0.013, \chi(1) = 1137.6, p <$

[1] If the number of available faces is odd, then Equation 4 requires that rounding of $N_a/2$ to be consistent throughout the calculation.

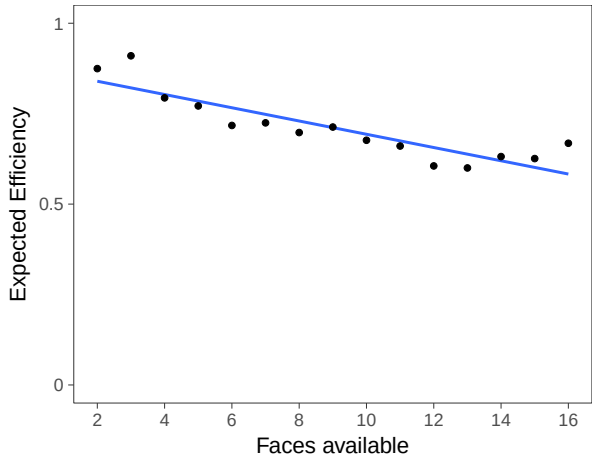

Figure 2: The expected efficiency of questions decreases as the number of available faces increases. The blue line indicates the best fitting linear model of the data and indicates a decrease in efficiency of 1.3% for each additional face. Points indicate the mean expected efficiency across all questions,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

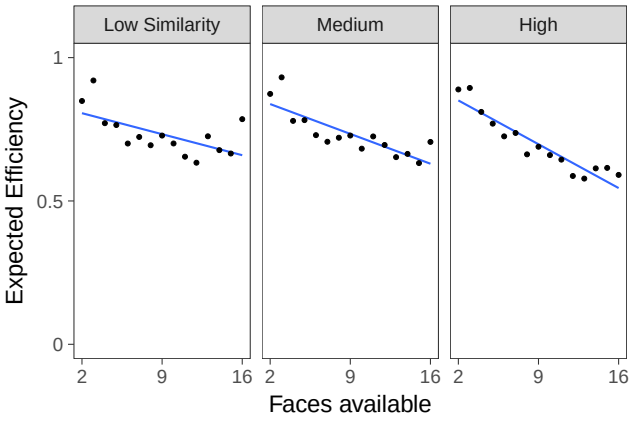

Figure 3: The expected efficiency of questions across the number of available faces and the similarity of the faces. Though overall expected efficiency was lowest in the high similarity condition, the results indicate an interaction between face similarity and the number of available faces. Specifically, 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.

0.001).[2] This suggests that questions asked with fewer available faces were expected to be more efficient and each additional available face decreased the expected efficiency of a question by 1.3% (Figure 2).

Face similarity Second, we considered the effect of face similarity on question efficiency. Similarity affected efficiency relative to an intercept-only baseline model ($\chi(2) = 174.85, p < 0.001$). 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 ($M = 0.71, SD = 0.32, t(17654.7) = 2.00, p = 0.046$),[3] and the HIGH SIMILARITY condition ($M = 0.71, SD = 0.32$) resulted in a reliably lower efficiency than the MEDIUM SIMILARITY condition ($t(17675.7) = 9.52, p < 0.001$).

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.

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 improved the model ($\chi(2) = 344.09, p < 0.001$). The parameter estimates (Table 1) showed a trade-off between the intercept for each similarity condition and the slope across availability. The intercept of the LOW SIMILARITY condition was reliably lower than the MEDIUM SIMILARITY condition ($t(17784.3) = 5.10, p < 0.001$) while the slope across availability of the LOW SIMILARITY condition is reliably higher than the slope of the MEDIUM SIMILARITY condition ($t(17652.5) = 9.95, p < 0.001$). Similarly, the intercept of the HIGH SIMILARITY condition is reliably higher than the MEDIUM SIMILARITY condition ($t(17782.6) = 5.1, p < 0.001$) while the slope across availability of the HIGH SIMILARITY condition is reliably lower than the slope of the MEDIUM SIMILARITY condition ($t(17660.4) = 7.11, p < 0.001$).

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).

Demographic effects Third, we considered the impact of race on the expected efficiency of questions. Specifically, we

[2] 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] 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).

restricted the conditions to those in which all faces were from the same race, either White or Asian, and to participants who self-reported their race as either White (N = 870) or Asian (N = 218).

The race of the participant affected expected efficiency relative to an intercept-only baseline model ($\chi(1) = 12.48, p < 0.001$) 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.[4]

Finally, we considered the impact of gender on the expected efficiency of questions. Specifically, we restricted the conditions to those in which all faces were of the same gender. The gender of the faces affected expected efficiency relative to a random-intercept baseline model ($\chi(1) = 19.33, p < 0.001$), 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.[5]

Discussion

The current work evaluates the effect of similarity and set size on the efficiency of testing hypotheses about faces. Specifically, we contrast these effects in hypothesis testing with previously established results in the domains of visual search and recognition memory. We find agreement across domains for some factors but differences for others.

Most notably, the similarity between faces has a qualitatively different impact on performance in hypothesis testing when compared to recognition memory. Increasing similarity has been shown to decrease recognition accuracy (Tanaka et al., 2012) and visual search speed (Barras & Kerzel, 2017), yet in the hypothesis testing task questions were increasingly more efficient as faces became more similar. This disagreement may be the result of completely different uses for discriminative features. Dissimilar faces result in many features that uniquely identify the target face, improving processing in the speeded search tasks, but selecting which of these many features is the most optimal feature to ask a question about may be difficult and result in people reverting to heuristic search and selecting sub-optimal questions (Oaksford & Chater, 1994).

The similarity between the faces and the participant, the own-race and own-gender effects, was also different in the 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 the consistency of these effects in meta-analyses of visual search (Levin, 2000) and recognition (Herlitz & Lovén, 2013), suggests that the perceptual advantage for faces similar to yours do not extend to more deliberate decision tasks.

The impact of set size, by contrast, did seem 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. Despite this agreement, the hypothesis testing task showed an unexpected interaction between set size 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 high similarity and high set size both contribute to increasing the difficulty of searching for most efficient question by increasing possible features and the combinations of faces across features exponentially.

The non-linear relationship between the number of available faces and the possible combinations of faces that would be eliminated by a feature highlights the systematic non-linear relationship between efficiency and set size in all panels of Figure 3. The data suggests a non-linear (perhaps quadratic) function might provide a better fit to describe the relationship between expected efficiency and number of available faces. This possibility will be evaluated in future work.

These results have several theoretical implications for our understanding of how identity is represented. Most theories of person identity represent an individual identity having an encoded set of features in face space (Lewis, 2004) that vary along dimensions. The intersection between visual search and language descriptors is relatively unexplored in face space models, yet seems a necessary area for understanding how people solve the complex task of identifying someone during a hypothesis testing task.

The results reported here also have implications for design of biometric systems. The development of Verbal Question Answering in Artificial Intelligence represent important developments in the way that humans and machines will interact. For example, face biometric systems are often used to search for persons of interest in large arrays returned by the system before experts are shown the best matches (Ross, Nan-

[4]The formula for the full linear mixed-effect model was: `efficiency ~ subject.race * face.race + (1 | subject.ID)`. When the number of available faces was included with no interaction effects the model comparison showed 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.

[5]The formula for the full model was: `efficiency ~ subject.gender * face.gender + (1 | subject.ID)` When the number of available faces was included in all models the comparison showed a similar result.

dakumar, & Jain, 2006). It remains unclear what information from an automated system people find useful when searching for the correct identity of a target in a large array or how people querying an automated system to get information they want could then use that information to narrow their search (Heyer, MacLeod, Carter, Semmler, & Ma-Wyatt, 2017), but this work may help contribute to understanding what features would be most useful for human-machine systems to suggest to improve hypothesis testing decision processes.

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.

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.

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

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.

Klayman, J. (1987). *An information theory analysis of the value of information in hypothesis testing* [Working paper no. 171]. University of Chicago, Graduate School of Business, Center for Decision Research.

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.

Lewis, M. (2004). Face-space-r: Towards a unified account of face recognition. *Visual Cognition*, *11*(1), 29–69.

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.

Ross, A. A., Nandakumar, K., & Jain, A. K. (2006). *Handbook of multibiometrics* (Vol. 6). Springer Science & Business Media.

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

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.

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.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

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.