Title
Ensemble Coding and Perceptions of Belonging

Author
Goodale, Brianna Mae

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Ensemble Coding and Perceptions of Belonging

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

By

Brianna Mae Goodale

2018

ABSTRACT OF THE DISSERTATION

Ensemble Coding and Perceptions of Belonging

by

Brianna Mae Goodale
Doctoral Candidate in Psychology
University of California, Los Angeles, 2018
Professor Kerri Johnson, Chair

Prior research has shown human observers extract perceptual summaries for sets of items after brief visual exposure, accurately judging the average size of geometric shapes (Ariely, 2001) or the walking direction of a crowd (Sweeny, Haroz, & Whitney, 2013). Beyond actuarial summaries, I hypothesized that observers extract social information about groups that may influence downstream judgments and behavior. In Study Set One, I showed that humans accurately perceive the sex ratio of a group after only 500 millisecond visual exposure. I then tested whether these percepts bias judgments about the group's social attitudes and affect the perceiver's sense of belonging. As the ratio of men to women increased, perceivers judged the group to harbor more sexist norms, and judgments of belonging changed concomitantly, albeit in opposite directions for men and women. Using reverse correlation, I also demonstrated that majority-male groups elicit a mental representation perceived as angrier and less approachable than the average majority-female group member.

In Study Set Two, I examined whether the ensemble coding process replicates with real-world groups. Using stimuli derived from panels presenting at the 2015 Society for Personality and Social Psychology Annual Meeting, I showed that perceivers rapidly extract actuarial and social summaries from 4-person groups. I next probed whether ensemble coding judgments extend to attributions about the group's broader impacts, finding that male-dominated panels were rated as less likely to mentor underrepresented minorities and disseminate their work publicly. Thus, observers judge a group's sex ratio from a mere glimpse, inferring from it social attitudes and interpersonal affordances.

In the final chapter, I described a novel open-source Python framework I created to test my hypotheses. Platform-independent, the Social Vision Toolkit facilitates experimental stimuli presentation and exists free of charge. Although researchers must write their own scripts, the Social Vision Toolkit includes several demo experiments and code annotation to aid new programmers. It can be downloaded from an open-access GitHub repository, allowing users to adapt the code and customize features to suit their needs. The Social Vision Toolkit enables greater methodological transparency and collaboration, as well as potentially broadens the scope of what is experimentally possible.

The dissertation of Brianna Mae Goodale is approved.

Peter John Lamberson

Margaret Joan Shih

Scott Pratt Johnson

Kerri Johnson, Committee Chair

University of California, Los Angeles

2018

Dedicated to my LP, my 33⅓ , Samuel Brotherton,

who never stopped believing in me.

Table of Contents

Acknowledgements

First and foremost, I could not have completed this academic feat without the support and patience of my mentor, Kerri. Thank you, Kerri, for taking a chance on me and for allowing me to rediscover the joys of research. You helped me find my voice in graduate school, indulging my interests in learning both R and Python. I will forever be grateful for the compassion, mentorship, and academic home you have given me.

I also wish to thank my parents, Dianne and William Goodale, Jr., as well as my sister, Cortney Proulx. They encouraged me to take a risk six years ago, moving across the country in pursuit of my career dreams. Only making slight fun of me, they have put up with my stats obsession and picked up the phone when I needed a dose of New England love. They help keep me grounded, while pushing me to reach for the stars.

Another compelling force driving me towards completing my doctoral education was my undergraduate mentor, Dr. J. Richard Hackman. Sharing a rural upbringing, he invested in my professional development and was the first person to encourage me to consider a career in academia. Thank you, Richard, for refusing to write that law school recommendation many years ago. I miss you immensely but try to honor your memory by working hard, reaching out to lift up younger students, and promising one day I'll learn how to fly fish.

Thank you as well to my graduate school support system, Christina Schonberg and Dr. Anthony Rodriguez. Beyond epic Rock Band marathons, you taught me the true value of friends and made me laugh during some of the hardest semesters. Christina, thank you for taking me for retail therapy (despite my reluctance) and for secretly enjoying reality TV (strictly as an important psychological case study, of course). Anthony, I am grateful for your willingness to share your stats notes, your advice on parenting, and time with your statistically significant other.

I am incredibly appreciative of the support from my labmates, Nicholas Alt and Jessica Shropshire. I have grown immeasurably as a researcher because of our conversations; you have pushed me to think critically and creatively, as well as tolerated buggy scripts and provided feedback on early versions of the Social Vision Toolkit. You have been patient as I learned a new literature and never hesitated to send me articles relevant to my interests. All in all, you have made me feel like I have a place in our lab and for that, I am incredibly grateful.

I would be remiss without thanking the National Science Foundation (NSF) and the UCLA Center for the Study of Women (CSW) for their financial support. Receiving the NSF's Graduate Research Fellowship enabled me to pursue my research goals exploring barriers facing women's entry into STEM. Similarly, because of the CSW's travel award, I could share my research at the 2015 Society for Personality and Social Psychology's Annual Meeting.

I also wish to acknowledge my co-authors, Nicholas Alt, Dr. David James Lick, and Dr. Kerri L. Johnson, for their contributions to Study Set One, Groups at a glance. A version of that chapter is currently under review following an invited revise & resubmit, at the Journal of Experimental Psychology: General. The study methodology and hypotheses were my ideas, with my co-authors providing feedback on the initial design, the pilot findings, and drafts of the manuscript. Dr. Lick helped with the reliability analyses for Study 1a, sharing his SAS script so I could run the subsequent analyses myself.

Finally, thank you to my life partner, Samuel Brotherton, for being my cheerleader, personal chef, and best friend. I could fill another 180 pages describing the ways in which this would not have been possible without you; however, I will spare you (and me) that embarrassment. Suffice to say, I love you and am looking forward to tackling the rest of our adventures together. Bring on the trad climbs, mountain biking, and parenting—we got this!

Brianna Mae Goodale
Curriculum Vitae
https:github.com/bgoodale

EDUCATION

M.A. (December 2013) *University of California, Los Angeles*, Los Angeles, CA
Major: Social Psychology
Minor: Quantitative Psychology

B.A. (June 2009) *Harvard University,* Cambridge, MA
Cum Laude, Major: Psychology

AWARDS AND HONORS

2015 Student Research Award, Honorable Mention,
Association for Psychological Science
2014 Graduate Student Poster Award, Runner-up,
Society for Personality and Social Psychology

GRANTS AND FELLOWSHIPS

2014 – 2017 National Science Foundation Graduate Student Fellow ($32,000/year)
2015 UCLA Center for the Study of Women Graduate Student Travel Grant ($300)
2014 Research Initiative for Diversity and Equity Research Grant ($13,800)
Co-Principal Investigator with Hua W. Ni
2013 Graduate Dean's Scholar Award ($14,500), UCLA
2012 Graduate Dean's Scholar Award ($8,500), UCLA
2012 Distinguished University Fellowship ($21,000), UCLA

PUBLICATIONS

Alt, N. P., **Goodale, B. M.,** Lick, D. J., & Johnson, K. In press. Threat in the company of men: Ensemble perception and threat evaluations of groups varying in sex ratio. *Social Psychological and Personality Science.* http:doi.org/10.1177/1948550617731498

Shilaih, M.,*** Goodale, B.M.***, Falco, L., Kübler, F., De Clerck, V., & Leeners, B. In press. Modern fertility awareness methods: Wrist wearables capture the changes of temperature associated with the menstrual cycle. *Bioscience Reports.* http://doi.org/10.1042/BSR20171279
*Note: * indicates co-first authors*

Goodale, B. M., Alt, N. P., Lick, D. J., & Johnson, K. (2018). Groups at a glance: Perceivers infer social belonging in a group based on perceptual summaries of sex ratio. Invited revise and resubmit. *Journal of Experimental Psychology: General.*

Schonberg, C. & **Goodale, B. M.** (2018). *Predicting preschool enrollment among Hispanic WIC participants in Los Angeles County.* Manuscript submitted for publication.

CONFERENCE SYMPOSIUM AND POSTER PRESENTATIONS

Goodale, B. M., & Johnson, K. (2017, January). Groups at a glance: How perception of sex ratio influences belonging. In Farley, S., & Hall, J. (Chairs), *Nonverbal Preconference.* Symposium conducted at the annual meeting of the Society for Personality and Social Psychology, San Antonio, TX.

Goodale, B. M., Ni, H. W., Huo, Y., & Johnson, K. (2017, January). *Socioeconomic status shows: How dynamic body motion increases stereotype-consistent endorsements of trait judgments.* Poster presented at the annual meeting of the Society for Personality and Social Psychology, San Antonio, TX.

Goodale, B. M. & Shih, M. (2016, January). *Strength in numbers: How group confrontation of sexism bolsters women's math performance and decreases stereotype threat.* Poster presented at the annual meeting of the Society for Personality and Social Psychology, San Diego, CA.

Goodale, B. M. & Shih, M. (2015, May). *Stifling Silence: How failure to confront increases stereotype threat among women in STEM.* Poster presented at the annual meeting of the Association for Psychological Science, New York, NY.

Goodale, B. M. & Shih, M. (2015, August). How conflicting or congruent stereotypes contribute to the "Asian Advantage" in STEM. In Gu, J. & Aquino, K. (Chairs), *Opening governance to Asian Americans.* Symposium conducted at the annual meeting of the Academy of Management, Vancouver, British Columbia, Canada.

TEACHING AND MENTORING

University of California, Los Angeles

Winter 2017 – *Teaching Fellow*, Social Networking, Student Evaluation: 6.90/9.00
Spring 2016 – *Teaching Associate*, Psychological Statistics, Student Evaluation: 7.25/9.00
Winter 2016 – *Teaching Associate*, Social Networking, Student Evaluation: 6.30/9.00
Spring 2014 – *Teaching Asst.*, Research Methods in Psychology, Student Evaluation: 8.44/9.00
Winter 2014 – *Teaching Asst.*, Human Motivation, Student Evaluation: 7.79/9.00
Fall 2013 – *Teaching Asst.*, Introduction to Social Psychology, Student Evaluation: 8.26/9.00

Harvard University

Spring 2010 – *Teaching Fellow*, Group Decision Making, Student Evaluation: 4.40/5.00
Fall 2010 – *Teaching Fellow*, Abnormal Psychology, Student Evaluation: 3.15/5.00
Fall 2009 & Fall 2010 – *Teaching Fellow*, The Social Psychology of Organizations, Student Evaluation: Fall 2009, 4.33/5.00; Fall 2010, 3.27/5.00

PROFESSIONAL AND DEPARTMENTAL SERVICE

2014 – 2017 Graduate Adviser, Undergraduate Research Journal of Psychology, UCLA
2014 – 2017 Graduate Studies Committee Representative, Psychology Graduate Students Association, UCLA
2015 Graduate Sponsor, Psychology Research Opportunity Programs (PROPS), UCLA
2013 – 2014 Reviewer, Psychological Science, APSSC RISE Award Competition
2013 Reviewer, Psychological Science, APSSC Student Research Award

Significance of Research

> [T]o appreciate an Impressionist painting one has to step back a few yards, and enjoy the miracle of seeing those puzzling patches suddenly fall into one place and come to life before our eyes.
>
> – Ernst Gombrich, *The Story of Art*

A 19th-century European art movement, Impressionism was demarcated by its distinct painting style; up close, the canvas appeared covered in tiny individual dots. As British art historian Gombrich (1950) notes, however, stepping back from the painting revealed a larger image. Whether intentional or not, this artistic style mimics a visual process now known and widely recognized as ensemble coding. An Impressionist admirer does not register each paint stroke in the bucolic picture before him; similarly, in ensemble coding, a perceiver may not register each individual object in the scene. Instead, the singular objects fade to form a broader impression transcending its parts.

Having had a variety of names across multiple literatures (Alvarez, 2011; Alvarez & Oliva, 2008), *ensemble coding* refers to the visual process of extracting and encoding summary statistics about a set of objects. It often occurs at the expense of local details about the constituent objects (Alvarez, 2011; Alvarez & Oliva, 2008). Prior research has demonstrated how humans can, upon short exposure to a set, give an accurate indication of mean descriptive details (e.g., average size; Chong & Treisman, 2003, 2005a, 2005b). Some evidence exists pointing to ensemble coding as higher-order processing, demonstrating perceivers can holistically encode a set of human faces as well as geometric shapes (Haberman & Whitney, 2007, 2010). To date, most empirical work has probed ensemble coding's boundaries and the descriptive mechanisms underlying it. We know ensemble coding occurs when observing a

group of people (de Fockert & Marchant, 2008; Haberman & Whitney, 2007, 2009a; Sweeny et al., 2013). However, we do not yet know how these instantaneous judgments affect our decisions to interact with or enter that group.

The goal of this dissertation was to probe the ways in which ensemble coding may inform perceivers' downstream social judgments. Through two series of studies and a methods piece establishing a novel stimuli presentation program, I aimed to understand how statistical summaries of sets of objects contribute to instantaneous feelings of fit and belonging. Past research in cognitive and vision science led me to predict that changes in the sex ratio of a group would be rapidly and accurately processed by perceivers. Additionally, I expected this perceptual change to result in differences in trait judgments about the group and its constituent members. This research has important behavioral implications for minority group members, including women in science, technology, engineering, and mathematics (STEM). It suggests behavioral intentions to engage with a stereotyped domain may arise earlier than previously believed. I sought to test whether social judgments and feelings of fit can be evoked from a half-second glimpse of a group.

History of Ensemble Coding

Early Research on Scene "Gist" Extraction

As early as the 1970s, vision scientists wondered whether scene perception occurred piece-wise, feature-by-feature, or in an overall Gestalt manner. Biederman, Glass, and Stacy (1973) showed 36 participants 112 black and white slides of everyday locations like streets, kitchens, and store counters. The researchers divided every picture into six equally-sized segments. Each participant saw half the pictures in the "coherent" condition and half the pictures in the "jumbled" condition. In the coherent condition, participants saw each picture as it had

been taken. In the jumbled condition, participants saw four of the six picture segments shuffled out of order. Biederman et al. used a two-alternative forced choice task (2AFC), requiring participants to indicate with a button press whether or not they had seen a given object in the prior scene. For a third of all trials, the target object was in the scene (e.g., a teacup on the kitchen counter). For another third of all trials, the target object was not in the scene but would not have been out of place there (the "possible-no" condition; e.g., a teacup as the target object but absent from the kitchen scene). In the final third of trials, perceivers rendered judgments about "impossible-no" targets: objects that would not have made rational sense in the scene (e.g., a car in the kitchen).

Although testing specific object recognition, Biederman et al. (1973) were interested in how scene generalization impacted speed of recognition. They found that participants rendered impossible-no judgments faster than possible-no judgments. Biederman et al. concluded perceivers may have a schema for what a scene should look like. The authors hypothesized that this script allows perceivers to identify misplaced objects more quickly. In later studies, Biederman, Rabinowitz, Glass, and Stacy (1974) demonstrated perceivers could abstract an overall impression of a scene and whether objects belonged there in as little as 100 milliseconds (ms). Although testing response speed and not a summary statistic about the scene per se, Biederman and colleagues provided early insight into the possibility of holistic processing.

The Global/Local Processing Framework

Turning towards the how behind holistic scene processing, Navon (1977) theorized that perceivers see the overall scene crudely at first before decomposing it into fine grained constituent pieces. He hypothesized the existence of a "global-to-local" visual system, which he tested across two studies. In Experiment 1, Navon manipulated whether dots of light comprised

letters, a rectangle, or a random pattern. The visual stimuli flashed onscreen simultaneously or shortly after the auditory stimuli. Perceivers had to indicate whether the visual stimuli matched the auditory stimuli they heard. In Experiment 2, participants again saw the letters "H" and "S" as well as a rectangle and indicated whether the stimuli matched the letter they had heard. Each letter or rectangle was made up of composite smaller letters. Navon found participants responded to the visual stimuli on a global level, matching the ensemble letters with their respective constant sounds. Most participants were surprised to later learn that smaller, local letters had been present within the bigger shapes; they could only recall seeing the global features. Although Navon framed his research in terms of Gestalt processing and overall perception, his work constitutes some of the earliest research to consider how sets of objects may be processed holistically with little regard for their constituent parts.

Other researchers (Alvarez & Oliva, 2008; Kimchi, 1992) have since disagreed with the global versus local hierarchal distinction made by Navon (1977). In a review paper drawing from studies across vision science, Kimchi (1992) argued that a global and local features framework erroneously assumes that the two features are independent levels of visual processing. Navon conceived global and local processes as operating in near-identical perceptual units, with global features perceived first. Kimchi, however, cites evidence from her earlier studies (Kimchi, 1983, 1988, 1990; Kimchi & Merhav, 1991; Kimchi & Palmer, 1982, 1985) refuting the perceptual similarity between the two features. Her research suggests that local features may have different perceptual units entirely from global processing.

Interdependence in the Holistic Processing Model

Kimchi (1983, 1988, 1990, 1992) preferred to conceptualize the relationship between perceiving sets and constituent parts in spatial terms. She posited that *wholistic processing* [sic]

differs from global processing in that it encompasses the interdependence between constituent parts. While the relationship in global/local processing is top-down and hierarchical, in Kimchi's framework the holistic extraction is neither inherently a priori nor does it "dominate" component feature processing. Kimchi notes the similarity between her conceptualization of holistic processing and *configural* or *emergent* processing (Garner, 1978; Pomerantz & Pristach, 1989; Rock, 1986; Treisman, 1986). Since the early 1980s, other researchers have drawn on this idea to examine comparable interdependent, spatial relationships between constituent and entire sets (e.g., *Spatial Envelope properties;* Oliva & Torralba, 2001). The holistic model of visual processing proposed by Kimchi (1992) pushed the field further towards ensemble coding, suggesting an interplay between local features and a set or scene of objects as a whole.

Statistical Processing of Object Sets

Alvarez and Oliva (2008) considered both the holistic and global/local features models. Ultimately, they favored referring to statistical summaries of sets as *ensemble visual features.* They argued that prior terms were used interchangeably with low spatial frequency; the term ensemble visual features, however, has no spatial-frequency band limit. Parallel terms have developed in other fields. Ariely (2001) researched perception of *sets* of similar items, drawing on statistics and descriptive summaries as inspiration for his hypotheses. He reasoned that the visual system can abstract a representation of the set's mean and distributions instead of summing constituent objects. Chong and Treisman (2003, 2005a, 2005b) similarly referred to summary statistics when testing whether perceivers can judge mean shape size across sets of circles. Alvarez and Oliva (2008), however, criticized the conceptualization of descriptive set processing for not accounting for the relationship between constituent objects. They argued sets may be seen as discrete objects, not inherently grouped or interdependent. Ensemble coding, in

its current incarnation, maintains the statistical theory underlying summary sets, but needs further research to adequately address Alvarez and Oliva's critique.

Overlapping Conceptual Research

In addition to converging research from statistical sets, ensemble coding draws on similar concepts as the property *numerosity* in vision science. Numerosity refers to "a nonverbal capacity for the apprehension of number, in humans" (p. 425, Burr & Ross, 2008). Burr and Ross (2008) were concerned with the number of constituent objects in a set. They reasoned that if numerosity was a primary visual property, it should show adaptation aftereffects. In line with their hypothesis, they found a bidirectional adaptation effect for numerosity; perceivers reported seeing fewer target objects in a set after a negative adaptation and more target objects after a positive adaptation. Burr and Ross concluded that like color or size, humans have a general sense of "numberness." Perceivers appear able to guess the approximate number of objects in a set, although environmental factors may influence their accuracy. In follow-up studies, Ross and Burr (2010) demonstrated numerosity's independence from density and texture. Although narrower in its defining features (e.g., average number of objects) than ensemble coding, research on numerosity demonstrates a similar higher-order process through which perceivers may quickly abstract summary statistics about a set of objects.

Defining Ensemble Coding

Despite differing names, early research across several domains shows converging evidence for a visual process evaluating a group of constituent items as a singularity. For the purposes of this dissertation, I define ensemble coding as a top-down process while acknowledging the object interdependency central to holistic processing. Furthermore, my definition primarily considers how perceivers extract descriptive summary statistics about a set

of targets; it does not, however, exclude the possibility that perceivers may rely on numerosity judgments to determine the frequency of targets in a set. Additionally, I have proposed that downstream judgments and behavioral consequences can occur instantaneously as a result of ensemble coding. In line with findings from evolutionary and cognitive psychology, ensemble coding may exist as an essential mechanism by which humans form primary judgments about an environment and the people in it.

Indirect Evidence of Ensemble Coding

Potential Biological and Evolutionary Advantages

Uses fewer biological resources. It is widely accepted that, given finite neural resources, the visual system has evolved to be more efficient (Attneave, 1954; Geisler, 2008). To compensate for biological limitations like a limited number of retinal neurons, human retinas and other early parts of our visual pathway have adapted to code lots of information quickly (Geisler, 2008). Furthermore, human visual processing cannot attend to more than a handful of objects at once (Luck & Vogel, 1997; Simons & Levin, 1998). By encoding images more sparsely through holistic processing, the body can conserve metabolic energy (Brady & Alvarez, 2011; Geisler, 2008; Olhausen & Field, 1997). Prior research suggests that the human visual system has evolved to reduce biological demands and support quicker appraisal of information. Thus, an ensemble coding conceptualization of set processing makes neurological and cognitive sense.

Provides a potential evolutionary advantage. In addition to biological compensations, a visual processing system that acts quickly and prioritizes summary statistics in decision making may have given our ancestors an evolutionary advantage. Knowing instantaneously whether a water source or group of people were safe to approach could have made the difference between life and death (Johnson, Iida, & Tassinary, 2012). Research has demonstrated how ensemble

coding leads to more precise judgments of information than evaluating items in a set individually. In descriptive judgment tasks, responses based on the mean of a set of objects proved more accurate (e.g., closer in size, number, or degree) than evaluations based on a single item, even when that item equaled the actual mean of the set (Alvarez, 2011; Haberman & Whitney, 2007; Leib et al., 2014). Ensemble coding also creates more efficient working memory, increasing processing speed and precision (Brady & Alvarez, 2011). Ensemble coding may have allowed our evolutionary ancestors to make more accurate snap judgments and react to a new visual scene as it unfolded.

Statistical advantage to ensemble coding. Averaging across constituent items in a set also increases judgment accuracy. From a statistical perspective, each judgment rendered by a perceiver will inherently have some amount of individual error (Alvarez, 2011; Sweeny et al., 2013). If an individual makes a judgment about one object's mean size, the judgment will consist of the object's true size plus or minus an amount of error. A given individual may both overestimate certain objects' size and underestimate other objects' size; this random bidirectional error will cancel out and leave the resulting mean measurement closer to the true mean of the set (Alvarez, 2011; Im & Halberda, 2013; Sweeny et al., 2013). As the number of objects in a set increases and more subsequent errors cancel out, perceivers' ensemble coding accuracy becomes more precise (Ariely, 2001; Leib et al., 2014; Sweeny et al., 2013; Sweeny & Whitney, 2014). Reproductive fitness depends on surviving long enough to pass on one's genes; it makes sense then that humans more quickly and more accurately able to determine a scene's safety would have had a greater chance of survival and thereby helped contribute to future generations' ability to ensemble code.

Ensemble coding could allow for faster in-group recognition. Another example of where ensemble coding may have biological advantages derives from face identification. Traditionally, researchers have deemed searching for specific faces in a crowd a slow and difficult task (e.g., Brown, Huey, & Findlay, 1997; Kuehn & Jolicoeur, 1994; Nothdurft, 1993). Memorizing all the faces in one's tribe or high school class, for example, may prove time consuming and require a lot of cognitive energy. Rather than rely on recall of specific faces, ensemble coding may have developed as a process through which perceivers could more easily recognize in-group members. Biederman et al. (1973) showed perceivers can scan a scene and quickly determine whether something does or does not belong. Extracting the mean face of a novel group, perceivers could compare it to the schematic gist of their in-group; if it matches, they may have deemed it safe to approach. In line with this rationale, Haberman and Whitney (2007) found perceivers recognize and rapidly discount a deviant face from the group mean. This process may have proved useful in enabling our ancestors to identify potentially hostile out-group members.

Models of Holistic Processing in Person Perception

Extracting summary information from a set of objects is not a novel concept in social cognition; research in person perception has consistently shown perceivers process faces holistically rather than by the sum of their parts (Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher, McDermott, & Chun, 1997; McKone, 2004; Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Holistic person perception, in both bodies and faces, seems analogous to ensemble coding: perceivers gather multiple cues simultaneously before representing the target face or body as a single unit.

Perceivers process human faces configurally. Tanaka and Farah (1993) defined holistic representation as the mean representation of an item set. They predicted that facial features like the eyes, nose, or mouth would be more readily recognized by perceivers in the context of the target's whole face than when shown individually. Participants saw six two-dimensional pictures of faces in succession and blocked by condition; the facial features were either in their respective anatomical positions or had the eyes, nose, and mouth scrambled out of order. During the test phase, participants had to recall via a 2AFC which facial feature or whole face came from a named target. Perceivers responded correctly most often when seeing the facial feature in the context of the target's whole, anatomically correct face; they were significantly worse at recalling name-face associations when the feature was presented devoid of a face. Based on their results, Tanaka and Farah argue that perceivers do not encode each facial element individually. Instead, the authors suggest perceivers abstract the face as a whole, sacrificing memorization of exact details in the process.

Farah, and colleagues (1998) provided further empirical support for holistic face processing. Across four studies, they tested how similar stimuli potentially interfere with facial encoding. Participants in each study saw a test face, followed by a partial or whole mask, and then a probe face. Perceivers had to indicate whether the probe face was the same or different as the test face. Farah et al. found that perceivers were more accurate at reporting similarities or differences in two faces when they saw only a partial face mask compared to a whole face mask. The whole mask appeared to interfere with encoding and recall of the prior test face. In contrast, partial facial features were seemingly encoded by a separate mechanism or in such a way that the gist of the first face remained intact.

Other researchers have since expanded on Farah et al.'s results; additional evidence now suggest that inverted or scrambled faces are processed piece-wise while upright, anatomically accurate faces are processed holistically (Maurer, Le Grand, & Mondloch, 2002; Moscovitch, Winocur, & Behrmann, 1997; Robbins & McKone, 2003; Tanaka & Sengco, 1997). These findings parallel early research on scene perception. Biederman et al. (1973), for example, showed that perceivers have a harder time recognizing the presence of a target object and take longer to summarize scrambled scenes than non-scrambled scenes. Perceivers seem to encode faces and scenes as informational sets, extracting a contextual "gist" about what should appear there. Set-based encoding may increase the difficulty of recognizing constituent features of the ensemble compared to when target stimuli are presented in an atypical or piece-wise fashion.

Extending holistic processing to ensemble coding. Given that our visual system has developed to encode faces holistically, it may be that it can encode *sets* of human faces holistically as well. While some researchers theoretically question this potential overlap (e.g., Haberman & Whitney, 2009a), further empirical work is needed to definitively establish or refute the relationship between individual and ensemble face processing. For now, it seems plausible that a process akin to configural or holistic facial processing could occur with groups of individuals.

Ensemble Coding with Sets of Objects

Numerosity and subitizing sets

At its most descriptive level, ensemble coding can enable perceivers to know the average number of items in a given set. As described previously, perceivers can accurately detect whether a set of geometric shapes has relatively more or fewer constituent objects than a previous set (Burr & Ross, 2008; Kaufman, Lord, Reese, & Volkmann, 1949; Mandler & Shebo, 1982; Ross

& Burr, 2010; Trick & Pylyshyn, 1993). Piazza, Mechelli, Butterworth, and Price (2002) demonstrated that, rather than raw counting, perceivers may rely on the separate process of subitizing. The authors aimed to test the oft-contested distinction between counting (defined as "the error prone and slow process of serially counting more than five objects," p. 435) and subitizing ("the ability to enumerate a small group of four or fewer objects fast and accurately," p. 435). Piazza et al. turned to neural imaging via PET scanners to provide insight into the relationship between these two processes. Overall, their results suggest a common neural network in the extrastriate middle occipital and intraparietal brain regions. However, participants showed greater activation in the occipitoparietal network when counting dots on the screen than when subitizing. Little regional activation when abstracting the numerosity of a small set is consistent with an evolutionary explanation for ensemble coding; decreased neural activity when gathering a perceptual "gist" mirrors the finding by Brady and Alvarez (2011) that ensemble coding conserves cognitive energy.

Moving beyond an average set count, researchers have begun to examine the constraints and boundary conditions of numerosity. Halberda, Sires, and Feigenson (2006) wanted to know how many sets perceivers can enumerate simultaneously. The researchers conducted experiments varying the number of colored dots (from one to six) in an onscreen display. In the before-probe condition, the experimenters revealed in the instructions that participants would need to report how many dots of each color they had seen. In the after-probe condition, the color of the dots and the desired task was not brought up until after the display ended. Regardless of probe condition, participants accurately reported the mean number of overall dots. Halberda et al. examined the error between probe-before and probe-after conditions to determine the maximum number of sets a perceiver could encode in parallel; even after controlling for dot area and circumference, they

found that most perceivers could generate the numerosity of up to three colors at once. Their research aligns with ensemble coding findings showing perceivers are just as accurate at assessing average facial emotion when an ensemble group contains four members as when it contains twelve members (Haberman & Whitney, 2007). Although more items in a set to encode results in the cancellation of more random measurement error, numerosity and ensemble coding may see only logarithmic returns; accuracy in summary statistics appears to plateau after a certain number of ensemble group members has been achieved.

Perceivers accurately recall average object size

Researchers have shown perceivers can accurately extract the average object size for a set of geometric shapes. Hypothesizing perceivers can abstract mean object size from a set of circles without knowing a priori details about the individual set members, Ariely (2001) had two participants complete three experiments. Each experiment required participants attend to a set of 4, 8, 12 or 16 similarly colored circles for 500ms. After the stimuli presentation, participants then completed a yes/no task to indicate whether the target circle shown was in the prior set (Experiment 1) or a 2AFC task to indicate which of the two target circles was in the prior set (Experiment 2). Ariely found that participants gave more "yes" responses when the target circle was within the range of circles shown in the stimuli set. Participants could not discriminate better than chance, however, which of two target circles were in the prior set. Ariely's findings suggest perceivers encode summary statistics about an overall set rather than cataloguing characteristics about individual targets. In the third experiment, Ariely had participants indicate whether the target circle was larger or smaller than the mean of the prior set's component circles. Participants accurately indicated the relative size of the target circle compared to the prior stimuli set, with precision increasing as set size increased.

Ensemble coding of average object size appears impervious to numerous task fluctuations; it cannot be easily explained away by attentional changes or demand characteristics. Drawing methodological inspiration from Ariely (2001), Chong and Treisman (2003) had participants complete a 2AFC task. Perceivers saw a divided screen on which twenty-four circles appeared, with twelve on each side of the screen. In some trials, the separate sets were presented simultaneously while in other trials, they were presented sequentially. Participants had to indicate which side of the screen had the set with the larger mean circle or which side of the screen had a larger circle presented independently after the sets. Even after decreasing target exposure to as little as 50ms or waiting up to 2000ms before allowing participants to respond, Chong and Treisman found participants could still detect and accurately discriminate between the mean object size for two separate sets. Chong and Treisman (2005a) sought to test the borders of holistic processing in follow-up studies. They similarly found no difference in perceiver accuracy across the component number, density, color scheme, or presence of a distractor set in group-based ensemble coding. Regardless of sequential or simultaneous presentation of constituent targets, perceivers can accurately estimate the mean size of objects in an ensemble (Albrecht & Scholl, 2010; Chong & Treisman, 2005a; Joo, Shin, Chong, & Blake, 2009).

Ensemble objects need not remain statically still for perceivers to make accurate assessments about their size. Albrecht and Scholl (2010) conducted several studies to determine whether perceivers can encode dynamic, moving targets. In Experiment 1, they showed 10 perceivers a white disc which expanded and contracted randomly; on half of the 200 trials, it moved through 9 anchor point placements while in the other half of trials, its placement remained central on the screen. The stimulus disc was presented for 250ms, prior to a 1ms grey

back mask followed by the test disc. Participants had to increase or decrease the test disc's size to align with the prior trial's average stimulus disc size. Consistent with prior work, perceivers were accurate in their representation of disc size; whether the target moved as it changed size or remained stationary did not moderate this effect. Albrecht and Scholl also varied whether the disc expanded or contracted over time as well as the rate at which it changed shape through each anchor point. Again, changes in the disc's expansion and movement speed did not alter participants' ability to accurately report the average disc size. Their findings suggest participants encode summary statistics continuously, rather than anchoring average judgments on a few select time points.

Perceivers Accurately Encode Average Objects' Orientation within the Set

Whether observing moving or stationary targets, the visual system can nevertheless encode the average position of set members within the ensemble. Parkes, Lund, Angelucci, Solomon, and Morgan (2001) showed participants a series of Gabor patches[1] arranged in a 12-item circle. Depending on the trial, the researchers presented participants with 1, 2, 3, 4, 7, or 9 Gabor patches tilted either clockwise or counterclockwise from the horizon; the remaining spots in the circle were filled with distractor Gabor patches. In a 2AFC, participants indicated whether the whole ensemble tilted clockwise or counterclockwise. Regardless of the number of distractor items included, Parkes et al. found participants pooled their judgments to account for the orientation variance across the set. Furthermore, perceivers could not recall the exact position of the tilted Gabors. Consistent with ensemble coding theory, participants generated accurate, fast summary statistics about the set of items but lost specifics about each Gabor in the process.

[1] Gabor patches are low-level stimuli, designed to specifically target V1 receptor cells (Haberman et al., 2015). They appear as a set of five striped black and white parallel lines, seemingly out of focus and fuzzy around the edges.

Alvarez and Oliva (2008) similarly explored how perceivers process and encode the positions of target images within an ensemble. They had 24 participants (8 per experiment) watch a series of target and distractor circles move around on a screen. Experiment instructions identified the target circles before each trial, attempting to direct perceivers' visual attention. Ostensibly, each experiment required participants to note how many times a subset of the circles crossed two red lines. In reality, however, Alvarez and Oliva were interested in perceivers' ability to judge details about the sets of objects. They also asked participants to indicate via mouse click the location of the target or distractor set's centroid, which they defined as "the center of mass of a collection of objects" (p. 392). Across all experiments, Alvarez and Oliva found perceivers identified a set's centroid location better than chance; this effect held regardless of whether participants made judgments about the target or distractor circles. In contrast, participants made fewer location errors when having to identify a single target circle compared to a distractor circle. The authors concluded that perceivers, although unable to give precise local details, nevertheless maintain rough summary statistics about objects and sets outside their primary attentional focus.

Ensemble coding appears to occur across multiple perceptual domains. It enables perceivers to accurately extract average size, speed, position, and orientation from a set of inanimate objects. While early theorists argued summary statistics could be derived from sampling one or two items in a set (e.g., Myczek & Simons, 2008), consistent convergent research suggests ensemble coding occurs as a higher-order visual process. Missing from this review thus far, however, is a consideration of how ensemble coding interacts with sets of people; how, if at all, does group-based visual processing to influence our social attitudes and affordances?

Ensemble Coding with Human Sets

Perceivers Encode Sets of Individuals by Average Facial Features

Transitioning from ensemble coding research with sets of objects to groups of people introduces more potential variance. Alvarez and Oliva (2008), for example, could control extraneous variables by limiting the size or direction of circles. In contrast, researchers examining human groups cannot viably recruit constituent set members with exactly the same nose or eye shapes. Each constituent target varies naturally along multiple facial dimensions. Thus, the first question considered by ensemble group researchers was how perceivers make sense of constituent group members varying in facial features: can they extract a mean facial representation in a similar fashion to how perceivers extract mean object size from an ensemble?

To answer this question, de Fockert and Wolfenstein (2009) recruited 18 participants who saw 60 sets of 4 faces presented simultaneously onscreen for 2000ms. Participants then saw a morphed face of the four individual faces shown, a morphed face of four individuals not shown, a member of the set shown morphed with itself, or another face not previously shown morphed with itself. Participants were asked to indicate whether the face had appeared in the set. De Fockert and Wolfenstein found that participants were significantly more likely to say that the morphed image of the four-person set was present than when an actual member of the set was presented; they conclude that perceivers automatically and quickly encode a group of faces using summary statistics, consistent with ensemble coding research on object sets.

Expanding on prior work, Kramer, Ritchie, and Burton (2015) questioned how perceivers encode ensemble images of the *same* target (rather than different targets). They collected seven target images of 10 American and 10 Australian celebrities from a Google search. Four randomly selected pictures of the same celebrity comprised a given ensemble group. During each trial,

participants saw the ensemble group briefly onscreen before indicating via a 2AFC whether the test stimulus had been present. The test stimulus was always one of the following: a single target image from the ensemble; a single target image not in the ensemble; the average morphed face from the ensemble; or an average of three target images not shown. Experiment 1 presented the four target images for each celebrity simultaneously while Experiment 2 presented them sequentially. In both experiments, the authors found participants selected "present" significantly more often for the matching ensemble average than the non-matching average of the three remainder images. Their results suggest ensemble coding operates with a level of specificity; perceivers formed a statistical summary of each celebrity based on the specific trial images they saw. Ensemble coding appears to generate a mental representation of the average of the targets within a given set, rather than create a broader conception of the person overall.

Perceivers Accurately Summarize Mean Ensemble Group Emotion

Beyond average descriptive characteristics, Haberman and Whitney (2007) wondered whether perceivers could extract relevant social cues from an ensemble as well. They recruited three participants to complete a series of experiments designed to test limitations of group-based facial processing. Across studies, Haberman and Whitney showed participants four faces morphed from happy to sad or male to female. Each ensemble set appeared onscreen for 2000ms. In the first experiment, participants then indicated whether the test face was happier or sadder than the preceding set of ensemble faces. Haberman and Whitney found perceivers formed accurate, better than chance mental representations of the average emotion in each set. In Experiment 2, Haberman and Whitney tested whether attention to individual constituent group members drove this effect. Participants again completed a 2AFC task, choosing the test face they believed came from the prior set. Participants did not perform better than chance, however, and

were unable to determine which of the two faces they had seen previously. Increasing set size to 16 members in Experiment 3, Haberman and Whitney replicated their earlier findings. Perceivers could tell the group's relative emotionality but remained unable to indicate the emotion of a single target individual.

Building on their prior research, Haberman and Whitney (2009b) designed six experiments to test whether ensemble coding of human sets requires high-level visual processing. In Experiment 1, they demonstrated that observers could accurately identify the mean emotion of a given set better than chance; varying set size from 4 to 16 people did not affect participants' ability to detect the average emotion. Similarly, in Experiment 2, the authors showed that observer accuracy in mean emotion did not change as a result of length of exposure to the set. Even at 50ms, most observers were still more accurate than chance at identifying the mean emotion of 4- or 16-person sets. Haberman and Whitney next refuted alternative hypotheses by probing whether observers encoded set members individually. They found that participants were similar or worse than chance at encoding an individual target's location within the stimulus array. As the number of individuals in a set increased, participants became less accurate at recalling where in the set a specific face had been presented or even if it had appeared at all. Lastly, Haberman and Whitney showed that observers have a harder time processing the mean emotion of a set when the individual faces are scrambled than when they appear unaltered. Their results suggest that ensemble coding relies on configural facial processing. Haberman & Whitney argued that ensemble coding, seemingly impervious to increases in set size, must therefore act via higher-order processing.

Perceivers Accurately Summarize Ensemble Facial and Body Directional Cues

Average group emotion is not the only visual cue perceivers can extract using ensemble coding. Sweeney and Whitney (2014) contended that humans may show sensitivity to mean eye gaze; they reasoned that social interactions and joint attention requires the ability to perceive where others are looking. Sweeney and Whitney hypothesized that perceivers rely on ensemble coding to determine the average gaze direction of a crowd. Eight participants saw 1, 2, 3, or 4 computer-generated faces staring into space. There were 4 possible eye gaze directions and 4 possible head rotations, which combined to form 16 different stimuli. Each trial randomly selected four of the 16 images and then, depending on condition, showed participants all or only some of the faces in the set. The authors found that as the subset increased in size, the average gaze direction more closely approximated the ensemble's actual mean gaze (determined from the full set of four targets). Sweeney and Whitney reported a less strong effect when the stimuli faces were inverted, providing convergent evidence that group facial processing occurs holistically. In Experiment 2, Sweeney and Whitney replicated their findings from Experiment 1 and demonstrated that ensemble coding occurs in as short a duration as 200ms. The authors noted that their findings may have important downstream behavioral implications; however, they did not test these directly.

In addition to eye gaze, perceivers may also deduce other summary statistics about an ensemble group's intention or direction of movement. Attempting to measure perceiver attention to a single target flanked by distractors, Thornton and Vuong (2004) inadvertently showed the supremacy of ensemble coding. They asked participants across three experiments to indicate whether a central target was walking leftwards or rightwards. Eleven dots placed at approximate joint locations comprised the point-light drawings of each target's body. Depending on trial condition, perceivers saw the target figure walking alone onscreen, flanked by figures walking in

the same direction, or flanked by figures walking in the opposite direction. Participants were significantly slower at indicating target walking direction when it appeared flanked by other figures than when it appeared alone. Thornton and Vuong consequently concluded that task-irrelevant dynamic figures nonetheless interfere with perceptions of a single target. They further cautioned that flanking figures may influence perceivers' behavior and judgments. Their work reinforces the idea that perceivers extract summary statistics from a group of people and lose or are slower at retrieving individual-level information.

Focusing on perception of overall crowd movement rather than the direction of a single target, Sweeney, Haroz, and Whitney (2013) also probed how perceivers estimate a crowd's average walking direction. In Experiment 1a and 1b, Sweeney et al. found that response variability decreased significantly as the number of ensemble group members increased. If perceivers were using only a single target to determine crowd direction, the authors reasoned the variability in predictions should have remain constant across crowd size. Separate blocks asking perceivers to judge crowd direction from scrambled and inverted point-light displays added further evidence that ensemble coding occurs independently of lower-level processes. In Experiment 2, Sweeney et. al manipulated the variability in individual target's walking orientation; nevertheless, perceivers still saw the crowd as moving in a homogenous direction. Participants' response error actually decreased when the 12 targets moved in highly variable directions than when the 12 targets all moved in the same, actual mean direction. Sweeney et al. improved upon point-light display methodology employed by Thorton and Vuong (2004) to directly test whether perceivers can ensemble code targets' average movement.

Ensemble Coding in Human Sets Occurs Across Time

Yamanashi Leib and colleagues (2014) considered ensemble coding's temporal and spatial limitations; they questioned whether accuracy in encoding an average face could occur when targets were presented sequentially and from different angles. Noting that prior research had only tested two-dimensional face perception, the authors manipulated whether target faces were looking straight ahead, leftwards, rightwards, or in side profile. Participants saw between 4 and 18 faces per trial; each face appeared alone on screen for 50ms. They took as much time as necessary to rotate a front-looking face to match the "average" facial orientation of the prior images. Across studies, Yamanashi Leib et al. found a significant main effect of set size; as the number of targets in a set increased, perceivers were significantly more accurate and precise in identifying the "average" facial orientation. Further tests revealed perceivers were significantly more accurate at judging the "average" face than identifying a single face. Yamanashi Leib and colleagues attributed this increase in accuracy to the cancellation of noise from cross-set averaging. Their relatively short stimuli exposure and variance in facial position for individuals within a set led the authors to classify ensemble coding as a higher-order visual process. Presenting targets individually and from different angles does not appear to impede perceivers' ability to extract summary statistics about a group of people.

The Effect of Individual-level Target Attributes on Ensemble Coding

Wondering about subitizing within ensemble coding, Nagy, Zimmer, Greenle, and Kovács (2012) examined how perceivers respond to subgroup differences within a set. In their first four studies, Nagy et al. showed participants 8-person ensembles varying in sex ratio. In Experiment 1, the authors manipulated whether participants had to categorize a specific target face (focal condition) or the mean of all the faces in the set (global condition) as male or female. Participants performed at chance in the focal condition, not differing in their classification of the

specified face based on the ensemble sex ratio. In the global condition, however, the proportion of times a participant responded "male" changed as a function of the ensemble's sex ratio. A follow-up study revealed that perceivers respond to the statistical gender averages of composite faces rather than the raw number of men or women in the ensemble. Nagy et al. also found evidence for ensemble coding's vulnerability to group-wide after-effects. In Experiment 3, after viewing a male or female-dominated ensemble, perceivers saw a target face as less or more masculine respectively. Experiment 4 illustrated how exaggerating the polarity of the group's gender composition (e.g., to be 50% male and 50% female) alleviated this effect. The authors' final fMRI study provided initial evidence that the bilateral fusiform face area may underlie ensemble-wide facial adaptations. This area is also predominant in the processing of single target faces and bolsters the argument that ensemble coding operates on the same mechanism as face-based processing. Among the first to consider individual-level characteristics of target group members, Nagy et al.'s research offers convergent neurological and perceptual evidence for ensemble coding.

Individual-level Perceiver Attributes Affect Ensemble Coding

Social cognitive scientists widely recognize that individuals' own biases and stereotypes can affect their perceptions of target stimuli (e.g., Fiske, Cuddy, Glick, & Xu, 2002; Johnson, Freeman, & Pauker, 2011; Kunda & Thagard, 1996). Ensemble coding research has only begun considering how perceiver attributes impact statistical summaries of a group. Thornton, Srismith, Oxner, and Hayward (2014) attempted to translate other-race effects often present in individual perception to ensemble coding. Thornton et al. showed 40 participants a 16-person ensemble whose array order changed repeatedly during its 4000ms presentation. Using a quasi-experimental design to compare White and Asian participants, the researchers varied the ratio of

Asian to White target faces in the ensemble group. After each trial, participants had to indicate via a 2AFC whether there were more Asian or more White targets in the group. Thornton et al. found Asian perceivers significantly underestimated the proportion of Asian targets in the group compared to White perceivers. Perceivers' own race appears to shape ensemble coding accuracy, with more research necessary to determine the cause of this discrepancy.

Bai, Leib, Puri, Whitney and Peng (2015) found a similar bias in summary statistic estimation based on perceivers' gender. They wondered whether gender differences inherent in individual perception would extend to group-level processing. The methodology used by Bai et al. closely mirrored prior ensemble coding studies. Participants saw counterbalanced blocks of White faces varying in number, gender, and orientation. Their findings replicated prior work by Haberman and Whitney (2009b); participants were more accurate at identifying the mean face from a set when the target faces appeared upright than when they appeared inverted. Across their four studies, Bai et al. (2015) also found a robust effect of perceiver sex. Female perceivers made fewer mistakes than male perceivers overall. A step towards better understanding perceiver-level variables, research by Bai et al. calls attention to the need to measure and report gender differences in ensemble coding.

Proposing an Integrated Theoretical Model

To date, prior research on ensemble coding has not been gathered into a cohesive, single theory and empirically validated. Phillips, Weisbuch, and Ambady (2014) offered a conceptual model for how ensemble coding could operate. They couched their three-part model in literature from organizational behavior, vision science, and social psychology. Their first step, Selection, posits that the perceiver to automatically identifies a group of targets. In the second step of their model, the authors noted that Extraction occurs. They drew on many of the ensemble coding

studies from vision science to explain how summary statistics result in spontaneous extractions of the group's central tendency and dispersion. Perceivers make judgments about the group based on these summary statistics in the final step of their model, Application. Phillips et al. did not claim that the last stage of their model happens instantaneously. Instead, they suggested it interacts with prior semantic knowledge to influence and weight perceivers' judgments. The authors concluded that their model serves an important role in explaining how organizational culture, norms, leadership, and team task assignment/performance may emerge. Despite not testing their theory empirically, Phillips et al. (2014) raised a call for future research to consider how summary statistics contribute to downstream behavioral consequences.

Adapting Prior Methodology

I have integrated methodology from ensemble coding studies conducted in psychophysics and vision science with norms from the field of social psychology to meet the final objective of my dissertation. Vision science research deviates from social psychology methodology in several key ways. First, most experiments rely on relatively small sample sizes (e.g., Chong & Treisman, 2003; de Fockert & Wolfenstein, 2009; Kramer et al., 2015; Navon, 1977; Sweeny & Whitney, 2014; Williams & Sekuler, 1984) with many hundreds or thousands of trials per participant (e.g., Thornton et al., 2014). Over the last few years, however, social psychology has grown increasingly concerned with low sample size. Editorials in leading journals have called for larger sample sizes, citing small experiments as more prone to Type I errors and less likely to be replicated (Button et al., 2013; Munafò et al., 2017). This difference in field norms may arise in part from psychophysics primary reliance on within-subject repetition to drive power. In contrast, many social psychology studies have mixed or between-subject designs. I have

attempted to find a solution catering to both fields by recruiting more participants than power analyses would suggest necessary while decreasing the average number of trials per study.

In addition to small sample sizes, many psychophysics experiments do not recruit naive or hypothesis-blind participants. Williams and Sekuler (1984), for example, recruited 4 participants for their study, one of whom was an author on the paper. Knowing the hypotheses of a study can lead to unintentional response bias (Pfungst, 1911); current best practices in social psychology suggest recruiting naïve participants when possible and keeping the experimenters themselves blind to hypotheses. For similar reasons, researchers should also avoid using the same participants across studies. To maximize generalizability and increase confidence in the validity of my findings, participants in each study presented in this dissertation were ineligible to participate in future studies.

Finally, psychophysics researchers test their hypotheses by using analyses that capitalize on distributional variance (e.g., repeated-measures analysis of variance, see de Fockert & Wolfenstein, 2009; Kramer et al., 2015; or, Thornton et al., 2014). This statistical analysis may be problematic for two reasons. First, analyses of variance (ANOVAs) cannot account for nested data structures; they look for differences in *mean* values across conditions (Judd, Westfall, & Kenny, 2012). Often, ensemble coding experiments have more than two trials or time points per condition. To conduct an ANOVA thus requires experimenters average across all stimuli or across all participants (Judd et al., 2012); unlike mulit-level or mixed modeling, ANOVAs cannot account for random, systematic variance in both perceivers and stimuli.

Additionally, the statistics behind ANOVAs assumes independence of errors. Because the same person sees different trials across time, their responses are not truly independent; they will be correlated and share some variance brought to each observation by the perceiver herself (J.

Krull, personal communication, January 6, 2016). Multi-level modeling allows researchers to partition out this shared error variance. Failure to do so may unintentionally inflate the seeming effect of condition; using a repeated-measures ANOVA approach inflates the Type I error rate and may lead researchers to erroneously conclude a random artifact constitutes a significant effect (Judd et al., 2012). Implementing multi-level modeling does involve a methodological tradeoff, however. Mixed models handle missing data better than ANOVAs (Judd et al., 2012). However, the need to estimate more parameters in a mixed model decreases statistical power (Westfall, Kenny, & Judd, 2014) and thus presents another reason to recruit larger samples than typically used in psychophysics. In my dissertation studies, I have sought to bring statistical and methodological norms from social psychology to bear on concepts previously studied in the domains of vision science and psychophysics.

Research Overview

I have developed a novel open-source Python framework and conducted two sets of studies to examine how the ensemble coding of a group's sex ratio influences perceivers' downstream judgments and behavioral intentions. The first set of studies sought to replicate prior research showing perceivers can accurately report descriptive summary statistics about a group of people. It then built on this work by testing whether perceivers can extract subjective summary statistics like feelings of fit and belonging from a mere glimpse of the group. Study Set Two probed the external validity of the proposed ensemble coding process by having perceivers view and respond to real-world groups. I also examined how a group's sex ratio contributes to its perceived intellectual merit and ability to create broader impacts. Lastly, the methods paper describes the novel methodology I used to conduct my dissertation studies. It establishes the Social Vision Toolkit as a free resource implementable by other social scientists and walks the

reader through how to code their own experiment script. Together, the two study sets and the methods paper aimed to forward the investigation of ensemble coding, applying social psychological methods to better understand how actuarial summaries of groups may drive human behavior.

Study Set One: Groups at a glance

Groups at a glance: Perceivers infer social belonging in a group based on perceptual summaries of sex ratio

Brianna M. Goodale, Nicholas P. Alt, David J. Lick, and Kerri L. Johnson

Abstract

Human observers extract perceptual summaries for sets of items after brief visual exposure, accurately judging the average size of geometric shapes (Ariely, 2001), walking direction of a crowd (Sweeny, Haroz, & Whitney, 2013), and the eye gaze of groups of faces (Sweeny & Whitney, 2014). In addition to such actuarial summaries, we hypothesize that observers also extract social information about groups that may influence downstream judgments and behavior. In four studies, we first show that humans quickly and accurately perceive the sex ratio of a group after only 500 ms of visual exposure. We then test whether these percepts bias judgments about the group's social attitudes and affect the perceiver's sense of belonging. As the ratio of men to women increased, both male and female perceivers judged the group to harbor more sexist norms, and judgments of belonging changed concomitantly, albeit in opposite directions for men and women. Thus, observers judge a group's sex ratio from a mere glimpse and use it to infer social attitudes and interpersonal affordances. We discuss the implication of these findings for a heretofore overlooked hurtle facing women in male-dominated fields (e.g., science, technology, engineering, or mathematics; STEM): how the ratio of men to women provides an early visible cue that signals an individual's potential fit.

Keywords: ensemble coding, sex ratio, belonging, group norms, social vision

Groups at a Glance: Perceivers Infer Social Belonging in a Group based on Perceptual Summaries of Sex Ratio

> I think the worst time was when I was all alone, after Sandra left. The public perception saw eight-men and then there was this little woman, hardly to be seen. But now, because I'm so senior, I sit towards the middle. I have Justice Kagan on my left, Justice Sotomayor on my right… So the public will see that women are all over the bench. They are very much part of the colloquy.
>
> —United States Supreme Court Justice Ruth Bader Ginsburg, *2015*

Reflecting on her appointment to the United States Supreme Court in an interview at Georgetown University, Justice Ginsburg discussed the balance of men and women on the bench. At the time of her appointment, Justice Ginsburg was the second woman ever confirmed to the Supreme Court and has, at various points, served as the sole female voice. With the presence of women on the Supreme Court growing to three under President Obama's administration, Justice Ginsburg's comments gestured toward a profound perspective underlying the pursuit of equality: that the ratio of men to women on the Supreme Court was not only salient to the public, but that it also carried important social implications for society's perceived fit of women as arbiters of law.

Justice Ginsburg's reflections may characterize percepts that extend well beyond the judiciary. In a similar way, the ratio of men to women has yet to reach parity in other traditionally masculine fields. In science, technology, engineering and mathematics (STEM), for example, women comprise only 24% of the American workforce (Beede et al., 2011), a statistic that indicates that observers of and participants in those fields are likely to encounter 3 times as many men as women. Here we test how one's first glimpse of a group's sex ratio shapes

perceivers' mental representations of a group and their feelings of fit therein. We bring to bear existing research and methods from the vison and cognitive sciences to probe how the spontaneous perception of a group's sex ratio impacts perceivers' judgments of the group's social attitudes and interpersonal affordances (i.e., judgments about whether a group offers the individual opportunities for belonging). We contend that these group percepts and the inferences that they evoke have broad implications for individuals in a range of fields, including the courtroom, and extending to tech start-ups and beyond.

The notion that merely perceiving a group might influence judgments of social attitudes and interpersonal affordances first requires that observers can (and do) achieve some degree of accuracy from minimal visual information, more generally. Indeed, people readily and rapidly form impressions about others, often achieving a remarkable level of accuracy (Hugenberg & Wilson, 2013; Macrae & Quadflieg, 2010). From merely glimpsing a face, people can judge a person's demographic identities (sex, age, race) and social traits (trustworthiness, dominance). These early percepts compel perceivers to approach or avoid others based upon appearance alone (Johnson, Iida, & Tassinary, 2012). Thus, the causes and consequences of social perception involving isolated faces are well understood. In daily life, however, people are often encountered in groups. Here we test both the accuracy and interpersonal implications of visually perceiving a group of unknown others. As has been observed in research probing the perception of *individuals*, we hypothesized that perceivers deduce similar characteristics of a group of people. Specifically, we predicted (and found) that observers' actuarial summaries about a group occur quickly and accurately, and that these summaries have important implications for subsequent social inferences involving a group's social attitudes and the interpersonal affordances the group provides.

Interpersonal Affordances: The Centrality of Belonging

Feelings of belonging are fundamental to the human experience (Baumeister & Leary, 1995). Indeed, belonging is so important that individuals will engage in goal-directed activity to satisfy the need to belong, and if their need goes unfulfilled, people actively seek out new groups or different relationships (Baumeister & Leary, 1995). Furthermore, persistent lack of belonging portends negative mental and physical health outcomes. For instance, lack of belonging fosters feelings of loneliness, anxiety, and depression (Baumeister & Tice, 1990; Conte, Weiner, & Plutchik, 1982; Hagerty & Williams, 1999; Lofland, 1982), and social isolation has been linked to increased cortisol activity (Kiecolt-Glaser et al., 1984), higher blood pressure (Cacioppo et al., 2002; Cacioppo & Hawkley, 2003; Hawkley & Cacioppo, 2003; Uchino, Cacioppo, & Kiecolt-Glaser, 1996), poorer immune functioning and healing ability (Cacioppo & Hawkley, 2003; Kiecolt-Glaser et al., 1984, 1987; Kiecolt-Glaser, Glaser, Cacioppo, & Malarkey, 1998), greater occurrence of psychosomatic health problems (DeLongis, Folkman, & Lazarus, 1988) and decreased sleep efficacy (Cacioppo et al., 2002; Cacioppo & Hawkley, 2003). Efforts to enhance social belonging appear to be effective, showing enhanced self-reported health and less frequent healthcare visits over a three year period (Walton & Cohen, 2011). Belonging interventions also enhance motivation (Walton, Cohen, Cwir, & Spencer, 2012) and academic achievement (Cohen, Garcia, Apfel, & Master, 2006; Shnabel, Purdie-Vaughns, Cook, Garcia, & Cohen, 2013; Walton & Cohen, 2011). Thus, felt belonging confers numerous emotional and physical benefits.

The determinants of an individual's feelings of belonging or fit with a group are multifaceted. In some instances, the group's own self-characterization provides potent outward facing signals. For instance, an explicit statement of a company's philosophy in a job-fair

brochure formed the basis of African American job applicants' trust in the company (Purdie-Vaughns, Steele, Davies, Ditlmann, & Crosby, 2008). More subtle cues to exclusion yield similar effects. In one study, observers responded to job postings and mock interviews that used either gender-exclusive (e.g., referring exclusively to "he" or "him") or gender-inclusive (e.g., "he or she" or "his or hers") language (Stout & Dasgupta, 2011). Among women, gender-exclusive language decreased felt belonging and motivation to pursue the position, but increased anticipated workplace sexism. The opposite was true among men; gender-exclusive language enhanced their motivation to pursue the position. Thus, both explicit and subtle language signaled distinct workplace environments to potential applicants, thereby affecting their likelihood to engage with the group. Moreover, once immersed within a group, individuals tend to distance themselves from traits that appear to be incongruent with the group's identity (Pronin, Steele, & Ross, 2004). Specifically, Pronin et al. (2004) found that women who enrolled in a high percentage of male-dominated courses (e.g., mathematics) eschewed feminine characteristics (e.g., wearing make-up), in an effort to enhance their fit within their social environment.

While most research to date has tested perceivers' response to social cues about the group, some evidence indicates that visual cues can change individuals' identification with a group or within a certain context. Dubbed "ambient belonging," this sense of social fit arises based solely on the presence of inanimate objects placed within the environment. In a computer classroom, simply replacing Star Trek posters with nature posters boosted women's interest in pursuing computer science as a major (Cheryan, Plaut, Davies, & Steele, 2009). Similarly, the presence of more minority individuals in a job pamphlet led African American participants to report feeling more organizational trust and to anticipate being more comfortable in that work

space (Purdie-Vaughns et al., 2008). Thus, in addition to more subtle linguistic indicators, visible cues can also serve to either enhance or curtail feelings of belonging.

Collectively, therefore, feelings of belonging show widespread importance, and they are informed by multiple factors ranging from subtle linguistic cues to incidental visual cues. For an observer attempting to discern whether a group affords the possibility of belonging, the earliest available information might be obtained visually from merely a glimpse, a possibility to which we now turn.

The Social Vision of Groups

Observers' visual perceptions of social spaces and the groups of people that inhabit them are likely to inform a range of subsequent judgments and behaviors. While some evidence indicates that perceivers attend to visual cues when making judgments about a social space, the process by which perceivers aggregate across groups of objects and/or people to draw inferences remains poorly understood. Speaking to this point, (Phillips et al., 2014) called for more empirical research into *people* perception to augment the sizable literature in *person* perception, arguing that people perception occurs much more frequently and has important implications for organizational behavior that rely on individuals' assessments of a social space (e.g., organization acculturation and team leadership). The current research is among the first to empirically test these processes, by examining how *visual* percepts (cf. verbal descriptions; Hamilton et al., 2015) about a group inform downstream social judgments of the group's likely social attitudes and interpersonal affordances.

Visually extracting the "gist". Early evidence from vision science provides support for our hypothesis that inferences about groups may have roots in objects processed as a collective set. In one study designed to probe the mechanisms of scene perception, participants briefly

viewed pictures of everyday locations (e.g., streets, kitchens, and store counters) and indicated whether or not they had seen a given object (Biederman, Glass, & Stacy, 1973). Participants' judgments were more efficient when an object was logically improbable for a given scene (e.g., a car in the kitchen) than when it was commonplace (e.g., a teacup in the kitchen). The researchers concluded that perceivers may rely on a schema or script for a scene's appearance, allowing them to quickly rule out objects that do not appear to belong there. Subsequent research demonstrated that this process could occur in as little as 100 ms (Biederman, Rabinowitz, Glass, & Stacy, 1974) and that specific details about an object's shape or identity are not necessary for accurate scene categorization (Oliva & Schyns, 2000; Oliva & Torralba, 2001; Torralba, 2003; Torralba & Oliva, 2003). Furthermore, perceivers are more accurate at summarizing spatial details when extracting the overall scene gist than when recalling the locale of individual objects within it (Alvarez & Oliva, 2008, 2010). Thus, prior research implies that perceivers can rapidly extract the gist of a scene, making holistic judgments that incorporate prior knowledge quickly and efficiently at some cost to local detail (for a more thorough review of prior research on scene gist extraction, see Oliva & Torralba, 2007).

Whether and how perceivers similarly extract the gist of a group of objects/people has become a focus of more recent research. Given the vast amount of visual information available to perceivers on a moment to moment bases, perceptual mechanisms evolved to ease the processing burden (Attneave, 1954; Barlow, 1961; Geisler, 2008). In one such manifestation, the visual system tends to aggregate when it encounters group of similar objects to yield a perceptual summary, or *ensemble representation* (Alvarez, 2011). This process, *ensemble coding,* facilitates quick and accurate summary percepts for object groups. This process appears to be extensive, informing a range of aggregate judgments for a group of objects, including their size (Ariely,

2001; Chong & Treisman, 2003, 2005a), the ratio of specific shapes (Burr & Ross, 2008; Ross & Burr, 2010), and their physical orientation (Parkes et al., 2001). Importantly, similar ensemble representations occur for social percepts. From brief exposures, perceivers accurately summarize a group's average facial emotion (Haberman & Whitney, 2007, 2009a, 2009b), eye gaze direction (Sweeny & Whitney, 2014), gender (Haberman & Whitney, 2007), race (Thornton et al., 2014), and walking direction (Sweeny, Haroz, & Whitney, 2013).

Ensemble coding appears to guide visual representations for both descriptive features of inanimate objects and overarching group characteristics. Still, the interpersonal consequences of these descriptive summaries remain untested. In the same way that ensemble coding eases one's perceptual burden by producing actuarial summaries (Olshausen, 2003), it might also ease one's social burden by providing a heuristic for estimating a group's social affordances. In particular, we hypothesized that ensemble coding may provide a means of judging whether a group affords personal belonging—a fundamental social need (Baumeister & Leary, 1995).

Thus, research has extended the concept of ensemble coding from summaries of simple visual objects to summaries of complex person characteristics, although the interpersonal consequences of these summaries remain unexplored. Theoretically, perceivers might perceptually group targets to increase the efficiency of decision-making (Olshausen, 2003). Summary representations of a group may therefore guide broader social judgments about a group in a heuristic fashion, enabling perceivers to estimate the behaviors, attitudes, and beliefs of individual group members given little more than the composition of the group itself. We hypothesized that this heuristic inference process may be especially important in the domain of personal belonging. Deciding in advance whether or not one belongs in a group presents a difficult challenge. Ensemble perception may provide an efficient solution to the challenge:

Upon seeing a group, perceivers may extract a summary representation of the "average" group member and use that representation to guide assumptions about the group overall.

The Present Studies

We tested these possibilities by integrating established methods from social, cognitive, and vision sciences. In Studies 1a and 1b, we first examined whether groups differing in the ratio of men to women were accurately perceived and then tested whether they elicited different perceptions of belonging from men and women. We predicted that perceivers would: a) accurately perceive the group's sex ratio following brief presentation; b) draw inferences about a groups attitudes based on sex ratio (i.e., the endorsement of sexist norms); and, c) utilize this information to gauge their own feelings of belonging within the group. Study 2 replicated the observed effects using a design that affords inferences about the temporal process. Study 3 used a different methodology to extend our observations by testing perceivers' mental representations of the average group member as a function of a group's sex ratio.

Study 1a

Methods

Design. Study 1a aimed to test how observers extract actuarial summary statistics and draw social inferences from brief exposure to social ensembles. A key goal in our approach was to test how the ratio of men to women in an ensemble affected judgments of social attitudes (i.e., harboring sexist attitudes) and interpersonal affordances (i.e., a sense of personal belonging). Study 1a employed a 2 (Perceiver Sex: male v. female) x 5 (Actual Sex Ratio: 0/8 men, 2/8 men, 4/8 men, 6/8 men, 8/8 men) mixed-model, quasi-experimental design in which each participant provided multiple judgments for each stimulus. As such, data were nested within stimuli and

participant. Along with all subsequent studies described here, Study 1a received research ethics committee approval prior to the collection of data.

Participants. Ninety-one U.S.-based Internet users (57% women) completed an online study via Amazon's Mechanical Turk (mTurk). Sample size estimates and power analyses for nested data require knowing numerous parameters including, but not limited to: the number of level-one groups (e.g., how many trials seen per participant); the size of the effect; each random effect's variance; covariance estimates for random effects; regression coefficients; and the design effect (Aguinis, Gottfredson, & Culpepper, 2013; Snijders, 2005; Westfall et al., 2014). Given the novelty of our theory and the adaptation of vision science techniques, we had no prior effect sizes and variance estimates on which to base an a priori power analysis. We therefore recruited what we judged to be a conservatively high number of participants, given other samples utilized in social vision research. Consistent with scientific norms in vision science, traditional work in ensemble coding used a psychophysics approach and relied on repeated measures experimental designs; furthermore, the within subject design and high trial numbers (often exceeding 10,000 trials) in these studies allows for sufficient power even with much smaller samples than we typically see in social psychological research (e.g., ranging from 2 to 55 observers at most; see Albrecht & Scholl, 2010; Alvarez & Oliva, 2010; Ariely, 2001; Bai, Leib, Puri, Whitney, & Peng, 2015; Haberman, Brady, & Alvarez, 2015; Haberman & Whitney, 2010). By such standards, we sought to recruit a large sample participants for a completely within-subjects design, given the cross-classified nature of our hypotheses. As a starting point for moving forward, we conducted post-hoc power analyses using the parameter estimates derived from our model. As shown in Table 1, post-hoc power analyses revealed we were fully powered to find

our hypothesized results (achieved power for main effects and their interaction ranged from 0.872 to 1.0).[2]

Stimuli. Stimuli included twenty social ensembles depicting 8 White faces. Ensembles varied in the ratio of men:women: 0:8, 2:6, 4:4, 6:2, 8:0 (see Figure 1). Target faces were drawn randomly from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) and arranged at random into a 2 by 4 array. This procedure was repeated four times for each condition, yielding 20 unique ensembles, presented randomly within each counterbalanced block. Each ensemble image was saved as 2386 pixel by 1791 pixel JPG image to ensure standardization across trials and participants.

Procedure. Participants learned that they would provide judgments about groups of people. Upon entering the survey hosted by Qualtrics, participants provided informed consent.

Participants provided judgments in four counterbalanced blocks. Within each block, ensembles were presented individually and in random order for 500 ms prior to a judgment prompt. Prior research in vision science has consistently relied on 500 ms display time to allow for the extraction of ensemble summaries while preventing the serial processing of individual targets (Bai et al., 2015; Haberman et al., 2015; Haberman & Whitney, 2009b; Sweeny & Whitney, 2014).

Two blocks probed participants' perceptual acuity in ensemble coding. In one block, participants estimated the Perceived Sex Ratio in each ensemble using a visual scale that depicted schematic men and women that varied in the ratio of men:women from 0:10 to 10:0,

[2] We used a custom R script to run 10,000 Monte Carlo simulations for each hypothesized model, separately testing each predicted main effect and interaction. Additional follow-up post hoc analyses revealed that, given our effect sizes and experimental design, we would have had adequate power to find our predicted effects with only 8 participants (4 men and 4 women, power = 0.96). The R script for our post-hoc power analyses can be accessed via our GitHub repository at: https://github.com/UCLASoComLab/Publications

increasing by one for each scale item (see Figure 2, panel A). In another block, participants judged the gendered appearance of the average face in an ensemble. This Perceived Facial Masculinity scale depicted 21 computer-generated faces that varied anthropometrically in face appearance from highly feminine to highly masculine (see Figure 2, panel B).

Two additional blocks probed participants' inferences about each ensemble's social attitudes and affordances. In one block, participants provided judgments about the norms within the group, defined to participants as "a group's spoken or understood rules about member behavior; they explain how members ought to act," with an example of raising one's hand in class. For each ensemble, participants rated the perceived importance of 8 norms related to sexist ideology, presented in random order, using a 7-point scale anchored by "*Extremely Unimportant to the Group*" and "*Extremely Important to the Group*" (see Appendix for complete scale). In a fourth block, participants indicated the extent to which they thought they would personally "fit in" and "belong" in each ensemble using 9-point scales anchored by "*Not at All*" and "*Extremely Well*" and "*I Wouldn't Belong at All"* and "*I Would Definitely Belong Here*," respectively. Because participants provided repeated responses within each of these trial blocks, we measured each scale's internal reliability by proportioning participants' response variance into its subcomponents and estimating how much variance was accounted for by the repeated measures design, the actual stimuli themselves, and the items within each scale (for more details on how to estimate these variances, see Cranford et al., 2006). Observed reliability of change (R_C) was 0.62 and 0.75 for judgments of norms and belonging, respectively. As such, we averaged each scale to yield two composite indices—Perceived Sexist Norms and Belonging.

Upon completion, participants provided basic demographic information and were debriefed.

Statistical Analysis. Because responses were nested within perceiver and target, we used cross-classified random coefficient models incorporating random intercepts and slopes in all analyses. Across all studies, perceiver sex moderated effects only where noted. In an effort to provide some context for the magnitude of our significant effects, we report both the Intraclass Correlation Coefficient (ICC) for the null model, as well as the pseudo R^2 statistic for each significant effect. The ICC provides an indication of how much variance in the dependent variable occurs between participants (Hox, 2002; Kreft & de Leeuw, 1998; Lahuis, Hartman, Hakoyama, & Clark, 2014; Raudenbush & Bryk, 2002); we would expect this to be generally small, as we hypothesize most of the differences in perceiver judgments will depend on changes in the participant's response to a group (e.g., a different response to seeing 0 men in an ensemble versus seeing 8 men). The ICC helps us understand how much variance in a model is explained by between-subjects as compared to within-subject effects. To provide additional context, we also report the R^2 derived from multilevel variance partitioning (R^2_{MVP}). While some debate surrounds the best way to represent effect sizes in multi-level models given the complexity of the data (Aguinis et al., 2013; Lahuis et al., 2014; Peugh, 2010; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2014), the R^2_{MVP} has been shown to be the least biased and most efficient effect estimator in cross-level, random-slope models (Lahuis et al., 2014); it should be interpreted as the amount of overall variance in the dependent variable explained by the model and thus will only be reported once per model.

We used the "lme" function with residual maximum likelihood estimation (REML) in the open-source R packages "lme4" and "lmerTest" to model our four outcomes of interest, as all our criterion were continuous (Bates, Maechler, & Bolker, 2015; Kuznetsova, Brockhoff, & Christensen, 2015). Perceiver Sex was dummy coded in all models, with 0 representing

participants who self-identified as male and 1 representing participants who self-identified as female. The Actual Sex Ratio variable was coded to reflect the number of men in each ensemble: 0, 2, 4, 6, 8. In each statistical model for Study 1a, we first tested the main effects of Actual Sex Ratio and Perceiver Sex before adding their interaction term in a stepwise fashion. Our reported results include the unstandardized regression coefficients for each effect as well as its corresponding significance tests. All data and statistical analyses scripts described in Study 1a as well as in subsequent studies are publicly available at: https://github.com/UCLASoComLab/Publications

Results and Discussion

Accuracy in encoding actuarial summaries. To test the perceptual acuity of ensemble representations, we regressed Perceived Sex Ratio and Perceived Facial Masculinity (separately) onto Perceiver Sex, the Actual Sex Ratio, and their interaction. The ICC for the null model predicting Perceived Sex Ratio and Perceived Facial Masculinity was 0.01 and 0.32 respectively; most of the observed variance in actuarial summaries occurs due to within participant differences. Testing our hypotheses, we found that as the ensemble's ratio of men to women increased, Perceived Sex Ratio also increased, B=1.03, SE=0.06, t=18.16, p<.001, R^2_{MVP}=0.78, and judgments of the average face became decidedly more masculine, B=0.73, SE=0.11, t=6.71, p<.001, R^2_{MVP}=0.37 (see Figure 3, panels A and C). Thus, both of these measures show a high level of sensitivity to the sex ratio in an ensemble, indicating that observers extract such information readily and rapidly, on the basis of 500 ms of visual exposure.

Social Attitudes and Affordances. Next, we tested our novel prediction that perceivers infer social attitudes and affordances from ensemble representations by regressing Perceived Sexist Norms and Belonging (separately) onto Actual Sex Ratio, Perceiver Sex, and their

interaction. The ICC for the null models of Perceived Sexist Norms and Belonging were 0.39 and 0.32, respectively. As the ensemble's ratio of men to women increased, groups were judged to harbor more sexist attitudes, B=0.15, SE=0.03, t=5.89, p<.001, R^2_{MVP}=0.32, and perceivers' sense of belonging increased, B=0.25, SE=0.06, t=4.50, p<.001, R^2_{MVP}=0.39. The effect on Belonging varied by Perceiver Sex, however, interaction B=-0.71, SE=0.07, t=-10.65, p<.001 (see Figure 4, panels A and C). As the ensemble's ratio of men to women increased, Belonging increased among men, but decreased among women, Bs=0.25 and -0.46, SEs=0.05 and 0.06, ts=4.68 and -8.10, ps<.001. Thus, observers formed accurate representations of a group's sex ratio and drew inferences about its attitudes from brief exposures. Male-dominated groups were judged to be especially sexist, and perceptions of personal belonging increased as representation of one's own sex increased.

Study 1b

Study 1b sought to replicate the discoveries from Study 1a employing a more rigorous design and using larger 12-person groups with variable ratios of men:women. In particular, we developed a fully randomized stimuli creation paradigm to remove (rather than control for) potential variance that could be explained by using the same target stimuli across participants. Furthermore, Study 1b tested whether ensemble coding can occur with more complex stimuli sets.

Methods

Design. Study 1b sought to replicate findings from Study 1a using novel methodology and employed a 2 (Perceiver Sex: male v. female) x 5 (Actual Sex Ratio: 0/12 men, 3/12 men, 6/12 men, 9/12 men, 12/12 men) quasi-experimental mixed-model design.

Participants. Seventy-six undergraduates from a large, public West Coast university participated in exchange for course credit. One participant who indicated their Sex to be "non-binary/genderqueer" was excluded from our analyses, leaving a final sample of 75 participants (75% women). As with Study 1a, a priori power analyses was not estimable given the theoretical novelty of our study design. Study 1b sought to improve upon Study 1a by fully randomizing each ensemble, removing the need for specifying cross-classification in the multi-level model. Since power analyses for multi-level models relies on the structure of nested data (Snijders, 2005; Westfall et al., 2014), we lacked reasonable prior estimates for effect sizes and variance estimates of data *without* cross-classification. For this reason, we again recruited what we judged to be a conservative sample size and calculated post-hoc power analyses via a customized R script. As shown in Table 1, our post hoc analyses revealed we were adequately powered to find our hypothesized effect, with achieved power for all main effects and interactions equal to or greater than 0.995.

Stimuli. In Study 1a, we manually created 20 social ensembles using a random number generator. In Study 1b, we created a custom Python script to automate the process. Individual faces were drawn randomly from the White targets within the Chicago Face Database (Ma et al., 2015) and placed into a 3 by 4 array to form 12-person social ensembles. As in Study 1a, the ratio of men:women varied: 0:12; 3:9; 6:6; 9:3; 12:0. The fully randomized stimulus generation in Study 1b occurred in real time; the custom Python script simultaneously generated and tracked the cumulative number of each Actual Sex Ratio condition each participant had seen. It drew separately from the population of men and women in the overall stimulus set to generate each unique ensemble (sampling without replacement within each trial, but with replacement between trials). Furthermore, the Python script ensured uniform presentation of individual target stimuli

such that each target face within the overall ensemble was 1280 pixels wide by 960 pixels high. Within each block, participants judged 10 unique ensembles for each Actual Sex Ratio, presented randomly, resulting in 50 unique stimuli.

Procedure. Participants reported in-person to the lab, where they provided informed consent before beginning the study on one of three iMac desktop computers. The computers used in Study 1b and all subsequent in-lab studies reported here had a 27" 5k Retina display with a 5120-by-2880 pixel resolution. Participants learned that they would provide judgments about groups of people presented briefly on the computer screen. Each trial consisted of a centered fixation cross for 500 ms, a randomly generated ensemble for 500 ms, a blank screen for 500 ms, and finally, a judgment prompt. Before beginning judgment trials, participants completed three preview trials to familiarize themselves with the procedure. Preview trials depicted individual cartoon faces from a famous, recognizable 1970's cartoon series rather than pictures of actual people and were intended to alert participants to the pace of presentation.

As in Study 1a, participants completed four blocks in counterbalanced order. In each block, participants judged a total of 50 ensembles (10 per sex ratio). Judgment items were identical to those in Study 1a, with one exception. For Perceived Sex Ratio, we updated the visual scale to depict 12 schematic men:women to reflect our increase in ensemble size. Once again, participants provided basic demographic information and were debriefed.

Statistical Analysis. We replicated our previous findings using the same analytic strategy described in Study 1a. We included an additional set of analyses in Study1b to measure the error in perceived number of men, as the Perceived Sex Ratio measure reflected the actual number of individuals in each ensemble. We calculated the Error in Perceived Number of Men for each trial by subtracting the actual number of men in the ensemble from participant's response to

Perceived Sex Ratio. This provided us with a measure of whether participants had over or underestimated the number of men in each ensemble. We also calculated the Absolute Error by taking the absolute value the difference between Perceived Sex Ratio and the actual number of men in each trial. We then separately regressed the Error in Perceived Sex Ratio and Absolute Error on to perceiver sex, ensemble sex ratio, and their interaction in a step-wise fashion. Given our full-randomization of target stimuli, we no longer needed to account for target cross-classification so the Study 1b models below specified only an observation's nesting within participant.

Results and Discussion

Accuracy in encoding actuarial summaries. The ICC, calculated based on variance components of the null model with random intercepts for Perceived Sex Ratio, was 0.00. This indicates that all the variance in responses was within participant; that is to say, all participants responded nearly identically to the measure. For Perceived Facial Masculinity, the ICC equaled 0.34. Consistent with findings from Study 1a, as an ensemble's ratio of men to women increased, Perceived Sex Ratio also increased, B=0.48, SE=0.04, t=11.66, p<.001, R^2_{MVP}=0.39, and judgments of the average face became more masculine, B=0.26, SE=0.07, t=3.54, p<.001, R^2_{MVP}=0.06 (see Figure 3, panels B and D).

The ICC for both Error in Perceived Number of Men and Absolute Error equaled 0.02, suggesting most variance in responses occurred between participants. Additionally, there was a significant main effect of ensemble sex ratio on Error in Perceived Number of Men, B=-0.50, SE=0.02, t=-23.88, p<.001, R^2_{MVP}=0.38. Participants overperceived the number of men in the ensemble, although this error decreased as the actual number of men in the ensemble increased (see Figure 5, panel A). When the ensemble was all-female, perceivers reported, on average,

seeing at least 2 men in the ensemble (M_{0men}=2.81, *SD*=2.33, *Median*=2.00). However, when the ensemble was all-male, perceivers reported, on average, seeing 3 fewer men in the ensemble than were actually present (M_{12men}=-3.42, *SD*=2.70, *Median*=-3.00). There was also a significant main effect of actual sex ratio on the Absolute Error, such that as the number of men in the ensemble increased, participants made more errors in reporting Perceived Sex Ratio, *B*=0.50, *SE*=0.02, *t*=3.21, *p*=.002, R^2_{MVP}=0.01 (see Figure 5, panel C). When there were zero men in the ensemble, perceivers were off in estimating the Perceived Sex Ratio by about 2 individuals (M_{0men}=2.81, *SD*=2.33, *Median*=2.00). The rate of misperception increased when the group was all men, with perceivers misjudging the sex ratio by about 3 individuals (M_{12men}=3.42, *SD*=2.70, *Median*=3.00). On average, across all conditions, perceivers misestimated the ratio of men to women by about 2 individuals (*M*=2.57, *SD*=2.05, *Median*=2.00).

Social Attitudes and Affordances. The ICC for Perceived Sexist Norms and Belonging in the null models equaled 0.24 and 0.28, respectively. We again replicated our findings from Study 1a; as the ensemble's ratio of men to women increased, groups were judged to harbor more sexist attitudes, *B*=0.09, *SE*=0.02, *t*=5.24, *p*<.001, R^2_{MVP}=0.16 (see Figure 4, panel B). Once again, the effect of Actual Sex Ratio on Belonging varied for men and women, interaction *B*=-0.26, *SE*=0.04, *t*=-6.76, *p*<.001, R^2_{MVP}=0.08. As an ensemble's ratio of men to women increased, Belonging increased among men, but decreased among women, *B*s=0.10 and -0.16, *SE*s=0.04 and 0.02, *t*s=2.60 and -8.10, *p*s=.018 and < .001 (see Figure 4, panel D).

Collectively, Studies 1a and 1b found that ensemble representations accurately reflect the sex ratio of groups and that variations in sex ratio inform inferences about a group's sexist attitudes and personal belonging. That said, these studies could not test causal links between these judgments because of the randomization of blocks, a temporal factor that precludes tests of

mediation (Baron & Kenny, 1986). We set out to test our hypothesis that actuarial summaries of a group drive perceivers' feelings of belonging in Study 2.

Study 2

In Studies 1a and 1b, we demonstrated that perceivers accurately extract actuarial summaries from a group of people. Furthermore, we showed that the group's sex ratio significantly impacts feelings of belonging and perceptions of sexist norms. Although offering convergent evidence that ensemble representations arise from instantaneous judgments about a group, Studies 1a and 1b could not offer insight into the underlying mechanisms; in Study 1a, while targets were identical across blocks, the data's cross-classification structure prevented us from running a mediation model testing our hypothesis. Study 1b no longer utilized a cross-classified structure, however, participants did not see the same targets across blocks because each stimulus was unique, preventing us from yoking participants' responses to a given image. To more directly test our hypothesis required a test of moderated mediation, so we designed Study 2 to avoid cross-classification and allow participants to see the same ensembles across blocks.

Consistent with Studies 1a and 1b, we expected both male and female perceivers would accurately perceive increases in the ensemble sex ratio as the actual number of men in the ensemble increased. We predict that these actuarial summaries would differentially mediate the effect of the Actual Sex Ratio on Belonging.

Methods

Design. Study 2 aimed to test the mechanism by which differences in an ensemble's sex ratio impacts one's sense of belonging within the group. Study 2 employed a 2 (Perceiver Sex: male v. female) x 5 (Actual Sex Ratio: 0/12 men, 3/12 men, 6/12 men, 9/12 men, 12/12 men)

quasi-experimental mixed-model design in which participants provided multiple judgments for each stimulus.

Participants. Estimating desired sample size a priori for multi-level mediation analyses is a notoriously difficult task; while custom scripts are recommended as best practice to estimate sample size for a single multilevel regression equation (Muthén & Muthén, 2002), mediation analysis requires knowing threefold the amount of information a priori (Krull & MacKinnon, 1999). Additionally, mediation can occur at the individual and/or group level (e.g., for a given observation as well as for a participant across trials; Kenny, Kashy, & Bolger, 1998; Kenny, Korchmaros, & Bolger, 2003), and variance matrices will, as a result, differ across each component of the mediation analysis (Krull & MacKinnon, 1999). At the time of data collection, there was no readily available script or practical guidelines for estimating a priori sample size for multilevel mediation analysis. Solving this problem was beyond the scope and purpose of this paper. Thus, we recruited what we considered a conservative sample size, based off recommendations from Krull and MacKinnon (1999); according to their simulations, larger sample sizes (e.g., 50 participants or more) and greater than 30 trials per person decreased bias in estimating the coefficients in the mediator model. Based on their simulation findings, we recruited fifty-eight college undergraduates (81% women) who completed Study 2 in a laboratory setting in exchange for course credit. Participants were only eligible if they had not participated in Study 1b.

Stimuli. Testing our proposed mediation required us to eliminate the stimulus cross-classification in our design while still ensuring each participant saw the same unique-to-them set of ensembles. We therefore created a custom Python script that randomly generated and stored 100 unique ensembles (20 per condition) as described in Study 1b. Here, however, the exact

ensembles that were generated in the first test block were repeated in subsequent blocks. Thus, each participant viewed a randomly generated set of ensembles that was repeated within participant throughout subsequent blocks. This continuity in the design was essential for testing our hypotheses involving multilevel mediation without stimulus cross-classification.

Procedure. After providing informed consent, participants completed a series of preview trials identical to those used in Study 1b. After completing the preview trials, participants completed two blocks of trials in which they judged a total of 100 ensembles (20 per sex ratio) for Perceived Sex Ratio and Belonging. The presentation order of dependent variables remained the same across participants. All other aspects of stimulus presentation were identical to Study 1b.

Statistical Analysis. We used the statistical R package "mediation" (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) to test how an ensemble's sex ratio compels a sense of belonging. We recoded both Actual Sex Ratio and Perceived Sex Ratio to reflect the number of same-sex faces in each ensemble, relative to Perceiver Sex (e.g., ensembles of 3 men:9 women were coded as 3 for male participants and 9 for female participants), hereafter Actual and Perceived Same-Sex Others, respectively. This allowed us to collapse across Perceiver Sex to test a common causal process for both men and women. Our mediation model tested whether the relation between Actual Same-Sex Others and Belonging was mediated by Perceived Same-Sex Others. This model therefore tests whether one's sense of belonging stems, at least in part, from the first glimpse of a group of people.

The "mediation" analysis package in R requires the user to specify a control group and a comparison group. We compared the proposed mediation model across four separate contrasts: 0 v. 3 Same-Sex Others; 0 v. 6 Same-Sex Others; 0 v. 9 Same-Sex Others; and 0 v. 12 Same-Sex

Others. Finally, we analyzed the Error in Perceived Number of Men and Absolute Error as outlined in Study 1b.

Results and Discussion

We analyzed judgments using a multilevel mediation model. Using a quasi-Bayesian Monte Carlo approximation with 10,000 simulations, we estimated average direct effects and the average casual mediation effects of Actual and Perceived Same-Sex Others respectively on Belonging (Tingley et al., 2014). As predicted, the effect of Actual Same-Sex Others on Belonging was mediated by Perceived Same-Sex Others in the ensemble. Across all four contrast comparisons, Perceived Same-Sex Others partially mediated the effect of Actual Same-Sex Others on Belonging (see Table 2). The proportion mediated and relative effects were similarly sized across contrasts, suggesting a steady influence of actuarial statistics on judgments of social affordance. Of note, the perception of just three Same-Sex Others in a twelve-person ensemble group (proportionately only 25%) was sufficient to increase perceivers' sense of belonging (see Figure 6). As the number of Actual Same-Sex Others in the group increased, so did judgments of Perceived Same-Sex Others, again indicating sensitivity to the sex ratio of ensembles. These percepts accounted for 14% of the variance between Actual Same-Sex Others and Belonging.

The ICC for Error in Perceived Number of Men and Absolute Error equaled 0.03 for both null models. As in Study 1b, as the number of men in the ensemble increased, participants overperceived the number of men in the ensemble less, B=-0.48, SE=0.02, t=-24.36, p<.0001, R^2_{MVP}=0.41 (see Figure 5, panel B). When the ensemble was all-women, participants reported seeing about 2 men more than were present ($M_{0\text{men}}$=2.53, SD=2.05, $Median$=2.00). However, when the ensemble was all-male, participants reported seeing about three fewer men in the ensemble than were present ($M_{12\text{men}}$=-3.02, SD=2.39, $Median$=-3.00). Additionally, the Absolute

Error significantly increased as the ratio of men to women in the ensemble increased, B=0.05, SE=0.01, t=3.65, p=.000566, R^2_{MVP}=0.01 (see Figure 5, panel D). Participants miscalculated the Perceived Sex Ratio by about two individuals when the ensemble was all-male (M_{0men}=2.53, SD=2.05, $Median$=2.00) and by about three individuals when the ensemble was all-female (M_{12men}=3.02, SD=3.02, $Median$=3.00). On average, across all conditions, perceivers misestimated the ratio of men to women by about 2 individuals (M=2.44, SD=1.96, $Median$=2.00).

Study 3

Having established that perceivers accurately represent social ensembles within two individuals and use these percepts to draw inferences about groups' social attitudes, we sought to better characterize the mental representations resulting from ensemble coding. In Study 3, we used reverse correlation, a method that provides a visual approximation of participants' mental representations (e.g., an ethnic in-group member; Dotsch & Todorov, 2012; Dotsch, Wigboldus, Langner, & Van Knippenberg, 2008; Ratner, Dotsch, Wigboldus, van Knippenberg, & Amodio, 2014), to test how sex ratio affects one's mental representation of the average group member's face. Predominantly used by researchers in social cognition, reverse correlation methodology has also been successfully employed to test questions pertaining to perceptual judgments in vision science (e.g., configural face processing; Sekuler, Gaspar, Gold, & Bennett, 2004). The images used in reverse correlation derive from the same base face, with random patterns of sinusoidal noise added or subtracted to them; when collapsing across participants' trials, the random noise cancels out and leaves an image that has been systemically, albeit unconsciously, selected by the participant. While Studies 1 and 2 allowed us to examine explicit belonging based on actuarial

summaries, Study 3 allowed us to visualize perceivers' mental representations of groups varying in their sex ratio.

Stage 1 Materials: Generating mental representations of the average group member

Design. Study 3 employed a between-subject design in which we manipulated Condition; participants saw only majority-male (9 men, 3 women) or only majority-female (3 men, 9 women) twelve-person ensembles.

Participants. Prior studies employing reverse correlation have ranged broadly in sample size (e.g., from 28 to 176 participants; see Dotsch & Todorov, 2012; Dotsch et al., 2008; Ratner et al., 2014). While the first study to apply reverse correlation to group targets, we based our sample size off the approximate average sample size from prior reverse correlation studies. We thus recruited one hundred and ten college students, who received course credit for completing the study. We excluded data from 16 participants (1 due to computer failure, 3 did not follow instructions, and 12 had recently completed a near-identical task with full debriefing), leaving a final sample size of 95 (63% women).

Stimuli. Participants were randomly assigned to view ensembles that consisted of a majority men or majority women. A custom Python program fully randomized the selection of faces and their placement within the 12-person ensemble group. White male and female photographs were drawn without replacement within trial, but with replacement between trial, from the Chicago Face Database (Ma et al., 2015).

Reverse Correlation Image Classification Task. Using the R package "rcicr" (Dotsch, 2016), we generated 700 pairs of greyscale images that depicted variations of an androgynous base face. These image pairs served as the forced-choice alternatives for our measure. This technique imposes visual noise over the base image, thus occluding it in systematic ways. Each

image pair was constructed by both adding and subtracting a randomly generated visual noise pattern to/from the base image. Consistent with prior methods (Dotsch & Todorov, 2012; Dotsch et al., 2008; Ratner et al., 2014), we presented these image pairs simultaneously.

Procedure. In each of 700 trials, an ensemble appeared for 500 ms followed by a pair of reverse correlation test images. Participants indicated which of the two images best represented the average face in the group (see Figure 7, panel A). Using the R package "rcicr" (Dotsch, 2015), we aggregated the noise patterns first across trials within a participant and then across participants to obtain a composite image that visualized the mental representation of the average group member for majority-male and majority-female ensembles (see Figure 7, panels B and C).

Stage 2 Materials: Evaluating mental representations of the average group member from majority-male and majority-female ensembles

Design. We compared the mental representations from majority-male and majority-female ensembles generated in Stage 1 using a within-subject design. An independent group of participants categorized each composite image as male or female and rated them on interpersonal characteristics (e.g., Anger, Warmth, Approachability). We predicted that judgments obtained for composite images in Stage 2 would vary according to the sex ratios of ensembles presented in Stage 1.

Participants. We recruited one hundred and forty-two U.S.-based Internet users to complete our study from mTurk. Our sample size is consistent with Stage 2 sample sizes from prior reverse correlation studies (see Dotsch & Todorov, 2012; Dotsch et al., 2008; Ratner et al., 2014). We excluded data from 22 participants (18 failed the manipulation check, 3 failed to indicate their gender, and 1 had an identical IP address as an earlier respondent). This left a final sample of 120 participants (45% women).

Procedure. In Stage 2, participants judged the two composite images from Stage 1 in three counterbalanced blocks. In the Perceived Sex Category block, participants categorized the two mental representations as male or female. In the trait judgments block, participants judged how angry, friendly, warm, approachable, intelligent, competent, happy, and attractive each mental representation appeared using a series of 7-point Likert scales (anchored by 1= "*not at all*" and 7= "*very*"). Finally, in the Perceived Facial Masculinity block, participants indicated how masculine or feminine each mental representation appeared using a 7-point Likert scale (anchored by 1= "*extremely feminine*" to 7= "*extremely masculine*").

Statistical Analysis. To test whether gender-linked judgments reflected the Actual Sex Ratio of ensembles in Stage 1, we regressed Perceived Sex Category (using conditional logistic regression) onto Condition. We used paired samples t-tests to independently compare trait judgments and Perceived Facial Masculinity of the two mental representations. We used a Bonferroni correction to ensure the family-wise alpha level did not rise above 0.05; to do this, we divided the desired family-wise alpha-level of 0.05 by the number of paired samples t-tests (n=9) we were running to arrive at the revised significance level of less than or equal to 0.0056. We used this corrected alpha level when deciding whether a given trait judgment varied by Condition in Stage 2. Furthermore, perceiver sex did not moderate the effect of condition on any of the nine trait judgments.

Results and Discussion

First, we regressed Perceived Sex (using conditional logistic regression) and Perceived Facial Masculinity (paired samples t-test) onto Condition. Relative to judgments of the composite image derived from majority-female ensembles, the composite derived from majority-male ensembles was 4.67 times more likely to be categorized as male, B=1.54, z=2.42, p=.0155,

and was rated as significantly more masculine (see Table 3). Similar differences were obtained for gender-stereotyped judgments using paired samples t-tests with Bonferroni correction, such that composite images derived from majority-male, relative to majority-female, ensembles were rated as angrier, less warm, less approachable, less friendly, less happy, and less attractive. We observed no significant difference in perceived competence or intelligence between mental representations from majority-male and majority-female ensembles (see Table 3).

These findings reveal that, in addition to compelling accurate representations of a groups' sex ratio, conveying information about a group's attitudes, and informing one's sense of belonging within a group, ensemble coding also alters more implicit mental representations. As such, the mental representation of an average group member is imbued with gendered-linked appearance cues, including stereotypic cues to warmth, approachability, and anger.

General Discussion

Taken together, our findings indicate perceivers accurately discern the sex composition of a group of faces after brief visual exposure, forming mental representations of the average group member that embody gendered features and stereotypes (Broverman, Vogel, Broverman, Clarkson, & Rosenkrantz, 1972; Wood & Eagly, 2010). Studies 1a and 1b served first as successful conceptual replications of prior work in vision science, demonstrating perceivers accurately encode high-level information about a group's composition. Furthermore, we also found that these split-second percepts gave rise to more abstract social evaluative judgments. Perceivers felt a greater sense of belonging as members of their own sex increased numerically, and both men and women inferred that male dominated groups were likely to harbor sexist attitudes. Thus, we discovered that perceivers deduce group norms even in the absence of interaction with other group members; the ratio of men to women in an ensemble sends a rapid,

strong nonverbal signal about whether future gender discrimination is likely to occur. Study 2 demonstrated that perceptions of same-sex others in an ensemble partially drive observers' feelings of fit and belonging; our social attitudes and affordances shift in response to actuarial summaries of groups, even from brief visual exposures. Finally, in Study 3, reverse correlation techniques revealed perceivers' mental representations of the average group member and found that representations of majority-female ensembles tend to be perceived as more welcoming and approachable than the representations of majority-male ensembles. Collectively, findings from our four studies illuminate the efficiency with which human perceivers extract visual information about a group and use it as a foundation on which to base important social judgments.

Implications for Vision Science

These findings have important implications across multiple domains. Methodologically, this research stands to enrich the experimental repertoire available to both vision scientists and social psychologists alike. Drawing inspiration from early studies on ensemble coding, we developed new experimental protocols that help push the boundaries of vision science methodology. Using novel methods that we adapted from prior work utilizing a psychophysics paradigm (e.g., Haberman & Whitney, 2009a), we found that perceivers can accurately achieve summary percepts for a groups of individuals. While perhaps unsurprising, this provides an important conceptual replication of prior work using a more widely accessible alternate method that affords testing of larger sample sizes. Additionally, the range of measures explored throughout our work expands the scope of inquiry available to vision scientists that heretofore have focused exclusively on narrowly-focused aspects of perceptual accuracy. The development of these methods corroborates the robustness of the ensemble coding phenomena, thereby complementing and expanding upon earlier research.

Theoretically, this research also expands on prior work in vision science by demonstrating that, in addition to supporting accurate visual percepts, a mere glimpse of a group is also sufficient for perceivers to infer a broader array of social information. These overarching social impressions appear to occur just as rapidly and readily as the visual estimates, sparked by only a half-second visual exposure to an image. Thus, in addition to encoding the average distribution of various social categories within an ensemble (e.g., sex, Haberman & Whitney, 2007; or race, Thornton et al., 2014), observers also extract social meaning. Perceivers in our studies accurately extracted the *ratio* of men to women in an ensemble, and they used that knowledge to infer the groups' attitudes and estimate their own fit within the group. Their ability to do so was not immediately evident from prior research in vision science; we present novel evidence of the downstream social consequences of ensemble coding, with our findings suggesting that complex social inferences occur alongside lower-level processing of groups from the briefest of exposures. Future researchers may seek to tease apart the precise threshold at which social attitudes shift as a result of changes in the number of same-sex others in the group. While we found adding three same-sex others to an ensemble significantly increased felt belonging and perceived discriminatory norms, it may be that in much larger groups many more same-sex others are needed to change perceivers' social affordances. It seems unlikely that seeing only three same-sex others in a 100-person ensemble will elicit similar feelings of belonging. It may be, instead, that a particular ratio of same-sex to dissimilar others must be achieved before perceivers feel like they belong.

Although our research draws on methodology and theory promoted by prior ensemble coding researchers, an alternative hypothesis from vision science could partially explain our results. To date, little work has attempted to delineate the difference between numerosity and

ensemble coding. Numerosity, the ability to infer the relative number of objects in a set, has been demonstrated to be a robust primary visual process coming online developmentally early (Burr & Ross, 2008; Ross & Burr, 2010; Xu, Spelke, & Goddard, 2005). Developing the theory of numerosity, Burr and Ross (2008) conclude that, like color or size, humans have a general sense of "numberness"; perceivers can guess roughly how many objects comprise a set, although environmental factors influence their accuracy.

According to the alterative numerosity hypothesis, perceivers in Studies 1 and 2 could have based their responses to the dependent variables on a general sense of the number of men to women in each group, rather than extracting an actuarial summary of the ratio or average facial masculinity. However, most prior research on numerosity has asked participants to make comparative judgments between sets of objects (e.g., which of two images contains more circles, Kirjakovski & Matsumoto, 2016) rather than reporting the actual ratio or perceived average of items within the set. Furthermore, our work differs from prior research in numerosity by employing reverse correlation. It is unlikely that a numerical sense of the number of men and women in the ensemble could lead to the systemic choice preferences in average group member seen in Study 3. Instead, it seems more plausible that perceivers draw on a mental representation extracted from the composite target images when choosing from the test pair images. Indeed, this explanation is in-line with the theory underlying reverse correlation as a methodology, which posits that visual categorization relies on a holistic extraction of features rather than the summary of individual parts (Mangini & Biederman, 2004). It is a limitation of the current work—albeit one reflected more broadly in the field—that we cannot say without a doubt whether numerosity or ensemble coding processes underlie perceivers' response to groups varying in the ratio of men to women. Despite this limitation, the social consequences of seeing similar or dissimilar others

in a group remain the same: namely, increases in the number of men to women in an ensemble affect feelings of fit and lead all perceivers to rate the group as more likely to endorse sexist norms. While outside the scope of the current research, differentiating between numerosity and ensemble coding processes' effects on social judgments constitutes a very interesting and important avenue for future research.

Implications for Social Cognition

The current research also provides a viable means by which social cognitive scholars can probe the perceptual underpinnings of group judgments. Whereas it is widely recognized that the composition of a group is likely to impact social behaviors (Hackman, 1992), a majority of those observations have utilized immersive experience (e.g., Seger, Smith, Kinias, & Mackie, 2009). Here, we demonstrated that seeing a group, devoid of any other information, can be sufficient to influence perceivers' social judgments, using both explicit and indirect measures (e.g., reverse correlation). Additionally, from a theoretical perspective, our research highlights that inferences about a group's social attitudes and affordances may arise earlier than we previously knew. A perceiver need not be immersed in an environment, real or imagined, to feel the weight of being outnumbered; merely a glimpse of a group triggered consequential social percepts, including a broad sense of social belonging and inferences about gender-based attitudes. As such, this work complements prior observations that prolonged exposure to a seemingly hostile environment (e.g., women taking a math test in a room with geek paraphernalia; Cheryan et al., 2009) can lead to performance decrements and changes in gender attitudes (Pronin et al., 2004), by providing evidence for immediate shifts in social cognitions upon first espying a group. Our visual environment, including the people we see, rapidly and strongly impacts our perceptions of whether we belong and what behaviors to expect.

These findings form a foundation for answering more nuanced questions about group perception. To be sure, our manipulation of the sex ratio of ensembles provides an important first step toward understanding the broader social implications of group perception, but it is far from exhaustive. With these insights in hand, we are now poised to probe a range of factors that differ among individuals that inhabit a group and how they uniquely impinge on percepts. For instance, given that impressions of *individuals* exhibit a tight perceptual tethering between gender and race categories (Johnson, Freeman, & Pauker, 2012; Carpinella, Chen, Hamilton, & Johnson, 2015), percepts of ensembles that vary systematically along both dimensions simultaneously are likely to vary accordingly. The gendering of race may therefore bias both the accuracy of actuarial measures and the extremity of social inferences, depending on the specific racial composition of the group. Testing intersectional ensemble coding would provide a unique opportunity for future researchers to understand the relationship between race and gender in visual processing.

Additionally, our findings also provide a foundation for expanding the scope of inquiry to probe behavioral approach and avoidance tendencies when perceiving personally relevant or contextually situated groups. For example, based merely on a glimpse of a male-dominated STEM class, a woman may draw meaningful inferences that could impact her educational and professional trajectory: first in her perception of the sex ratio; followed by her impression of group norms and stereotypes; and finally manifesting in inferences about her own fit in the group. Similarly, percepts of real-world groups with established norms or and/or visual similarities (e.g., a math club or sports team) might yield more extreme summary impressions than those we have observed based on a random selection of faces, in much the same way that described similarity enhances judgments of entitavity (Hamilton et al., 2015). If correct, this implies that our findings may even underestimate the potency of a spontaneous impression of a

group. Thus, the visual heuristics that ease the task of social perception may fluctuate in strength as a function of personal relevance of social context, thus holding the potential to incur interpersonal costs well before observers interact with group members. This approach provides a viable way for testing such implications.

We wish to clarify, furthermore, that we do not believe this phenomenon unique to women. As our findings demonstrate, men also feel decreased fit in majority-female groups from only a mere glimpse. Although mental representations of majority-female groups may be perceived as less threatening overall (Alt, Goodale, Lick, & Johnson, 2017), nevertheless gender disparity may cause men to avoid female-dominated fields. In the 2014-2015 academic year, for example, only 12% of the Bachelor's degrees in registered nursing were awarded to men (U.S. Department of Education, 2017). Other fields stereotyped as more feminine in nature, like education or psychology, show a similar disparity in educational achievement (Forsman & Barth, 2017; Shen-Miller & Smiler, 2015; U.S. Department of Education, 2017). Although researchers in the last decade have focused primarily on increasing women's participation in historically male-dominated fields (Shen-Miller & Smiler, 2015), our findings shed light on a broader mechanism by which men in female-dominated fields may feel similarly disadvantaged and out of place. Increasing gender parity across fields may positively boost both men and women's participation, with further research needed to probe the long-term effects of seeing similar or dissimilar others on vocational achievement.

Concluding Remarks

Our research provides the first empirical evidence confirming not only that the ratio of men to women in a group is readily and accurately perceived, but also that it drives feelings of fit and belonging within the group and broader inferences about the group. Given these findings,

perhaps it is unsurprising that Judge Ginsburg, as one woman surrounded by eight men, felt "all alone." Indeed, such feelings are likely to be shared by naïve observers of any group exhibiting such an extreme ratio of men to women. Our findings also imply an important remedy. Each and every step toward equal representation on the Supreme Court (or any group for that matter) produces a commensurate shift in observers' actuarial representations of and interpersonal inferences about the group. In this same interview, Justice Ginsburg opined, "People ask me…When will there be enough women on the [Supreme] Court?" Her answer underscored the potency of the relationship between sex ratio and belonging: "When there are *nine*."

Study Set Two: Ensemble Coding in the Wild

Ensemble coding in the wild: Examining the effect of sex ratio on perceived competence and belonging among real-world groups

Brianna M. Goodale and Kerri L. Johnson

University of California, Los Angeles

Ensemble coding in the wild: Examining the effect of sex ratio on perceived competence and belonging among real-world groups

Study 4a

Study 4a sought to replicate with a smaller, real-world ensemble the findings from Studies 1-3. I was interested in understanding whether ensemble coding can occur with noisier stimuli sets, attempting to increase the external validity of our previous findings. I expected perceivers to extract similar summary statistics from a naturally-occurring group. In particular, I predicted that as the number of men in the group increases, perceivers would: a) accurately report a more male-dominated sex ratio; b) choose a more masculine face as representing the average group member; and, c) expect the group to have more sexist norms. Additionally, I hypothesized an interaction between perceiver sex and the group's sex ratio, such that perceivers would feel greater belong as the number of same-sex group members increased.

To answer my research questions, I drew on groups socially relevant to psychologists and about whom public information exists: conference panels. Presenting at academic conferences marks an opportunity to share one's research with a broader audience and provides a pulse on what topics are popular within the field. Most conferences offer online or in print copies of their programs. Making the materials publicly available allows anyone to go through and categorize panel talks. Curious about norms and perceptions of ensembles in a field dedicated to better understanding human interactions, I chose to examine panels from the 2015 Annual Meeting of the Society of Personality and Social Psychology (SPSP). According to the SPSP conference website (2015b), the organization reviews conference submissions based on several factors including: "scholarly/theoretical merit", "relevance to social and personality psychology," "significance, and originality." The submission guidelines note as well that, "Final selection

among submissions deemed meritorious will be made with an eye toward achieving a balanced and broadly representative program." The submission guidelines do not indicate that the review process is blind; instead, reviewers and conference organizers may use other characteristics to decide which panels should be invited to present. I wanted to probe what information could be extracted from the ensembles and how their sex ratios contribute to impressions of the group's work as original, meritorious, and significant to the field. Using the actual panels selected to present at the 2015 SPSP meeting, Study 4a explored how ensemble coding may impact perceptions of real-world groups.

Methods

Participants. Eighty-four college undergraduates from a large, West Coast university completed the study in exchange for course credit. I excluded three participants who reported identifying outside the gender binary from the analyses. Additionally, computer error caused one participants' data not to save. Eighty participants (78% female) comprised the final sample.

Materials and Design. Study 4a followed a 2 (perceiver sex: male v. female) x 5 (actual sex ratio of men/women in ensemble: 0/4 v. 1/3 v. 2/2 v. 3/1 v. 4/1) quasi-experimental, mixed model design. Participants saw a series of four-person ensembles drawn from the actual panels of presenters who participated in the 2015 Annual Meeting for the Society of Personality and Social Psychology (SPSP).

Target stimuli creation. Because Study 4a aimed to study naturally-occurring groups, the number of 4-person ensembles in each condition followed the actual distribution of 4-person panels at the 2015 SPSP meetings. The 2015 SPSP program listed a total of 81 4-person panels: 11 all-female ensembles; 16 ensembles with 1 man; 25 ensembles with 2 men; 23 ensembles with 3 men; and, 6 all-male ensembles. Seven undergraduate research assistants blind to

hypotheses created a password-protected database of panels from the 2015 SPSP meeting based on the program given to attendees. They then Googled each of the four presenters in a given panel (identified as the first-author listed for each talk) and pulled the most recent headshot they could find for that presenter. When possible, forward looking, close-up images of the presenter's head and shoulder were used. Each picture was labeled with a letter and number corresponding to their panel presentation date and time, as well as their order in the presentation; research assistants recorded this code next to the presenter's name in the database. This allowed experimenters to double check the identity of the person photographed without compromising the target's identity during the actual experiment or subsequent data analysis. A total of 323 presenters' photos were collected via this method. One presenter did not have a headshot available online; thus, his panel was excluded from the study design and subsequent analyses.

An eighth research assistant also blind to hypotheses then standardized each presenter's headshot using Adobe Photoshop. All images were cropped into an oval around the target's face, removing as much of the hairline, ears, and neck as possible. Editing software within Photoshop was also used to correct tilted heads so that each final image had completely vertical features. The oval was then placed onto a white background and cropped into a 144 x 144 pixel square. The facial oval for each image reached from the top to bottom of the square, measuring 12 pixels in height. Each standardized face was then sorted back into its respective digital folder with the other presenters who served on their same panel. This left a final sample of 80 target 4-person panels: 11 all-women panels; 16 panels with 1 man; 25 panels with 2 men and 2 women; 22 panels with 3 men and 1 woman; and, 6 all-men panels.

Pre-testing individual target faces. To verify perceived target sex and collect potential covariate information, Amazon Mechanical Turk (mTurk) workers (n=377) made trait judgments

about the standardized individual target faces in exchange for $0.50. The pre-test participants rated each of the 320 targets in randomized order. Participants were randomly assigned one of four dependent variable blocks in an effort to minimize experiment fatigue. Each target image appeared one at a time onscreen above the dependent variable; it remained visible until the participant indicated his or her answer.

In one block, participants (n=94) categorized the target's sex as male or female in a two-alternative forced choice (2AFC) task. For each target image, I used a χ^2 difference test to assess whether the proportion of respondents correctly identifying the target's sex was significantly different than chance (e.g., 50%). I found participants correctly identified the target sex better than chance (all p's $\leq$.05) for all but three images. [3]

The second block asked participants (n=94) to rate the attractiveness of each individual target face on a 7-point Likert scale (1=*Extremely unattractive* to 7=*Extremely attractive*). I averaged together the participant-mean attractiveness ratings for each of the four panel members to arrive at a rating of overall ensemble attractiveness. By asking pre-test participants about these trait judgments instead of the main study sample, I created normed, control data impervious to ensemble presentation.[4] The pre-test averages of ensemble attractiveness comprised an important covariate included in subsequent analyses of my main hypotheses.

[3] Each of the three target images were categorized by participants no differently than one would expect due to random chance ($p\geq$.07). I ran all subsequent analyses with and without the 3 target ensembles included. In all cases, the results remained unchanged. Thus, I opted to present the results including the full 80 target images in this paper, as it constitutes a more conservative test of our hypotheses.

[4] Although yet to be tested empirically, I expected a group-level biasing of individual trait judgments when targets are presented together in an ensemble. I predict this effect will appear similar to aftereffects, with perceivers seeing a single individual in ensemble as more attractive if the group has a higher average attractiveness rating. I offered a more conservative measure of covariates by soliciting ratings for each individual target before averaging them across the ensemble. Future research by myself and colleagues will examine this proposed group-derived shift in individual trait perceptions.

The third pre-test block asked participants to estimate each target's age. Since older individuals are stereotyped as wiser or more knowledgeable (Brewer, Dull, & Lui, 1981; Levy, 1996, 2009; Levy, Ashman, & Dror, 2000), I wanted to control for the perceived mean age of ensemble targets in the main analyses. Unfortunately, participants (n=94) were unwilling or unable to provide accurate estimations of targets' age; their responses ranged in value from 2 to 31 years old. It seems unlikely that participants actually considered the target a toddler or pre-teen. It may be that perceivers felt more uncomfortable making age judgments compared to other pre-test tasks. Age is heavily stigmatized in contemporary American culture with older adults perceived as low-status and more likely to face discrimination (see Garstka, Schmitt, Branscombe, & Hummert, 2004); this may have rendered perceivers less willing to differentiate between younger and older targets. Alternatively, the pre-test age judgments may have proved more difficult to make than judgments about target's sex or attractiveness. In the first two blocks, participants chose from one of several options. However, in the age judgment block, questions were formatted as free-response. It may have taken longer for participants to type in an estimate for each target. Thus, perceivers in the age judgment block may have experienced greater experimental fatigue than perceivers in the sex categorization or attractiveness rating blocks. Given the wide variability in participants' responses for this block, I was unable to test age as a potential covariate in subsequent analyses.

The fourth pre-test block asked participants (n=95) to rate how feminine or masculine they found each target face on a 7-point Likert scale (1=*Extremely feminine* to 7=*Extremely masculine*). Although outside the scope of my primary research question, I am interested in potential differences between an ensemble's rating of target masculinity when presented in an

ensemble (e.g., as 1 4-person target group) versus when presented alone. I have not analyzed this data for the current paper and will not be presenting them here.

Ensemble stimuli generation. Having generated and standardized the individual target faces for each 4-person panel, a custom-made Python script then assembled the individual pictures into their respective ensembles in real-time during the main experiment. The script placed each individual presenter in a panel into a 1x4 array. To maximize external validity, within each ensemble picture order followed the presentation order as outlined in the 2015 SPSP program. A total of 80 ensemble stimuli were generated, matching the number of 4-person panels that presented in the 2015 meeting. See Figure 8, panels A-E, for sample stimuli matching each condition.

Procedure. Upon providing informed consent, participants were escorted by the experimenter to a computer on which the custom Python script was preloaded. They completed a similar set of test trials as outlined in Studies 1b and 2 to familiarize themselves with the task; to better mirror the actual trials, the test trials in Study Set Two were comprised of only 4 cartoon figures arranged in a 1x4 array instead of 12 cartoons in a 3x4 array. The 80 panel-based ensembles were then presented in random order in each of four counterbalanced blocks.

The first block measured Perceived Sex Ratio. It was identical to the Perceived Sex Ratio block used in Study 1b with the exception that the stick figures appeared in groups of four. Thus, participants selected the stick figure group best approximating the ensemble they last saw from one of five possible images.

The second block was identical to the Belonging block from Studies 1a, 1b, and 2. Participants indicated on two separate questions how much they felt like they "fit" with the group and how much they belonged in the group on 9-point Likert scales. Their answers to these

two questions were averaged together to give an overall measure of belonging (measure of internal reliability R_c=0.97; see Cranford et al., 2006 for more detail on how to calculate multi-level scale reliability).

Blocks three and four were also identical to measures employed in Studies 1a and 1b. Block three measured Perceived Sexist Norms via an eight-item scale. Reliability analysis for the eight-item scale revealed that the measures hung together moderately well (R_C=0.66) and justified using a single mean average for Perceived Sexist Norms. The final block measured Perceived Facial Masculinity. It employed the same 21-point FaceGen scale as in Studies 1a, 1b, and 2; participants indicated which picture best aligned with the face of the "average group member" from the prior ensemble. After participants completed all four blocks, they answered basic demographic questions.

Statistical Analysis. I analyzed all results using stepwise, multi-level mixed effects models with random effects and random slopes. Because observations were nested within participant as well as within stimuli, the models were also cross-classified. I separately regressed each of the four criterion variables onto perceiver sex, actual sex ratio, and their interaction. Perceiver sex only moderated the main effects of condition where noted. For each analysis, I also tested the model controlling for average ensemble attractiveness, specifying it as a random slope within the target stimuli term. I also examined the effect of perceiver sex, Actual Sex Ratio and their interaction on the error in estimating the Perceived Sex Ratio. To calculate the Error in Perceived Sex Ratio, I subtracted the Actual Sex Ratio from the Perceived Sex Ratio for each ensemble. Finally, I also took the absolute value of this difference to arrive at the Absolute Error, which I then tested for condition and perceiver sex effects.

Effect sizes were calculated via the Intraclass Correlation Coefficient (ICC) for the null model as well as the pseudo R^2 statistic (R^2_{MVP}) for the full model (see Lahuis et al., 2014). All analyses were run in R, using the "lme" function with residual likelihood estimation (REML) from the open-source R packages "lme4" and "lmerTest" (Bates et al., 2015; Kuznetsova et al., 2015). Perceiver sex was dummy coded in all models, with 0 representing participants who self-identified as male and 1 representing participants who self-identified as female. The Actual Sex Ratio variable was coded to reflect the number of men in each ensemble: 0, 1, 2, 3, or 4. The reported results include the unstandardized regression coefficients for each effect as well as its corresponding significance tests.

Results

Accuracy in encoding actuarial summaries. I calculated the ICC, which equaled 0.06, based on variance components of the null model with random intercepts for Perceived Sex Ratio. This indicates that 6% of the variance in responses was due within subject differences; most participants responded similarly to each ensemble across conditions. Study 4a replicated findings from Studies 1a and 1b with minimal groups. There was as significant main effect of condition on Perceived Sex Ratio, B=0.72, SE=0.03, t=21.18, p<.001, R^2_{MVP}= 0.55. As the number of men in the ensemble increased, perceivers correspondingly picked more male-dominated ratio representations. Female perceivers reported perceiving significantly more men in each ensemble than male perceivers, B=0.26, SE=0.07, t=3.54, p<.001; this effect, however, was relatively small (R^2_{MVP}=0.01; see Figure 9, panel A).

The ICC for Perceived Facial Masculinity equaled 0.41. Consistent with hypotheses and prior findings, as an ensemble's ratio of men to women increased, judgments of the average face became more masculine, B=0.54, SE=0.14, t=3.83, p<.001, R^2_{MVP}=0.04 (see Figure 9, panel C).

Almost all of the variance in the null models for Error in Perceived Sex Ratio and Absolute Error occurred between participants (ICCs=0.10 and 0.08, respectively). As the ratio of men to women in the ensemble increased, perceivers overestimated the number of men in the ensemble significantly less, B=-0.28, SE=0.02, t=-16.02, p<.0001, R^2_{MVP}=0.13 (see Figure 10, panel A). When the ensemble was all-female, participants were mostly accurate in their selection of Perceived Sex Ratio ($M_{0\text{men}}$=0.74, SD=0.94, $Median$=0.00). Similarly, participants appeared to make almost no errors in estimating the Perceived Sex Ratio in an all-male ensemble, although they were more likely to underestimate the number of men ($M_{12\text{men}}$=-0.32, SD=0.77, $Median$=0.00). There was also a small but significant main effect of perceiver sex, such that female perceivers significantly overestimated the ratio of men to women compared to male perceivers, B=0.26, SE=0.07, t=3.54, p=.000673, R^2_{MVP}=0.01. Across conditions, male perceivers tend to underestimate (M=-0.02, SD=0.99, $Median$=0.00) while female perceivers tend to overestimate (M=0.24, SD=0.84, $Median$=0.00) the actual number of men in each ensemble.

As the ratio of men to women in the ensemble increased, the number of Absolute Errors made by perceivers decreased, B=-0.09, SE=0.02, t=-4.88, p<.0001, R^2_{MVP}=0.02. This effect was qualified, however, by a significant interaction, B=-0.14, SE=0.04, t=-3.44, p=.000935, R^2_{MVP}=0.03 (see Figure 10, panel C). Simple effects analyses revealed that the absolute number of errors made by male perceivers did not vary as a result of the number of men in the ensemble, B=0.02, SE=0.06, t=0.32, p=.75, *n.s.* However, female perceivers made significantly fewer overall errors in Perceived Sex Ratio as the ratio of men to women in the ensemble increased, B=-0.12, SE=0.03, t=-4.51, p<.0001, R^2_{MVP}=0.04. Additionally, there was no difference in the amount of errors made by male or female perceivers when the ensemble was all-female, 1 man: 3

women, or 2 men: 2 women (all *p*s≥.13716). However, women made fewer estimation errors than men when the ensemble was majority- (3 men: 1 woman, *B*=-0.30, *SE*=0.08, *t*=-3.59, *p*=.000578, R^2_{MVP}=0.04) or all-male (4 men: 0 women, *B*=-0.26, *SE*=0.14, *t*=-1.93, *p*=.057883, R^2_{MVP}=0.03).

Social Attitudes and Affordances. The ICCs for Belonging and Perceived Sexist Norms were 0.27 and 0.21, respectively; this suggests that the majority of variance in participants' responses to social attitudes measures was within subject. There was a main effect of condition on Perceived Sexist Norms, *B*=0.32, *SE*=0.03, *t*=10.00, *p*<.001, R^2_{MVP}=0.20; as the number of men in the ensemble increased, perceivers believed the group significantly more likely to endorse sexist norms (see Figure 11, panel A).

As the number of men in the ensemble increased, perceivers also reported significantly less Belonging, *B*=-0.48, *SE*=0.07, *t*=-6.73, *p*<.001, R^2_{MVP}=0.11. This main effect was qualified by a significant interaction of perceiver sex and condition, *B*=-0.74, *SE*=0.11, *t*=-6.33, *p*<.001, R^2_{MVP}=0.15 (see Figure 11, panel C). Whereas female perceivers reported significantly less Belonging as the number of men in the ensemble increased (*B*=-0.64, *SE*=0.07, *t*=-9.03, *p*<.001, R^2_{MVP}=0.19), male perceivers felt similar levels of Belonging regardless of the number of same-sex others in the ensemble (*B*=0.10, *SE*=0.10, *t*=0.98, *p*=.339, *n.s*)[5]. Simple effect analyses also revealed that female perceivers felt significantly more Belonging than male perceivers when the ensemble was all-women (*B*=1.29, *SE*=0.39, *t*=3.28, *p*=.00155, R^2_{MVP}=0.09) and marginally more Belonging when the ensemble was 3 women and 1 man (*B*=0.58, *SE*=0.32, *t*=1.81,

[5] When controlling for mean attractiveness of each ensemble, the simple effects analysis revealed that men felt significantly more Belonging as the ratio of men: women increased, *B*=0.24, *SE*=0.10, *t*=2.37, *p*=.0285, R^2_{MVP}=0.03. Additionally, mean attractiveness was a significant covariate, *B*=0.95, *SE*=0.14, *t*=6.59, *p*<.001; male perceivers felt like they belonged significantly more in more attractive ensembles. However, the significant interaction and simple effects findings for female perceivers remained unchanged when including mean ensemble attractiveness as a covariate.

p=.0736, R^2_{MVP}=0.02). Similarly, female perceivers felt significantly less Belonging when the ensemble was majority- (3 men: 1 woman, B=-0.90, SE=0.31, t=-2.88, p=.00517, R^2_{MVP}=0.06) or all-male (4 men: 0 women, B=-1.74, SE=0.34, t=-5.06, p<.001, R^2_{MVP}=0.22). Male and female perceivers reported similar levels of Belonging when the ensemble's sex ratio was equal, B=-0.05, SE=0.31, t=-0.17, p=.866, *n.s.*

Discussion

Despite the smaller number of individuals in the ensemble in Study 4a compared to Studies 1a and 1b, the initial findings replicated. They provide compounding evidence that individuals extract summary social information as well as actuarial statistics from a group of people. I showed that perceivers are accurate in determining the sex ratio of an ensemble and that changes in group sex ratio maps onto significant changes in how perceivers envision the average group member's face. Furthermore, the number of same-sex others in the ensemble significantly affected perceivers' feelings of fit and belonging; in particular, women reported greater belonging in majority-female groups while men seemed to feel like they fit in almost any group. Finally, Study 4a again demonstrated that perceivers view majority-male ensembles as hostile spaces for women, with the group seemingly more likely to endorse discriminatory, sexist norms.

It is perhaps unsurprising that my hypotheses held and our findings replicated in Study 4a. The methodology remained the same, although the ensembles were smaller than in Studies 1a, 1b, and 2. Additionally, prior research has similarly shown perceivers can extract actuarial summaries from four-person sets (e.g., mean facial emotion; Haberman & Whitney, 2007). However, a decrease in the number of composite images per set does not automatically render ensemble coding perceptually easier (Ariely, 2001; Chong & Treisman, 2005b; Haberman & Whitney, 2007). More importantly, beyond its merit as a replication study, Study 4a provides

novel evidence of ensemble coding using real-world groups. To date, prior ensemble coding studies with human targets have relied on artificial group stimuli. These studies minimized extraneous noise and helped establish ensemble coding as a phenomenon. Past researchers compiled ensemble stimuli from carefully controlled individual sources: morphed images of real faces (Haberman & Whitney, 2007, 2010; Leib et al., 2014); point-light displays outlining the human figure (Sweeny et al., 2013); drawn images of human facial characteristics (Sweeny & Whitney, 2014); or, pictures of individual faces extracted from a public database (Alt et al., 2017; Thornton et al., 2014). Here, for the first time, I demonstrate that the process for extracting actuarial and social summaries from random groups holds when examining naturally occurring ensembles. In Study 4a, I sought out groups of people that were actively judged and evaluated together, namely on a conference panel in front of an audience of their peers. Thus, these findings extend prior work and demonstrate higher external validity; I can more confidently say, as a consequence of Study 4a, that ensemble coding is not a phenomenon unique to minimal or laboratory-contrived groups.

Although hypothesized, this was not a given expectation. Unlike stimuli in Studies 1a, 1b, 2, and 3, target images in Study 4a revealed only the facial features of each person. Each individual face in an ensemble had been cropped to remove as much hair, neck, and body cues as possible. Furthermore, the target images were extracted from headshots or larger pictures readily available online; they did not have the pre-processing and emotional control available for images taken from a specific dataset. Studies 1 through 3, for example, relied on groups with target images who were wearing identical clothing, making neutral expressions, and had been solicited specifically knowing their images would be used in future experiments (Ma et al., 2015). Despite attempts at standardizing stimuli, the target creation process in Study 4a invariably introduced

greater noise into the study than had I relied on morphed or database target images. Thus, replicating prior findings in Study 4a takes on larger significance when we consider that using real-world groups introduced more noise into the study and made it harder for the effects of interest to emerge.

Of note, the observed effects appeared less pronounced in Study 4a than in Studies 1a, 1b, and 2. While the proportional representation of the number of men to women in the ensembles remained the same across studies, the smaller overall ensemble size in Study 4a may have contributed to this diminished effect. Prior research has shown accuracy in actuarial summaries itself appears impervious to ensemble size (Haberman & Whitney, 2007). However, it may be that ensemble coding of social attitudes and affordances differently weights the number as well as the proportion of same-sex others in the group. For example, feelings of belonging may depend on the overall size of a group. Seeing the same proportion (e.g., 25%) of same-sex others around you may feel different if the group is four versus 100 people big. Future research may wish to consider the interacting effects of overall ensemble size and sex ratio. For now, it remains a possibility that the diminished effect size between Study 4a and Studies 1a-2 could be due to differences in the number of people in each ensemble.

The effects seen in Study 4a could also potentially be explained by a numerosity or counting hypothesis; instead of extracting actuarial summary statistics, it could be that participants were able to quickly count how many men and women were in each ensemble. Piazza et al. (2002), for example, contend that perceivers can rapidly subitize and accurately enumerate sets of four or fewer objects with little cognitive energy. Although each ensemble was presented onscreen for 500ms, theoretically a perceiver wishing to attend to each individual face in Study 4a would have had 125ms per face to do so. By contrast, subitizing the 12-person group

from Study 1b would limit individual attention time to just under 42ms per target. Furthermore, each ensemble in Study 4a was presented in the same location centered onscreen for each trial. Perceivers may have quickly learned where to look to see each target face; this strategy would have reduced the task's cognitive load and made it easier for participants to accurately extract descriptive summary statistics. Thus, Study 4a's effects could have resulted from numerosity and repetition rather than ensemble coding per se. I sought to address this potential limitation and alternative hypothesis in Study 4b.

Study 4b

The purpose of Study 4b was to rule out stimuli habituation and subitizing as alternative explanations for the effects found in Study 4a. Whereas in Study 4a participants viewed each ensemble in a centered 1x4 array, in Study 4b I randomized the placement of target images onscreen. Prior research has demonstrated that perceivers habituate to visually similar images and repetitive trials; this leads to decreased amygdala activity and a dampened response (Breiter et al., 1996; Carretié, Hinojosa, & Mercado, 2003; Romero & Polich, 1996). Practicing repetitive tasks, including facial recognition, decreases response latency (Forbach, Stanners, & Hochhaus, 1974; Itier & Taylor, 2004; Schweinberger & Neumann, 2016) and facilitates future similar task completion (Monsell, 2003; Schweinberger & Neumann, 2016). Thus, it may be the case that presenting only four stimuli in the same location onscreen each trial as I did in Study 4a could have constituted a relatively easy task for participants. Study 4b sought to increase task difficulty by randomizing the placement of individual stimuli within each ensemble, thereby undermining alternative explanations for these effects.

Despite the increase in cognitive load that could arise from Study 4b's methodology, I predicted that perceivers would still accurately extract descriptive and subjective summary

statistics about the ensemble. In particular, I hypothesized that as the number of actual men in the ensemble increased, perceivers would see the sex ratio as becoming more male-dominated, the average group member as more masculine, and the group overall as more supportive of sexist norms. I also expected a significant interaction of perceiver sex and actual sex ratio on feelings of belonging, such that as the number of men in the ensemble increased, male perceivers would report more belonging while female perceivers would report decreased belonging. Because the layout of the 4-person ensemble was to be shuffled across a 12-cell grid, I anticipated the effect sizes would be smaller than in Study 4a. Overall, I expected Study 4b to further validate ensemble coding's emergence as a robust phenomenon and show the same patterns as prior studies regardless of individual placement within each ensemble or the task's cognitive load.

Methods

Design. Study 4b followed a 2 (perceiver sex: male v. female) x 5 (sex ratio of male/female ensemble members: 0/4 v. 1/3 v. 2/2 v. 3/1 v. 4/0) quasi-experimental, mixed model design. It was a methodological replication of Study 4a, with the exception that the stimuli were presented in constrained but randomized order. Participants in Study 4b saw the 4 target faces from each panel randomly arranged in a 3x4 array; besides the 4 target faces, participants also saw 8 blank spaces on screen.

Participants. Eighty-five undergraduate students (76% women) at a large, West coast public university completed the laboratory experiment in exchange for course credit. Individuals who participated in Study 4a were not eligible to participate in Study 4b.

Materials and Procedure. Study 4b drew on the same SPSP panel stimuli created and used in Study 4a. The number of ensembles in each condition followed the same distribution and

ratio break-down as in Study 4a (e.g., 11 all-women ensembles; 16 1 man: 3 women ensembles; 25 2 men: 2 women ensembles; 22 3 men: 1 woman ensembles; and, 6 all-men ensembles).

As in Study 4a, participants gave informed consent and began by completing several test trials using popular 1970s cartoon characters. Instead of seeing each ensemble in a 1x4 array centered onscreen, participants viewed the four individual target images shuffled and spread out randomly in a 3x4 array. The placement of each target image within the array was determined in real time by an adapted Python script; since there are only four targets per ensemble, eight of the twelve spaces appeared empty. Which eight cells were empty varied randomly across trials and across participants (see Figure 8, panel F for example stimulus). Following a 500ms fixation cross, participants saw a given ensemble 500ms, then a 500ms delay, and finally the block-specific question(s). As in Study 4a, participants viewed each of the 80 ensembles once per block and responded to the same four dependent variable blocks: Perceived Sex Ratio; Perceived Facial Masculinity; Perceived Sexist Norms (scale reliability R_C=0.72); and, Belonging (scale reliability R_C=0.97).

Statistical Analysis. I ran a series of step-wise multi-level random effects regression models with random slopes and random intercepts to test how sex ratio impacted actuarial and social judgments about an ensemble, given the random presentation of target faces. I again drew on the covariate ratings of attractiveness and target gender categorization from the pre-test when conducting analyses. Since participants saw the exact same 4-people per a given ensemble, I included a cross-classification term for the target group image in my models. As in Study 4a, I separately regressed each of the four criterion variables onto perceiver sex, actual sex ratio, and their interaction; additionally, I tested the covarying effect of the ensemble's average attractiveness as both a fixed and random effects term (nested within the target group image

term). As in Study 4a, mean ensemble attractiveness only significantly changed the baseline models' effects where noted. Finally, I also examined the effect of perceiver sex and ensemble sex ratio, as well as their interaction, on the Error and Absolute Error in Perceived Sex Ratio (calculated for this sample as described in Study 4a).

Results

Accuracy in encoding actuarial summaries. ICC analyses for Perceived Sex Ratio and Perceived Facial Masculinity revealed that 5% and 36% of the variance in participants' responses was within subject, suggesting most participants responded similarly to each ensemble. As the number ratio of men to women in the ensemble increased, the Perceived Sex Ratio significantly increased, B=0.44, SE=0.02, t=20.64, p<.001, R^2_{MVP}=0.24, and the average target face was rated as more masculine, B=0.44, SE=0.11, t=4.01, p=.000103, R^2_{MVP}=0.02 (see Figure 9, panels B and D). Overall, compared to their male peers, women perceived significantly higher ratios of men to women in each ensemble, although this effect appeared relatively small, B=0.17, SE=0.07, t=2.56, p=.0122, R^2_{MVP}<0.01.

The ICC for Error in Perceived Sex Ratio equaled 0.04; almost all of variance in participants' responses occurred between subjects. Participants overestimated the number of men in the ensemble significantly less as the ratio of men to women increased, B=-0.56, SE=0.02, t=-31.10, p<.0001, R^2_{MVP}=0.29 (see Figure 10, panel B). When the ensemble was all-female, participants reported seeing about 1 more man in the ensemble than was actually present ($M_{0\text{men}}$=1.44, SD=0.99, $Median$=1.00). When the ensemble was all-male, however, participants reported seeing about 1 fewer man than was actually present ($M_{12\text{men}}$=-0.78, SD=0.88, $Median$=-1.00). Female participants also overperceived the number of men in the ensemble compared to

male participants, although this effect was relatively small, B=0.17, SE=0.07, t=2.56, p=.01223, R^2_{MVP}<0.01.

The null model for Absolute Error had an ICC equal to 0.03. As the ratio of men to women in the ensemble increased, the absolute number of errors in Perceived Sex Ratio decreased, B=-0.18, SE=0.02, t=-11.21, p<.0001, R^2_{MVP}=0.06. Perceiver sex significantly moderated this effect, however, B=-0.10, SE=0.04, t=-2.86, p=.0054, R^2_{MVP}=0.06 (see Figure 10, panel D). Simple effects analyses revealed that both female (B=-0.20, SE=0.02, t=-8.75, p<.0001, R^2_{MVP}=0.09) and male (B=-0.10, SE=0.03, t=-3.19, p=.00302, R^2_{MVP}=0.02) perceivers made increasingly fewer errors in estimating the Perceived Sex Ratio as the ratio of men to women increased. The rate of change and effect size were different depending on perceiver sex. There was no significant difference in Absolute Error between male and female perceivers when the ensemble was comprised of all-women, 1 man: 3 women, 2 men: 2 women, or 3 men: 1 woman (all ps≥.118). However, for all-male ensembles, women made fewer overall errors in estimating the Perceived Sex Ratio than men, B=-0.48, SE=0.12, t=-3.85, p=.00023, R^2_{MVP}=0.06.

Social Attitudes and Affordances. The ICC for Perceived Sexist Norms and Belonging calculated from the null models equaled 0.29 and 0.42, respectively. As the ratio of men to women in the ensemble increased, participants reported the group as significantly more likely to endorse sexist norms, B=0.21, SE=0.02, t=9.52, p<.0001, R^2_{MVP}=0.10. This main effect of condition varied based on participant sex, interaction B=0.12, SE=0.05, t=2.39, p=.01927, R^2_{MVP}=0.10 (see Figure 11, panel B). Simple effects analyses revealed that both men and women perceivers saw the group as significantly more likely to endorse sexist norms. However, the rate of change in Perceived Sexist Norms was significantly steeper for female compared to male perceivers (men B=0.12, SE=0.02, t=5.83, p<.0001, R^2_{MVP}<0.04; women B=0.24, SE=0.03,

t=8.78, p<.0001, R^2_{MVP}=0.12). Furthermore, the effect of sex ratio on Perceived Sexist Norms was about three times as large for female compared to male perceivers. Male and female perceivers reported similar levels of Perceived Sexist Norms when there were 1 or more men in the ensemble (all ps≥.14); when the ensemble was all-women, female perceivers reported significantly less Perceived Sexist Norms than their male peers, B=-0.32, SE=0.16, t=-2.00, p=.0493, R^2_{MVP}=.03.

As the ratio of men to women in the ensemble increased, both male and female participants reported significantly less Belonging, B=-0.31, SE=0.05, t=-6.73, p<.0001, R^2_{MVP}=0.06. A significant interaction qualified this main effect, B=-0.55, SE=0.08, t=-6.80, p<.0001, R^2_{MVP}=0.11 (see Figure 11, panel D). Female perceivers felt significantly less Belonging as the ratio of men to women increased, B=-0.44, SE=0.05, t=-9.38, p<.0001, R^2_{MVP}=0.11. In contrast, male perceivers felt marginally more Belonging as the ratio of men to women increased, B=0.11, SE=0.06, t=1.89, p=.0726, R^2_{MVP}=0.01[6]. There was no difference in reported Belonging based on perceiver sex when viewing an ensemble comprised of 0 men: 4 women, 1 man: 3 women, or 2 men: 2 women (all ps≥.155). However, women reported significantly less Belonging than men when the ensemble was majority- (3 men: 1 woman; B=-1.09, SE=0.35, t=-3.11, p=.00255, R^2_{MVP}=0.10) or all-men (4 men: 0 women; B=-1.62, SE=0.37, t=-4.40, p<.0001, R^2_{MVP}=0.22).

Discussion

Study 4b improved upon a limitation from Study 4a by randomizing the order and placement of individual targets within an ensemble. Instead of presenting the images in a 1x4

[6] The simple effect for male perceivers became significant when controlling for the ensemble's mean attractiveness rating, B=0.s11, SE=0.12, t=2.10, p=.0471, R^2_{MVP}=0.01.

array reflecting the actual panel line-up from the 2015 SPSP Convention, I wrote a Python script that randomly placed each of the four individual targets in a 3x4 grid. I once again replicated findings from Studies 1a and 1b. Participants detected shifts in sex ratio in line with each condition, choosing visual representations of the group that had a greater ratio of men to women and more masculine facial features as the number of men in the ensemble increased. Perceivers' feelings of belonging and anticipation of sexist norms also relied on the ratio of men to women in each ensemble. Replicating most of Study 4a's findings, Study 4b demonstrated ensemble coding's occurrence among real-world groups.

Despite similar overall trends, Study 4b did differ in small but significant ways from prior studies. Aside from Belonging, which had similar effect sizes in both studies (Study 4a R^2_{MVP}=0.15, Study 4b R^2_{MVP}=0.11), the effect sizes from Study 4b were almost half of the effect size of the same measures in Study 4a. This may be attributed to the randomization of stimuli in Study 4b; the study's purpose was an attempted replication of Study 4a despite a more challenging methodological presentation of stimuli. The reduced effect sizes for Perceived Sex Ratio, Perceived Facial Masculinity, and Perceived Sexist Norms suggests I successfully increased the task difficulty. Presenting images in a 3x4 array dampened, but did not completely undermine, perceivers' ability to encode summary actuarial and social information about each ensemble. This lends further credibility to the robustness of ensemble coding and again illustrates the presence of this process among real-world groups.

Less easy to understand, however, is the emergence of an interaction effect on Perceived Sexist Norms in Study 4b. While the main effect of actual sex ratio on Perceived Sexist Norms replicated across both studies, its moderation by perceiver gender was absent in Study 4a and the preceding studies. Closer inspection revealed that this interaction effect primarily resulted from

gender differences in the rate of change related to noticing potential discriminatory norms; the direction of the effect remained the same across perceiver gender. Both men and women agreed that as the number of men in an ensemble increased, the group would endorse more sexist norms. Women recognized an all-women ensemble as less likely to endorse sexist norms compared to men. Further, the effect of sex ratio on perceived sexist norms was three times as large for female perceivers compared to male perceivers. Thus, perceptions of sex ratios appear to matter more for women's social judgments of gender-related group norms than they do for men.

It is beyond the scope of the current paper to unequivocally determine why a significant interaction effect of perceiver sex and condition on Perceived Sexist Norms emerged in Study 4b but not prior studies. However, one possible explanation arises from the methodological change in stimuli presentation between the two studies. Participants in Study 4b had to scan a 3x4 array to extract the gist of the ensemble, whereas participants in Study 4a only needed to look at the same place onscreen to see each 1x4 array. It may be that this added step of having to find and process the four target faces contributed to the steeper change in Perceived Sexist Norms for women compared to men.

An alternative explanation could be that items on the discriminatory norms scale probed the potential mistreatment of *women* in the group (e.g., whether the group would treat women as "bad at math"). Research from organizational psychology suggests that groups form norms only around important values and behaviors that they want members to obey (Feldman, 1984; Hackman, 1992; Shaw, 1981); thus asking about these norms may have primed female perceivers to pay more attention to relevant visual cues. Prior research in social vision has found people subconsciously avert or seek out target images in their visual field that align with self-relevant goals (see Isaacowitz, 2006 for a detailed review). Thus, if female perceivers were

concerned with avoiding discriminatory groups, they may have actively sought out and paid greater attention to visual information signaling hostile norms.

In contrast, the items presented in the Perceived Sexist Norms scale do not suggest negative behavioral implications for men. The visual cues potentially sought out by female perceivers may have proven less socially relevant for male perceivers. Study 4b also required more work than Study 4a to process the same actuarial information. These factors together could have reduced male perceivers' motivation to work towards extracting social information about the ensemble's sex ratio and led to a significantly smaller effect size. Future research may wish to test this hypothesis in numerous ways. For example, experimenters could manipulate whether ensembles are presented randomly or in a 1x4 array within the same study, track the eye gaze of female versus male perceivers, and test the potential mediating effect of norms' self-relevance.

Study 5

Having demonstrated perceiver's spontaneous, accurate ability to ensemble code real-world stimuli in Studies 4a and 4b, I next probed how summary statistics bias perceivers judgments about the merit and impact of the group's work. Each of the 80 panels shown in Studies 4a and 4b gave talks at SPSP. I was curious whether each ensemble was viewed as equally academically rigorously or whether the group's sex ratio influenced how serious perceivers viewed their research. Consistent with sex stereotypes showing men as competent but not warm (Wood & Eagly, 2010), I predicted panels with more male members would be rated as higher in intellectual merit and less likely to make broader impacts. Conversely, I expected participants to rate female-dominated panels as lower in intellectual merit but more likely to share their knowledge with others (e.g., socially warm).

Methods

Design. Study 5 followed a mixed model, quasi-experimental design. It used the same sex ratio conditions and stimuli as Studies 4a and 4b: 0/4 men, 1/3 men, 2/2 men, 3/1 men, and 4/0 men per ensemble. I also examined the potential effect of perceiver sex (male v. female) on ensemble trait ratings, although I did not expect it to moderate the effect of sex ratio on either dependent variable.

Participants. Eighty-seven students at a large, public West Coast university completed the study in exchange for course credit; individuals who participated in Study 4a or 4b were ineligible to complete Study 5. Data from one participant who fell ill during the study and from one participant who reported their gender as outside the gender binary were excluded from analyses. Eight participants reported recognizing at least one of the faces in a given ensemble; since prior knowledge about the target may bias impressions about their work, I excluded data from these participants as well. The final sample was comprised of seventy-seven participants (73% women).

Materials and Procedure. Participants provided informed consent and were seated at one of three computers pre-loaded with a study-customized Python script. Participants completed the same practice trials as in Study 4a before moving onto the main experiment. The same 80 ensembles used in Studies 4a and 4b were reused in Study 5 (see Figure 8). Each ensemble was presented on screen for 500ms in a 1x4 array after a 500ms fixation cross; as before, a 500ms delay preceded the presentation of the criterion questions. Maximizing external validity as much as possible, individuals in the 1x4 array were laid out left to right in the order in which they spoke during their actual panel presentation (as indicated by the 2015 SPSP program). The 80 ensembles were presented in randomized order and shown once per block. Participants completed three counterbalanced blocks (and thus 240 trials) in total.

The first block assessed the ensemble's Perceived Masculinity. After each ensemble, participants were asked to indicate on a 7-point Likert Scale, "How masculine or feminine is the average group member of the group you just saw?" (1=*Extremely feminine* to 7=*Extremely masculine*).

In the second block, participants responded to 7 items assessing targets' seeming potential to make a broader impact on society. Items were based on the Broader Impacts criteria used by the National Science Foundation (NSF) in assessing fellowship and funding opportunities across multiple fields of science (The National Science Foundation, 2016). Each item asked participants to indicate on a 7-point Likert scale (1=*Extremely unlikely* to 7=*Extremely likely)* how likely members of the ensemble were to engage in certain activities. Sample items included: "How likely is work by members of this group to increase public engagement with science and technology, including scientific literacy?"; "How likely is work by members of this group to encourage underrepresented individuals like racial minorities, women, and persons with disabilities to participate in science?"; "How likely is work by members of this group to improve the well-being of individuals in society?" For a full list of the seven items, see Appendix B. An analysis of scale reliability following the procedure outline by Cranford et al. (2006) for multi-level repeated measures suggested that the seven items hung together reasonably well (R_C=0.84). Thus, I averaged together participants' responses for the seven items to arrive at an overall Perceived Broader Impacts score for each ensemble.

The third block assessed the intellectual merit attributed to each ensemble by perceivers. I again compiled the list of Intellectual Merit criteria put forth by the NSF to create seven related-items (see The National Science Foundation, 2016). After exposure to an ensemble group, participants were asked to indicate on a 7-point Likert scale (1=*Extremely unlikely* to

7=*Extremely likely*) group members' likelihood to engage in each behavior. Sample items included: "How likely are members of this group to advance knowledge and understanding *within their own* field/industry?"; "How likely are members of this group to advance knowledge and understanding *across different* fields/industries?"; "How likely are members of this group to transform the frontiers of knowledge?"; "How well qualified do members of this team seem to present their work to others?" See Appendix B for the full item list. The scale showed high internal reliability (R_C=0.91), so participants' responses to the seven items were averaged together to create a mean Perceived Intellectual Merit score for each ensemble.

After completing all three blocks, participants answered basic demographic questions and indicated whether they recognized any of the individual target faces in the study.

Statistical Analysis. Data from Study 5 was analyzed using a series of multilevel models with random slopes and random intercepts. Because observations were dually nested within perceivers and within stimuli, I specified cross-classification in the models. In a step-wise fashion, I regressed Perceived Scientific Rigor, Perceived Broader Impacts, and Perceived Intellectual Merit separately onto perceiver sex, actual sex ratio, and their interaction. I also tested for the covarying effect of ensemble attractiveness on the criterion. I report only the significant effects of the covariate below.

Results

Perceived masculinity. Most of the variance in participants' responses was across subjects (ICC=0.06), highlighting the reduced role of participant-level factors in determining Perceived Masculinity. As the ratio of men to women in the ensemble increased, perceivers reported the average group member as more masculine, B=0.73, SE=0.05, t=13.39, p<.0001, R^2_{MVP}=0.30 (see Figure 12). Female perceivers reported the average group member as more

masculine than male perceivers, B=0.36, SE=0.10, t=3.48, p<.0001, R^2_{MVP}=0.01; this finding is consistent with a conceptual replication of Perceived Sex Ratio from Studies 4a and 4b.

Perceived Broader Impacts. The ICC drawn from the null model predicting Perceived Broader Impacts equaled 0.40. There was a main effect of sex ratio, such that as the ratio of men to women in the ensemble increased, Perceived Broader Impacts decreased, B=-0.08, SE=0.25, t=-3.11, p=.00232, R^2_{MVP}=0.01. Perceiver sex marginally moderated the effect of ensemble sex ratio on perceptions of the group's likelihood to create work with broader impacts, B=-0.09, SE=0.04, t=-1.97, p=.0521, R^2_{MVP}=0.02 (see Figure 13, panel A). Men reported similar levels of Perceived Broader Impacts across sex ratio conditions, B=-0.01, SE=0.03, t=-0.51, p=.61. Women, however, believed groups with more men and fewer women were less likely to produce work that would impact their communities, B=-0.10, SE=0.03, t=-3.32, p=.00125, R^2_{MVP}=0.02.

Perceived Intellectual Merit. Calculating the ICC for Perceived Intellectual Merit revealed that more than two-thirds of the variance in participants' responses was due to differences between participants (ICC=0.32). Contrary to hypotheses, participants rated groups as having similar levels of intellectual merit, regardless of the ratio of men to women in the ensemble, B=-.00, SE=0.02, t=-0.15, p=.882, *n.s.* Overall, women rated the ensembles as having significantly more Perceived Intellectual Merit than men, B=0.32, SE=0.16, t=2.06, p=.0426, R^2_{MVP}=0.02 (see Figure 13, panel B).

Discussion

Findings from Study 5 provide further evidence that sex ratios within groups impact downstream perceivers' assumptions about target individuals and their work. Namely, majority female groups are perceived as significantly more likely to contribute to scientific literacy and to share their work with a broader audience. This finding aligns with gender stereotypes portraying

women as talkative, warm, aware of others' feelings, caring, patient, and motherly (Broverman et al., 1972; Ghavami & Peplau, 2012; Prentice & Carranza, 2002). Furthermore, these traits are consistent with occupations traditionally deemed "more feminine" (e.g., teaching or nursing; Miller & Budd, 1999). Study 5 highlights how groups rated as significantly less feminine by perceivers are also those significantly more likely to be seen as committed to increasing public engagement and mentoring underrepresented populations. From a half second exposure to an ensemble, perceivers extract socially relevant information and determine the group's usefulness in disseminating important knowledge.

Contrary to my hypotheses, however, I did not find any significant differences in perceptions of intellectual merit as a result of the ensemble's sex ratio. Given the prevalence of gender stereotypes suggesting women are incompetent and bad at STEM (Ghavami & Peplau, 2012; Miller & Budd, 1999; Wood & Eagly, 2010), I expected majority- and all-female ensembles to be rated as lower in perceived intellectual merit. Several competing alternative explanations may exist for why this pattern did not emerge. First, it may be that the distribution of perceiver gender in the sample biased the findings. Women comprised almost 75% of participants in Study 5. Research on in-group bias and favoritism suggests individuals have a tendency to overemphasize the positive attributes of their own group (Brewer, 1979). By this logic, female perceivers may have rated groups with more women as higher in Intellectual Merit. However, this was not the pattern that arose; instead, women rated all groups, regardless of sex ratio, as higher in Intellectual Merit than male perceivers.

A second potential hypothesis stems from social desirability theory; study participants may have responded to questions in a way that is socially desirable or renders them in a positive light. As a result, their responses may not reflect their true opinions or first impressions (Maher, 1992;

Reynolds, 1982). Given the public call for increased representation of women and underrepresented minorities in STEM fields (Beede et al., 2011), respondents may have been aware of negative gender stereotypes about women being intellectually inferior at STEM-related work. I did not explicitly state the ensembles came from a social science domain. However, participants may have surmised the research question of interest from the Perceived Intellectual Merit items themselves (e.g., "How rigorous is the science done by members of this group?"; see Appendix B for other items). In keeping with this explanation, perceivers may have hesitated or resisted reporting majority- or all-female groups as conducting less rigorous work. They may not have wanted to appear sexist, doctoring their responses at the expense of their true beliefs.

Alternatively, it may be that the Study 5 sample truly does not endorse gender-based stereotypes about women's intellectual merit. Participants came from a liberal-minded university in a county where 74.26% of constituents voted for a Democrat or Green Party candidate in the 2016 presidential election (Los Angeles County Registrar, 2016). It may be that if I had recruited a more nationally representative sample, Study 5 would have shown different results. Additionally, participants were all college-aged; most have had limited experience in the working world and may be less suited to judge the merits of other individuals' work. Thus, future research may want to consider recruiting a politically diverse sample of older adults.

Results from Study 5 imply that audience members at conferences make assumptions about female-dominated panels upon entering the room; even before seeing the presenters' data, perceivers may judge their work as more impactful and likely to encourage underrepresented minorities to pursue STEM fields. Study 5 has important potential implications across domains. For conference attendees and reviewers alike, greater attention and consideration for how the constituents of a panel affect its initial public reception could help in reducing a priori cognitive

bias. Results from Study 5 indicate that the overall diversity of panel presenters and the distribution of sex ratios in real-world scientific contexts matters.

General Discussion

Taken together, the results from Studies 4 and 5 demonstrate perceivers' ability to ensemble code real-world groups of people and demonstrate ensemble coding's external validity. The effects of Studies 1a, 1b, and 2 replicated in Studies 4a and 4b despite exposure to noisier targets. While I took some precautions to standardize the photos of SPSP panelists, on the whole, the images were taken at the target's discretion in the setting of his or her choice. The facial expressions were allowed vary naturally and more closely resemble the groups seen by perceivers in their daily lives. Study Set Two also shows that perceivers form downstream judgments about a group's ability to contribute to the scientific community and how much they want to interact with or join that group. Study 5 highlights the role of gender stereotypes in driving some of perceivers' behavioral intentions and attitudes towards ensembles, whether the perceivers realize it explicitly or not.

Future extensions of this work may consider ensemble coding's effectiveness in other, non-visual mediums. In Studies 4 and 5, participants saw each of the panels rapidly presented on screen. Most conference attendees, however, may glance through the symposium schedule and the list of presenters before deciding which sessions to attend. We know that women in science still face greater hiring discrimination and are judged more negatively than men with an identical resume (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Steinpreis, Anders, & Ritzke, 1999). Thus far, research has not considered whether sexist name-based differential treatment extends to groups of male and female scientists. It may be that perceivers ensemble code a list of presenters' names and base their decision to attend that panel on these

instantaneous summary statistics. I would expect the findings from Study 5 replicate across contexts, with male-dominated panel listings generating more interest and seen as having fewer broader impacts than female-dominated panel listings.

In addition to the medium by which perceivers gather information about an ensemble, researchers may also wish to consider how ensemble coding affects other downstream judgments. From a logistical perspective, conference organizers must assign each panel to a conference room. The size of these rooms may vary, however. It would be interesting to test whether a panel's sex ratio influences the room and amount of physical space assigned to it. Serving as a proxy for anticipated importance or significance of research, room assignment has important consequences for the panelists and the dissemination of their work. Smaller rooms cannot accommodate as many audience members and may even encourage some attendees to leave the panel if it becomes too crowded. Another non-manipulated factor to consider would be each panel's average H-index. The H-index constitutes a relatively objective measure of an individual's publication influence in their field. I wonder whether perceivers would rate ensembles with higher mean H-indices as doing more rigorous work, as having more intellectual merit, and more likely to positively impact society. Future work could consider what relationship, if any, exists between summary statistics about a group and more objective measures of their success.

Finally, it is worth noting that Studies 4 and 5 draw on panel presenters from a single conference in one specific year. For greater generalizability, future research should examine a variety of conferences across fields and time periods. It would be interesting to compare sex ratios of panels in the social sciences, for example, with panels drawn from conferences in the physical or biological sciences. Can perceivers accurately categorize each of the different panels

better than chance and if so, what cues do they draw on to do so? Additionally, I wonder how perceivers' judgments about an ensemble would change if they were primed to believe either social scientists or life scientists comprised the panel. I expect implicit gender stereotypes would contribute to an over perception of men in physical science panels and under perception of women. Replication across conferences and across STEM fields would lend even greater external validity to the findings from Studies 4 and 5.

Scientists rarely research in isolation; instead we are surrounded by colleagues, lab mates, journal reviewers, and public audiences who share in our work. It is important to understand how who we work with shapes the ways in which others see us and evaluate our merit. This becomes particularly salient as we consider best practices for getting more women and other underrepresented minority individuals into STEM fields. It may be that to improve gender parity, we need to simultaneously increase the visibility of the women present in proportion to the number of men while also continuing to combat negative gender stereotypes. This dual approach could serve to change women's feelings of belonging and reduce the strength of implicit effects discouraging perceivers from attending "less rigorous" majority-female talks. By focusing on increasing the perceived as well as the actual sex ratio of male-dominated fields, we may begin to change how young female scientists view their place in STEM.

Paper Three: The Social Vision Toolkit

The Social Vision Toolkit: A high-level, open-source stimuli presentation framework for social scientists

Brianna M. Goodale and Kerri L. Johnson

University of California, Los Angeles

The Social Vision Toolkit: A high-level, open-source stimuli presentation framework for social scientists

The purpose of this paper is to provide a brief overview of the Social Vision Toolkit, a new, open-source Python framework developed with social scientists in mind. In particular, the Social Vision Toolkit allows for full customization and random stimuli presentation in an experimental setting. It comes pre-loaded with multiple features which, out of the box, enable users to present several different types of stimuli and automatically records participants' input. It aims to be widely accessible, with the same code working across Android, macOS, and Linux platforms. Currently, the Social Vision Toolkit is available free of charge and stored in an online Git repository that anyone with an account can access. It is my hope that the Social Vision Toolkit further contributes to open science and greater transparency in experimental methods.

History Behind the Social Vision Toolkit's Development

The idea behind the Social Vision Toolkit arose from a need to create real-time random presentation of stimuli in early 2015. The authors wanted to run a study wherein participants saw fifty target images, with each image comprised of a random 12-person group of individual faces. Creating each stimuli by hand seemed unnecessarily tedious and would introduce potential bias into the results; it could be that the randomly created 50 imagines seen by each participant were driving the effect, rather than the variable of interest (e.g., the group's sex ratio). Although I could control for this variance by including a cross-classification term during data analysis, I wondered whether software existed that would create the images in real time for participants. Essentially, I was interested in removing, rather than controlling for, the variance of these target images. I turned to Python to create an application that would, from a pool of set images, select and combine a new group of target faces for each trial and each new participant. The Social

Vision Toolkit, although in its nascence then, enabled me to run several studies testing the social perception of a group's sex ratio in a cleaner, more controlled design than previously possible (for full description of the study designs and findings, see Studies 1b and 2).

Advantages of the Social Vision Toolkit

Prior to the Social Vision Toolkit's creation, several other similar options were available for experimenters' use; I considered each of these before ultimately rejecting them in favor of creating my own. There are several commercial packages which require minimal programming. Users instead click through a graphical-user interface (GUI) to design their studies. Arguably easier to use for experimenters with no or limited programming experience, these commercial packages streamline the process and ease interpretation of results (Vihinen, 2015). However, I was not in favor of designing my studies using commercial software for several reasons. First, companies are typically unwilling or unable to share the application code for their proprietary software. The program acts as a "black box", asking users to input their stimuli and tasks without seeing the background code that generates the experiment (Jones, 2013; Vihinen, 2015). Thus, the experimenter must rely on the company's word that the program operates as it says it does; they cannot check the quality of the code themselves or share it more broadly. In contrast, open-source programs, by nature, are structured such that anyone can view, use, and test them (Krogh & Hippel, 2006; Peirce, 2009). Resultantly, they encourage greater collaboration between researchers. They also enable a feedback loop wherein any potential bugs in a program can be spotted by users and readily resolved.

In addition to its hidden code, commercial software can be cost-prohibitive for some researchers. Even with an academic or university discount, a single license for SuperLab or ePrime software costs $695 and $995 respectively (Cedrus Corporation, 2017; Psychology

Software Tools, 2018). High costs unnecessarily disadvantages young researchers, faculty with limited budgets, and scholars from smaller universities (Jones, 2013; Vihinen, 2015). The final deciding factor for me was that I could not easily customize the scripts in the proprietary program as much as I would have liked. I could have randomized part of the design, but not to the degree and specificity that was necessary to run our studies. Thus, I turned to open-source software as a potential alternative solution.

The Social Vision Toolkit is not the first stimuli presentation framework available free of charge. Other open-source packages enabling experiment creation in Python exist and have similar cross-platform usability. They too can run on Windows, macOS, or Linux operating systems (Peirce, 2009). Some alternative frameworks may also generate stimuli in real-time on a frame-by-frame basis (Peirce, 2007, 2009). However, the Social Vision Toolkit differs from prior open-source applications in several key ways which I believe improve its accessibility and ease of use. First, it leverages the open-source Kivy framework (https://kivy.org/#home) for all graphics and input-related processing. Kivy was initially released in 2011 as a platform to make building GUIs easier (Kivy, 2017). It wraps OpenGL, thereby sidestepping the need to understand the details of low-level graphics code. With over 20 widgets available in its libraries, Kivy renders it easier to customize stimuli presentation than code that uses OpenGL alone. Its design makes it more accessible and "higher-level" than open-source packages where the user must write each component from scratch. For example, Kivy comes with an Image Grid widget, which, given a number of rows and columns, will automatically create an onscreen array of images with the specified dimensions. This allows users to accomplish in one line of code what would otherwise require extensive manual implementation and many lines of custom code. Kivy developers have also written extensive documentation about the library; these write-ups include

step-by-step instructions for implementing widgets and tutorials to get new users mastering the basics quickly (see Kivy Organization, 2012). Based on Kivy, I believe the Social Vision Toolkit has a lower bar of entry for researchers than other Python-based open-source frameworks for stimuli presentation.

To further maximize its usefulness and accessibility, I have designed the Social Vision Toolkit with some basic features pre-loaded into the application. The framework code (which most beginner programmers need not edit or worry about) includes functions to automatically save each participant's data; experiment scripts created using the Social Vision Toolkit will automatically back up data throughout the experiment in case the computer crashes or the participant quits the program. Additionally, the experimenter does not have to write a "print" or show function to get a stimulus or dependent variable to appear onscreen. The Social Vision Toolkit again does this automatically based on framework code and class specifications. Prior open-source applications require users write code within each experiment file that saves the data and tells the program what to show. By including these as automatic features, I aimed to streamline the code within each experiment file and require users to master fewer concepts.

The Social Vision Toolkit also has increased functionality compared to other open-source applications. With the Social Vision Toolkit, researchers write their experiment files directly in Python. Some other applications, in contrast, require the user to build their experiments in XML which is then compiled into Python. XML constitutes a mark-up language, where code must follow the exact format and options as specified by the original designer. Writing the experiment code directly in Python, however, allows researchers to customize their experiments in ways we, as the original authors, may have not even considered. Users can define their own functions and create their own classes of objects. Currently, the Social Vision Toolkit comes pre-packaged

with text and image stimuli presentation classes, with participants needing to respond via the keyboard. The Kivy library, however, includes widgets for audio and video stimuli presentation, plus the ability to record different types of input from the user (Kivy Organization, 2012). For example, researchers could customize their experiment using the Camera widget to record participants' nonverbal behavior during the course of the study (e.g. to document emotional responses to linguistic bias as in Stout & Dasgupta, 2011). Experimenters could also create radio buttons onscreen via the CheckBox widget. Additional widgets of potential interest to researchers include drop-down lists, a progress bar, Scatter (for rotating and scaling shapes using multitouch systems like an Ipad), sliders, text input boxes, and an onscreen keyboard (Kivy Organization, 2012). Although the Social Vision Toolkit does not currently integrate these widgets out of the box, the Kivy framework allows users to develop and customize experiments without having to write the manual code for each widget themselves. Thus, the Social Vision Toolkit may also appeal to more advanced programmers who can use it as a base for new, potentially unforeseen extensions.

The Social Vision Toolkit's final advantage for researchers derives from its declarative nature. This means that the code in each experiment file does not have to be organized in executable order. Open-source alternatives to the Social Vision Toolkit rely on an imperative programming model; researchers must write each line of code in the order in which it will be executed by the application. Should a researcher want to add a new block or change the order of tasks in their experiment, it would prove easier using the Social Vision Toolkit than a different open-source Python framework. I accomplished this by having the experimenter define the order of stimuli, tasks, and events, rather than explicitly dictating execution order. An imperative programming model reads like a recipe; each step is prescribed in the necessary order. By

contrast, an experiment file using the Social Vision Toolkit describes *what* an experimenter is trying accomplish instead of *how* it must be done. The how has been accounted for in the base framework code created by the authors.

Contributing to the Open Science Movement

Beyond technical improvements on prior open-source Python frameworks, the Social Vision Toolkit may contribute to shifting norms in publication and replicability. Over the last decade, social scientists have been calling for greater transparency in research methodology. In a controversial article generating several counter responses, Nosek and Bar-Anan (2012) called for psychologists to move beyond "17th-century technologies" by incorporating "21st-century scientific communication practices [to] increase openness, transparency, and accessibility in the service of knowledge building…" (p. 217). They challenged researchers to consider how the Internet can increase collaboration and to publish their articles in open-access journals free to the public. Similar recognition of the need for greater transparency and increased replicability in experiments arose at the same time in other fields, including political science (Jones, 2013), life sciences (Osborne et al., 2014; Perez-Riverol et al., 2016), and bioinformatics (Perez-Riverol et al., 2016; Vihinen, 2015). In a recent Nature: Human Behavior column, Munafò et al. (2017) point to "large numbers of scientists working competitively in silos without combining their efforts" as a key element contributing to the publishing of error-prone results (p. 1). He and his co-authors encourage researchers to seek external review of their methodology, collaborate with individuals outside their home institution, and make their data public for others to use.

We have created the Social Vision Toolkit in an attempt to further this collaborative effort. In addition to making data or one's analyses public, creating an open-source Python framework enables researchers to share their experiment scripts. In particular, I have published

the Social Vision Toolkit on the open access repository called GitHub. A "social coding platform" (Perez-Riverol et al., 2016), GitHub wraps Git which is a Version Control System (VCS). Any user can see, copy, or comment on public repositories on GitHub with a free account. While not solving all issues in the current replicability crisis (Ram, 2013), making the Social Vision Toolkit available on GitHub promotes several of the tenants of open science. First, it creates a dialogue between coders and, by extension, researchers using the Social Vision Toolkit; multiple remote contributors can collaborate and contribute to the same code (e.g., professors or students at different institutions; Blischak, Davenport, & Wilson, 2016; Perez-Riverol et al., 2016; Ram, 2013). Furthermore, as a VCS, GitHub allows individuals to "fork" a copy of the original repository and make their own edits to it without affecting the original code (Perez-Riverol et al., 2016; Ram, 2013). If the users believe their changes would greatly improve the original source code, they can submit a pull request for the authors to consider merging their changes into the original file. In this manner, the code base can be built collaboratively over time. Thus, it is my hope that the Social Vision Toolkit will be forked by other experimenters who then adapt the code to their own needs and share their improvements to class or task types back on GitHub.

In addition to collaboration via sharing of the original code, providing the full framework on GitHub also affords greater transparency to research methodology. Any user with a GitHub account can upload and share publicly numerous data or text files, including code from statistical analyses. Beyond making data open access, researchers may also consider uploading the experiment script written with the help of Social Vision Toolkit on GitHub. To be clear, this is by no means a requirement of accessing the Social Vision Toolkit; users can also fork the Social Vision Toolkit and edit it locally, without uploading changes to GitHub. However, should they

wish to share their experiment scripts as part of a publication submission, for example, they can easily do so. The Git VCS will seamlessly sync their local file with their GitHub repository at any time. Replicating findings may become that much easier, as researchers can review and run the same experiment code across institutions.

Publishing the Social Vision Toolkit on GitHub also answers the call for greater methodological peer review. With the experiment scripts made public, other researchers can see how an experiment was built and what questions the authors asked. GitHub also has a built-in issues tracker; this followers of a particular repository to submit bug reports should they notice errors in the code (Perez-Riverol et al., 2016). The issues can then be addressed by the repository's original authors and any necessary fixes integrated in future experiment scripts.

In both its implementation and design, the Social Vision Toolkit aims to contribute to changing norms in the social sciences. The advancement of technology in the last few decades has made it possible for experimenters to run more advanced, technical studies in lab and provided the tools for sharing these experiments online. Home to the Social Vision Toolkit, GitHub further provides opportunities for researchers to collaborate, comment, and review the methodology behind each study.[7]

How to Use the Social Vision Toolkit

Having established the context surrounding its development and its merits in advancing open science, I turn now to the logistics underlying using the Social Vision Toolkit. I will walk through the files assembled in the Social Vision Toolkit and describe how to use the built-in functions to create an experiment from scratch. I will also review the different types of object

[7] Readers seeking more detailed information on how to use GitHub, including how to commit changes from a forked repository, should refer to Perez-Riverol et al. (2016) and Blischak et al., (2016).

classes relied on by experiment scripts. Finally, I will interpret line-by-line one of the three demo experiment files included in the Social Vision Toolkit.

Contents of the Repository

As presented on GitHub, the Social Vision Toolkit repository contains many different type of files. Several files are the framework code which users can see but need edit; this is where I have coded built-in functions, like automatically saving participant data or defining the type of classes. These files include: data_recorder.py; event.py; experiment.py; stimulus.kv; stimulus.py; subject.py; task.kv; task.py; ui.kv; ui.py; and, util.py. Repository visitors will also see a folder labeled "widgets" (see Figure 14), which houses the most common Kivy widgets drawn upon by the Social Vision Toolkit. Again, most users will have no need to change the contents of this folder. Advanced users may wish to add additional widgets from the Kivy library here, should they want to implement different types of stimuli (e.g., audio files). Finally, the repository also includes a license outlining its open-access policies and a README file with how to cite the Social Vision Toolkit.

In addition to framework files, the Social Vision Toolkit repository also includes files which the user must access and edit in order to run an experiment. The actual experiment scripts created by the user will be saved in the forked version of the Social Vision Toolkit, alongside the "widgets" folder and framework files. Three demo experiment scripts are included with the Social Vision Toolkit and can be seen in Figure 14: example_experiment.py; between_subjects_reverse_correlation_demo.py; and, within_subjects_ensemble_coding_demo.py. When the user creates their own experiment file, it should be saved here in the main folder.

Additionally, the user should note the “assets” folder in the repository. It should contain all the task and stimuli jpg images called upon in each experiment file. For the purposes of this paper, I have included in “assets” a sub-folder labeled “demo”; it contains image files for the dependent variables in each of the three demo experiment scripts, as well as folders for the stimulus base faces called by two of the demo experiments. When creating and running an experiment, users must save any and all image or stimuli files to their local “assets” folder.

System Requirements and Dependencies

To run on any operating system, the Social Vision Toolkit requires users download Python version 2.7 as well as Kivy and its dependencies. Additionally, I recommend users write their experiment script in a free open-source text editor like Sublime Text or Emacs.

Experiment Structure

An experiment file written using the Social Vision Toolkit consists of several different classes. Namely, three main types of classes or objects make up an experiment file: Experiments; Event Blocks; and, Tasks (see Figure 15). I provide greater detail about each class and subclass type below. In short, Tasks consist of the items, stimuli, or text seen by the participant. One or more tasks comprise an Event Block; one or more Event Blocks comprise an Experiment. Event Blocks and Experiments are purely abstract concepts; they allow researchers to organize their experimental design in logical segments.

Experiment Class. The experiment class is defined only once, at the very end of the experiment script using the `class ExperimentName(Experiment)` variable. It tells the program the order in which Event Blocks should be presented and whether any events should be repeated. The command to define the event order within an Experiment class reads like a mathematical equation, with sequential Event Blocks indicated by the + sign and repetitions of

an Event Block represented by a multiplicative *. Users can also specify whether the Event Blocks should occur in sequential order, whether they should be randomized, or some combination of the two. The Between Subjects, Reverse Correlation experiment script included in the Social Vision Toolkit has an Experiment class that specifies a specific Event Block order (see Figure 16 and Appendix C). In comparison, the Experiment class in the Within Subjects, Ensemble Coding experiment script specifies the exact order of the first two Event Blocks while randomizing across participants the order of the third and fourth Event Blocks. All participants then see the same final Event Block (see Figure 17 and Appendix C).

The Experiment class may take additional arguments, should the user desire greater complexity in study design. For example, in a within subject experiment, the user may wish to vary the number of a type of stimuli seen by a participant. This can be done within the blocks argument of the Experiment class command. In the Within Subjects, Ensemble Coding demo, participants were exposed to five different conditions where the ratio of two types of images in a single Image Grid differed. I wished the user to see 10 different image types within each condition and wanted them to appear shuffled within a single Event Block. I specified this via the `blocks` variable, instructing the program to show all participants the Instruct Ratio Event, followed by the condition-randomized Ratio Event (see Appendix C).

Event Block class. The Event Block class can be thought of as the design blocks within the experiment. It is called in the experiment script via the `class EventBlockName(Event)` variable. Generally, users will want to create a separate Event Block for each of their dependent variables, for any opening onscreen experiment instructions, and for any demographic information they wish to collect at the end. Each Event Block should have its own easily identifiable name, specified by the `event_id` variable; this will be the

header for that Event Block in the comma-separated values (CSV) output file containing participants' responses.

Event Blocks are built using one or more Tasks. As with Experiment Classes, the order of Tasks within an Event Block can be randomized or performed in sequential order. The `stages` variable within each Event Block takes an array containing the order of the Tasks constituting that Event Block. In the Example Experiment demo script, the TestEvent Block contains three tasks: TestStimulus, wait(1000), and TestTask (see Figure 18). These will be shown to participants in the exact order in which they appear. In the Within Subjects, Ensemble Coding demo script, the BelongingFitEvent Block provides an example of partially randomized Task order. The `dvs` variable allows the experimenter to further specify which parts of the Event Block will be randomized. The GroupFitQuestion Task and the GroupBelongingTask comprise two items in a scale that participants must respond to after every stimuli. I wished the order of the two questions to be randomized from trial to trial, while ensuring that for each trial, participants first saw a Fixation Cross and a test stimulus before the randomized dependent variables. Thus, Block Events allow the experimenter to control the exact presentation order of Tasks.

Task class. Participants will see onscreen only objects belonging to the Task class. Tasks can span a wide range of types and purposes. Everything from instructions about the experiment to timed stimuli presentation can be represented via a Task object. There are several different types of Tasks, divided into sub-classes which have been pre-programmed to accomplish slightly different objectives.

Stimulus sub-class. The Stimulus sub-class of Tasks allows experimenters to present images for a specified amount of time. The experimenter sets the duration of the stimulus

presentation via a `duration` variable. The experiment will automatically record any keyboard input from the participant during the stimulus presentation; however, that is not the goal of this sub-class, and the stimulus will not disappear prior to the specified duration. The experimenter must specify the location of the jpg stimulus image as well as its file name in the `def render(self)` variable.

Simple Text Stimulus sub-class. A specific type of Stimulus object, Simple Text Stimulus sub-class renders experimenter-written text on the screen for a specified duration of time. In the Between Subjects, Reverse Correlation demo script, for example, I used Simple Text Stimulus tasks to create 15 second breaks between sets of trials. Once the specified duration expires, the Event automatically progresses to the next Task.

Task sub-class. The Task sub-class presents a named image (which should be saved in the local version of the `assets` folder) onscreen. The Task sub-class was designed to record participants' responses; thus it requires participant input before the program advances to the next Task or Event. In the Within Subjects, Ensemble Coding experiment script, Task images include the dependent variables of interest and 7-point Likert scales. As in the Stimulus sub-class, the experimenter must specify under the `def render(self)` variable the directory location and the exact filename of the image to be presented for that task. The Task sub-class also allows experimenters to specify `acceptable_keys`. If specified, the Event will only advance and the Task will only record a key listed in the array of `acceptable_keys`.

Simple Text Task sub-class. Best suited for presenting instructions, the Simple Text Task sub-class of Task objects only shows experimenter-determined text to the participant. It requires participants to press a key to advance the Event and for the next Task to appear.

Demo Experiment Scripts

The three demo files included in the Social Vision Toolkit represent different types of studies with varying levels of programming complexity. The Example Experiment file constitutes the most introductory file; it shows text onscreen for 1000ms, has a 1000ms delay, then asks the participant to press any key. This sequence of events repeats three times before the experiment ends. The Between Subjects, Reverse Correlation and Within Subjects, Ensemble Coding demo scripts were adapted from the full length scripts used to generate Study 3 and Study 1b, respectively, in this dissertation. They were chosen to demonstrate the breadth of what the Social Vision Toolkit can accomplish. Figures 16, 17, and 18 outline the Experiment design for each demo script; the Event Block and Task labels in each figure correspond to their respective names in the code. All three of the demo scripts will run in their current form, assuming the user has downloaded the proper version of Python and Kivy dependencies.

To better understand how to read and edit the demo script code, I will go line-by-line through the Example Experiment demo script now. It is rendered in its complete form in Appendix C. The script starts with a list of dependencies, which should appear at the top of every experimental script and should not be changed by the user:

```
from stimulus import Stimulus
from task import Task
from event import Event, wait
from experiment import Experiment
from ui import run_app

from kivy.uix.button import Button
```

These six lines of code call functions that were created in the framework files to the current experiment file; this allows the program to run them during the experiment. Next, the script defines the first Task it will use:

```
class TestStimulus(Stimulus):
```

```
 stimulus_id = "test_stimulus" # Identifies this stimulus in the results.
 duration = 1000     # Show this stimulus for 1000 ms.
 text="Test"
```

We have given the TestStimulus Task the stimulus id “test_stimulus”; this is the variable label it will have in the data output file. Additionally, because the Task belongs to the Stimulus sub-class, I must specify how long the image remains onscreen. Here I have called it for 1000 ms. Finally, the last two lines in the `class` variable define the Stimulus as the text “Test” and provide its layout on the page.

The next snippet of code defines a second Task as the TestTask. It labels the task as “test_task” for the results output and again allows the experimenter to edit the text seen by the participant. Because this is a Task with no sub-class, it requires input from the participant before the experiment will advance. There is no duration variable, as it is irrelevant to this Task.

```
class TestTask(Task):
 task_id = "test_task" # Identifies this task in the results.

 def render(self):
  self.layout.add_widget(Button(pos=self.layout.pos, text="I am a task. Press
any key to continue."))
 run_app(TestExperiment)
```

After defining the two Tasks, I create a new Event Block by calling the `class` variable again but specifying `Event` in parentheses. The `event_id` variable names the event block for the results output. I define the order of the Tasks I want to include in the Block Event in the `stages` variable, by providing a sequential array. Here, the program will present the TestStimulus Task, wait 1000ms, then present the TestTask. The `wait()` command has been built into the framework of the Social Vision Toolkit, so users need not define this variable in their own experiment script.

```
class TestEvent(Event):
 event_id = "test_event" # Identifies this event in the results.
 stages = [TestStimulus, wait(1000), TestTask]
```

Next, I define the Experiment class via the last `class` variable. The `events` variable underneath it specifies the order of the Block Event(s). In this demo script, I have only one Block Event (TestEvent), but I want it to repeat three times. The multiplicative asterisk indicates repetition, with the desired number of repetitions following it.

```
class TestExperiment(Experiment):
 events = [TestEvent] * 3
```

Finally, the last line of code only requires the user to update the `run_app` command with the correct name of their experiment. Because I defined the Experiment above as TestExperiment, I have written that in the `run_app` command. The rest of the `if` function should remain unchanged in order to ensure the script runs as intended when called.

```
if __name__ == '__main__':
 run_app(TestExperiment)
```

Executing Experimental Scripts

Having written and saved their script into a local forked version of the Social Vision Toolkit, experimenters can execute their scripts directly from the Terminal. They must navigate into the Social Vision Toolkit from their home directory. Once there, they can call the script via the command `python experimentname.py` . For example, to call the Example Experiment demo script, the experimenter would type `python example_experiment.py`

This will trigger a box containing the opening frame of the experiment to appear onscreen (see Figure 19). The opening frame includes a text entry box where researchers can record the id of the participant about to perform the experiment. It also allows the experimenter to select whether the participant is in Condition A or B for a between subjects study.

Each experiment created using the Social Vision Toolkit saves participants' responses in the same directory as the local repository location. For each session in which participants are run, the program saves a separate CSV file labeled with the date and time that the experiment began. The CSV files can be then processed and compiled using the experimenter's preferred statistical analysis software.

Discussion

The Social Vision Toolkit was designed to create a free, collaborative tool for social scientists to present experimental stimuli to participants. It supersedes the need for expensive, proprietary software, as it fulfills many of the same objectives. Additionally, it is provided free of charge for download from an open-access online repository and operates across platforms. Users may view the full code base, allowing them to see exactly what the program does and how it operates. Furthermore, it encourages input from other researchers who can adapt experimental scripts to their own needs and suggest overall improvements to the code. The Social Vision Toolkit differs from prior open-source stimuli frameworks in its declarative nature and higher-level functioning. Although lacking a GUI, I believe it presents a lower bar of entry for researchers newer to Python and comes pre-loaded with relevant features like automatic data recording. Thus, I hope the Social Vision Toolkit may help social scientists create novel, more intricate experiments using a range of stimuli.

Currently, the Social Vision Toolkit comes preloaded with three types of object classes which can be combined to create a between- or within-subject experiment. The Social Vision Toolkit allows researchers to easily present text or images to participants in a randomized or structured order and records participants' keyboard responses. Future improvements to the base package will include ready-to-go subclasses for presenting audio and video stimuli, the option to include an onscreen progress bar, and additional input classes like radio buttons and sliding scales. Written in Python, the Social Vision Toolkit also has the capability to integrate with other technologies like eye tracking or mouse tracking software. The Social Vision Toolkit is highly customizable and may suit many other researchers' needs beyond those originally envisioned by the authors.

In the spirit of open science, I encourage any interested developers to create their own Task sub-classes and submit a merge request on the GitHub repository. I have chosen to share the Social Vision Toolkit on GitHub in part to spur peer review and collaboration across institutions. I would hate to see other labs have to "recreate the wheel" or rely on expensive software, when a free alternative exists in the Social Vision Toolkit. Further, GitHub's forking feature makes it easier for researchers to share their experiment scripts, data and code for statistical analyses. Thus, other labs may find it easier to replicate prior findings and build upon these research questions. I hope that making the Social Vision Toolkit public further pushes social science towards greater methodological transparency and collaboration, as well as broadens the scope of what is experimentally possible.

Conclusion

The goal of this dissertation was to examine how individuals extract important, summary information from groups of people. Through two sets of studies and a methods paper, I sought to understand the ways in which changes in sex ratio influence subjective feelings and downstream judgments about the individuals in the group.

In Study Set One, I examined whether perceivers can accurately and rapidly extract actuarial and social summary statistics from 8 and 12-person ensembles. Participants briefly saw random groups varying in sex ratio from all men to all women. The results suggest perceivers can accurately encode changes in the proportion of male to female group members and view male-dominated groups as increasingly sexist. Additionally, perceivers feel less belonging as the number of same-sex group members decreases. Through reverse correlation, Study Set One also provided a glimpse into the mental representation generated when individuals see a male- or female-dominated group. The average group member from majority-male ensembles appears angrier and less welcoming than the average group member from majority-female ensembles. Findings from Studies 1-3 suggest ensemble coding evokes emotional responses in perceivers and may contribute to women's likelihood to approach or avoid male-dominated fields.

Study Set Two probed whether the ensemble coding process explored in Study Set One replicates with real-world groups. Participants saw 4-person arrays based on panels from a prominent social psychology conference. Despite exposure to noisier stimuli, perceivers again accurately reported changes in the perceived sex ratio and facial gender of the average group member across conditions. Furthermore, participants also anticipated that the group's likelihood to endorse sexist norms would increase as the ratio of men to women decreased. As in Studies 1a, 1b, and 2, perceivers' feelings of belonging depended on the number of same-sex others in

the group. Study 5 extended prior theory by examining whether perceivers encode information about the group's intellectual merit and potential for broader impacts gets encoded from a brief exposure to the ensemble. College students saw majority-male groups as less likely to engage in work with broader impacts, including mentoring underrepresented groups in STEM and increasing public engagement. Women perceived all groups, regardless of sex ratio, as having greater intellectual merit than men. Study Set 2 tested the external validity of my theory and demonstrated that real-world ensembles are nevertheless summarized in ways that may impact group members' goals.

The last chapter of this dissertation described the underlying Python framework I developed to test my theory: the Social Vision Toolkit. It presents the Social Vision Toolkit's history and context, contrasting it with other similar proprietary and open-source software. I outlined its advantages and highlighted its potential contribution to open science. Additionally, this chapter walked the reader how to design an experiment script using the Social Vision Toolkit. It also presented a line-by-line interpretation of a demo experiment included in the base package. Essential to answering research questions central to my dissertation, the creation of the Social Vision Toolkit may have the collateral consequence of contributing to the open science movement. Other researchers will be able to access and adapt our experiment scripts online via an open-access repository. The Social Vision Toolkit broadens the possibilities of research design and stimuli presentation, constituting a new methodological tool.

Taken together, my dissertation research highlights how instantaneous perceptions of a group's sex ratio drive immediate social attitudes and feelings of belonging. Even before interacting with a group, my findings suggest perceivers rapidly and accurately appraise the number of same-sex others present. This perception in turn drives beliefs about the group's

endorsement of sexist norms, its ability to contribute to societal advancement, and the average group member's approachability.

Research from Study Sets One and Two has important implications for women or men entering majority-single sex spaces. Prior studies have demonstrated how environmental cues can affect feelings of belonging and interest in counter-stereotypical professions. Cheryan et al. (2009), for example, showed that women who completed a career interest form in a room with Star Trek posters on the wall and video game boxes on the tables were significantly less likely to want to pursue a computer science career than their male peers. Follow-up studies revealed that feelings of belonging mediated women's interest in stereotypically "masculine" or geeky jobs (Cheryan et al., 2009; Good, Rattan, & Dweck, 2012). My research builds on these findings by demonstrating how feelings of belonging can occur after only a half-second exposure to a group. Even before an individual interacts with others or consciously processes environmental cues, she develops an innate sense of fit and anticipates a certain level of gender discrimination based on the group's sex ratio. This suggests, for example, that women walking into a male-dominated math classroom for the first time make immediate social judgments about whether they belong and whether the group will accept them. A similar process may also occur for men walking into a majority-female space, like a nursing or primary education class.[8]

Prior work has often framed questions of belonging in terms of stereotypical fit. However, my research cannot speak to this relationship explicitly. I did not ask study participants to categorize the type of group they saw (e.g., a math versus a humanities study group) or their interest in stereotype-consistent careers. However, Study 3 may touch on this question

[8] According to the U.S. Department of Education (2017), women still constitute the majority of B.A. recipients in nursing and early education.

tangentially. Participants' implicit mental representations of the average group member were rated by a naïve sample along traits consistent with gender stereotypes. In particular,4 the composite average face from a majority-male group was rated as angrier and less approachable than the composite image from a majority-female group. These characteristics align with stereotypes of men as more aggressive, more sexist, and less emotional than women (Broverman et al., 1972; Ghavami & Peplau, 2012; Wood & Eagly, 2010). Similarly, in Study 5, participants rated majority-female groups as more likely to share their work with their communities and mentor underrepresented minority individuals. These findings align with stereotypes of women as caring and motherly (Ghavami & Peplau, 2012).

Although results from Studies 3 and 5 hint at activation of stereotype-consistent perceptions, future studies should test this hypothesis directly. It may be that perceivers' endorsement of stereotype-consistent beliefs mediates the effect of a group's sex ratio on their feelings of fit and belonging. Alternatively, participants may have automatically and subconsciously categorized a group into a stereotype-consistent domain depending on its sex ratio; thus, their perceptions of belonging and discriminatory group norms could have been driven by these stereotypes themselves rather than the sex ratio per se. Follow-up studies will need to tease apart the exact relationship between gender stereotypes, social attitudes and affordances, and instantaneous perception of a group's sex ratio.

Future research should also consider testing systematic sampling of targets within a set as a competing hypothesis to ensemble coding. The current methodology does not allow me to determine unequivocally whether participants extracted a descriptive average of the group's sex ratio or whether they focused on a few individual targets in the group. Either strategy may result in similar outcomes, as participants who systematically sample only a few faces would be more

likely to see male targets in a majority-male group. This could then lead them to select an image with higher Perceived Facial Masculinity and more male stick figures on the Perceived Sex Ratio scale. To determine whether perceivers rely on the innate base rate of men to women or extract descriptive characteristics about the whole group would require different methodology than I have employed. For example, researchers could integrate eye tracking software into their study design to measure the onscreen location and duration of participant gaze. This would provide clarity on how perceivers visually process each ensemble; one could empirically test whether participants scan the whole group or focus on a subset of target faces. Despite an inability to identify the exact operating mechanism, the social consequences of my findings remain unchanged. As the number of men to women in the ensemble increased, perceivers felt concomitant shifts in belonging, judged the group to be more hostile towards women, and rated the group as contributing work with fewer broader impacts. While I believe the above studies present evidence for ensemble coding, future research should test competing processes including systematic sampling of target faces.

Drawing on concepts from social cognition and social psychology, ensemble coding has emerged as a way to conceptualize groups of objects. It explains gist abstraction through a statistical lens, complementing existing cognitive and evolutionary research about how humans perceive the world. Shown a set of objects, perceivers can accurately and rapidly summarize statistics about the items. This ability comes at the seeming expense of local-level perception; perceivers have more difficulty recalling information about individual targets compared to group-level characteristics. With concern for methodological validity and statistical rigor, my dissertation considered the downstream consequences triggered by ensemble coding judgments. Akin to appreciating an Impressionist painting, we have stepped back a few yards to recognize

that individual objects may instead be coded as a whole scene; our task is now to make sense of what the perceptual picture says and how it affects our behavior hereafter.

Tables and Figures

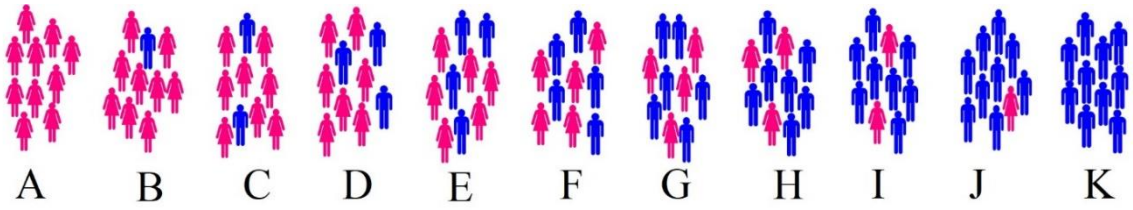
Table 1. Post-hoc power analyses of sample size from Studies 1a and 1b

Study	Tested Effect	Achieved Power
1a	Main Effect of Actual Sex Ratio on Perceived Facial Masculinity	1.0
1a	Main Effect of Actual Sex Ratio on Belonging	0.872
1a	Main Effect of Condition on Perceived Sexist Norms	1.0
1a	Main Effect of Actual Sex Ratio on Perceived Sex Ratio	1.0
1a	Interaction effect of Gender by Actual Sex Ratio on Belonging	1.0
1b	Main Effect of Actual Sex Ratio on Perceived Facial Masculinity	0.9982
1b	Main Effect of Actual Sex Ratio on Belonging	0.9958
1b	Main Effect of Actual Sex Ratio on Perceived Sexist Norms	1.0
1b	Main Effect of Actual Sex Ratio on Perceived Sex Ratio	1.0
1b	Interaction effect of Gender by Actual Sex Ratio on Belonging	1.0

Note: All power analyses indicate the proportion of times the modeled effect was significant ($p<0.05$) when tested on 10,000 randomly simulated datasets.

Figure 1. Sample stimulus from Study 1a in which Actual Sex Ratio varied within each ensemble. This example depicts an Actual Sex Ratio of 6 men and 2 women.

A "Please indicate by selecting the corresponding radio button below which PICTURE (labeled by letter) best represents the ratio of men to women you just saw."



B "Please indicate by selecting the corresponding radio button below which face best represents the AVERAGE group member in the group just shown?"

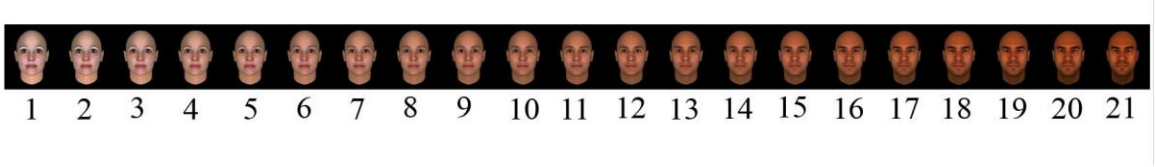


Figure 2. Dependent measures of actuarial summaries assessing Perceived Sex Ratio (panel A) and Perceived Facial Masculinity (panel B) used in Studies 1a, 1b, and 2.

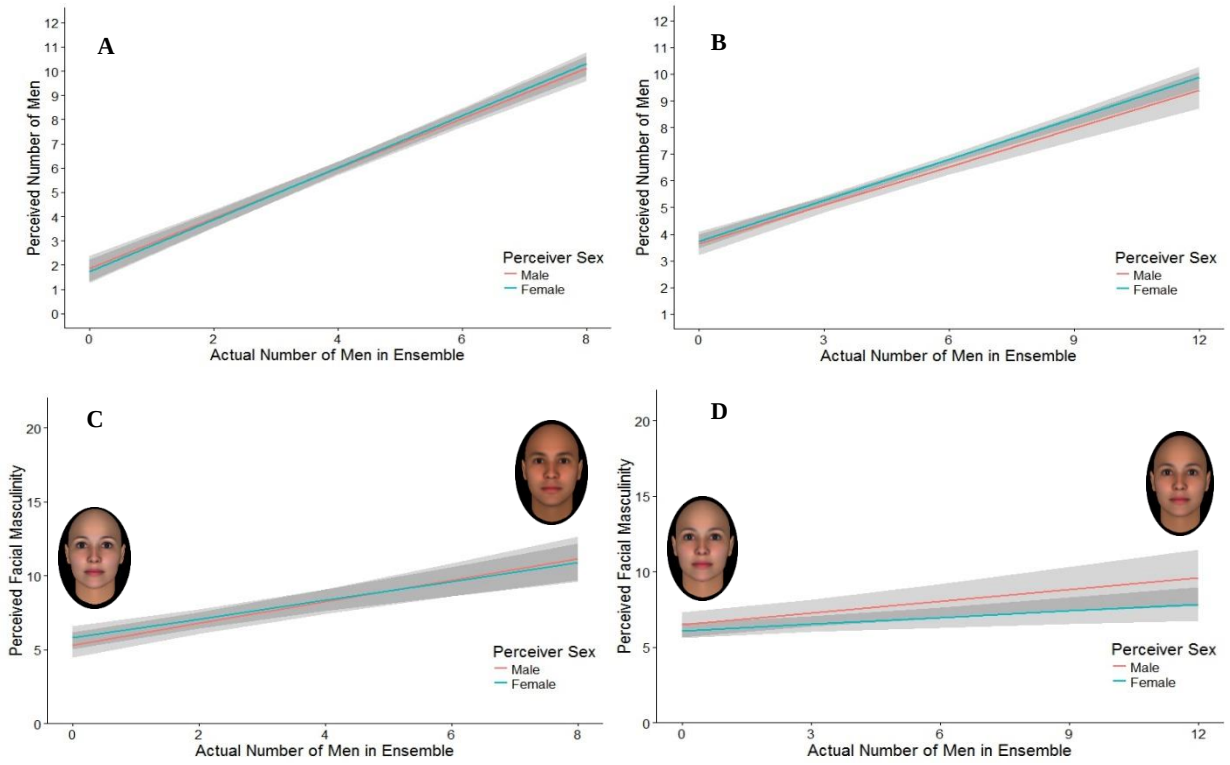


Figure 3. Perceived Sex Ratio and Facial Masculinity in Studies 1a and 1b. After 500 ms exposure, perceivers accurately report a group's sex ratio in Studies 1a (panel A) and 1b (panel B). Perceivers chose a more masculine-gendered face to represent the average group member in Studies 1a (panel C) and 1b (panel D) as the number of men in the group increased. The median face corresponding to participants' ratings for ensembles comprised of all-women or all-men are shown above their respective conditions.

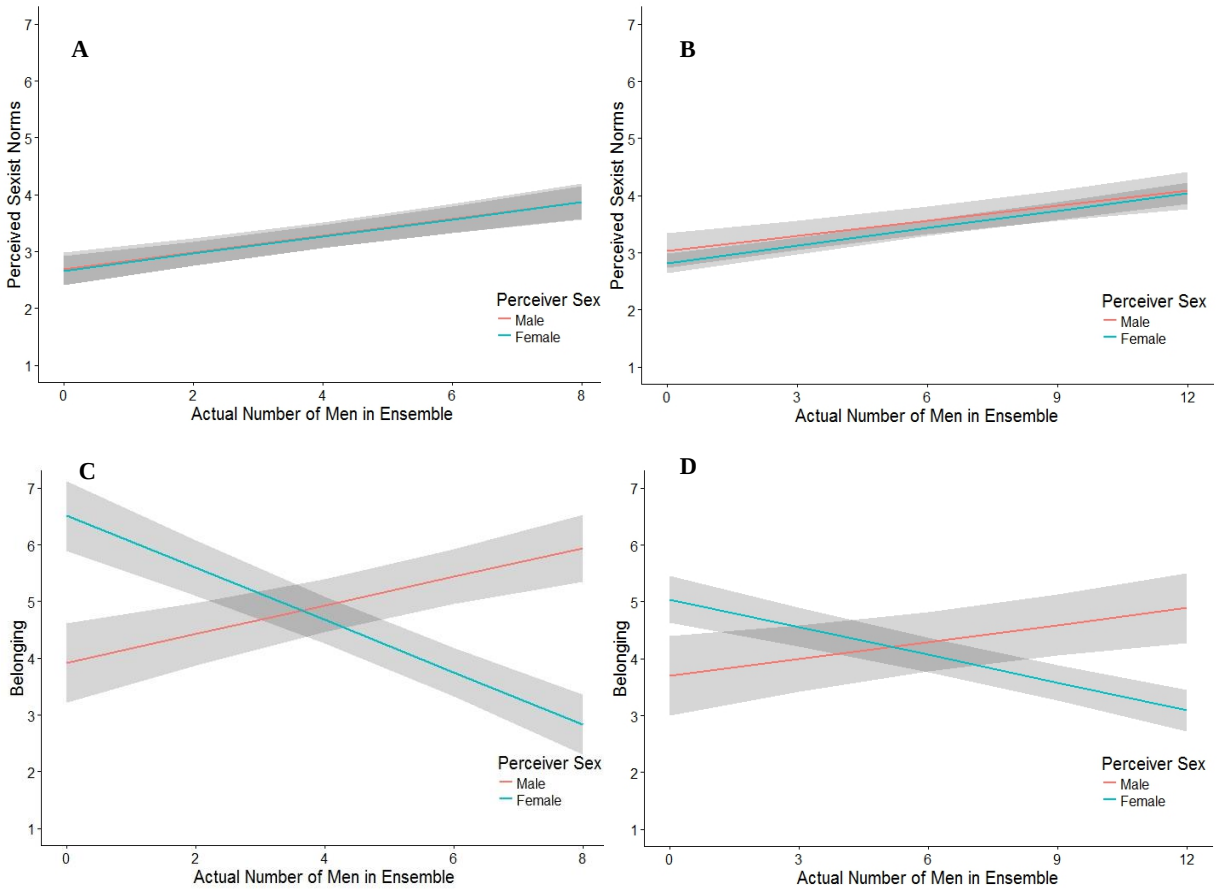


Figure 4. Perceived Sexist Norms and Belonging in Studies 1a and 1b. Participants view male-dominated groups as having more sexist norms in Studies 1a (panel A) and 1b (panel B). Their feelings of Belonging depended on the number of same-sex others in the group in Studies 1a (panel C) and 1b (panel D).

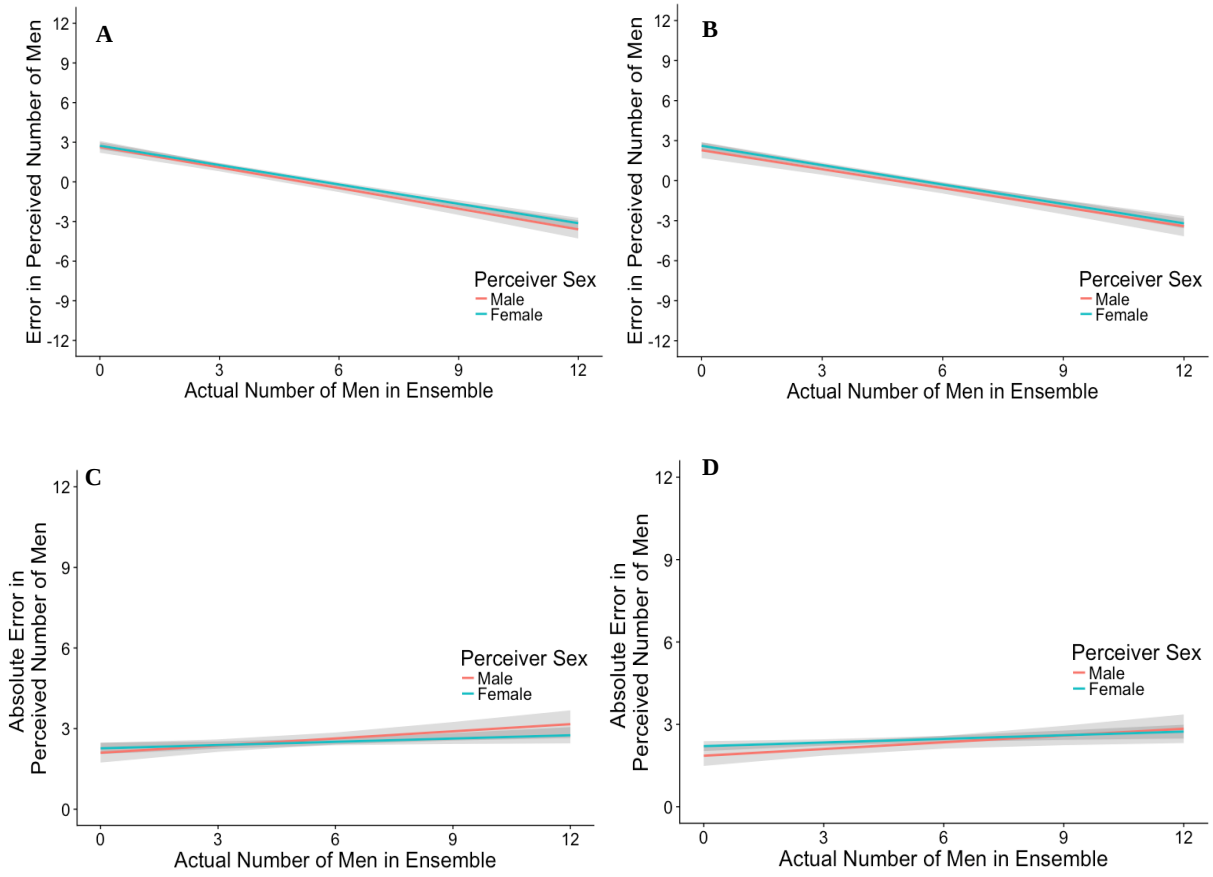


Figure 5. Error in Perceived Sex Ratio in Studies 1b and 2. Perceivers overestimate the number of men in an ensemble less as the ratio of men to women increases in Studies 1b (panel A) and 2 (panel B). Absolute error in perception, however, grew significantly as the sex ratio increased in Studies 1b (panel C) and 2 (panel D).

Table 2. Mediation models comparing the effect of Perceived Same-Sex Others on Actual Same-Sex Other's influence on Belonging

Number of Actual Same-Sex Others	Direct Effect	Indirect Effect	95% CI for Indirect Effect	Proportion Mediated
3	0.57**	0.09**	[0.05, 0.14]	0.14**
6	1.14**	0.18**	[0.09, 0.27]	0.14**
9	1.71**	0.27**	[0.13, 0.41]	0.14**
12	2.28**	0.36**	[0.18, 0.54]	0.14**

*Note: All mediation models compare the Actual number of Same-Sex Others in the ensemble to the control condition of 0 Same-Sex Others in ensemble; *p≤.05, **p≤.01*

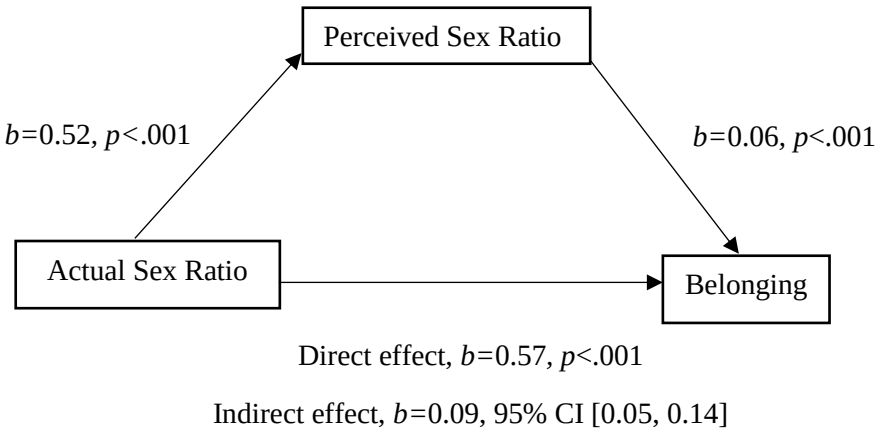


Figure 6. Study 2 Mediation Model. In Study 2, Perceived Sex Ratio mediated the effect of Actual Sex Ratio on feelings of belonging.

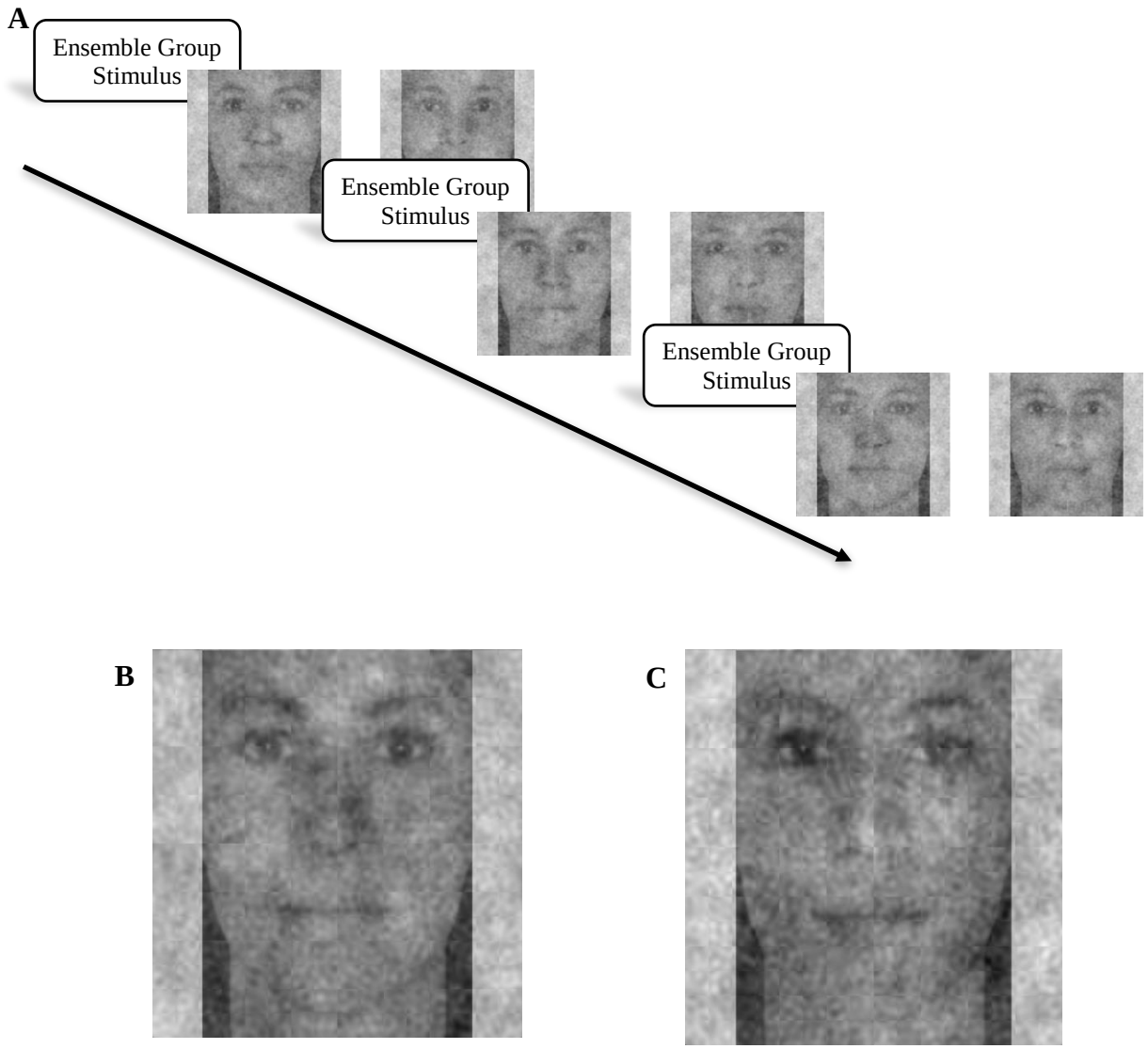


Figure 7. Study 3, Stage 1 design and Mental Representations of Majority-male versus Majority-female groups. In Study 3, Stage 1, participants completed 700 trials in which they saw a majority-male or majority-female ensemble. They indicated after each trial which of two faces most resembled the average group member (A). Their choices were aggregated into two composite images derived from the majority-male (B) and the majority-female (C) conditions.

Table 3. Majority-female ensembles rated more favorably than the mental representation of majority-male ensembles in Study 3

	Mental Representation of Majority-Male Ensembles		Mental Representation of Majority-Female Ensembles		
Traits	**M**	**SD**	**M**	**SD**	**t-test**
Angry	2.60	1.47	2.15	1.26	4.46**
Approachable	4.30	1.36	4.85	1.19	-5.05**
Attractive	4.16	1.32	4.83	1.18	-6.25**
Competent	4.46	1.21	4.55	1.12	-0.94
Friendly	4.19	1.28	4.78	1.12	-5.46**
Happy	3.92	1.34	4.68	1.20	-6.25**
Intelligent	4.42	1.19	4.55	1.22	-1.28
Masculine	4.00	1.28	3.09	1.30	6.71**
Warm	3.99	1.38	4.65	1.31	-5.25**

*Note: * $p \leq .0056$; **$p < .0001$, with Bonferroni correction*

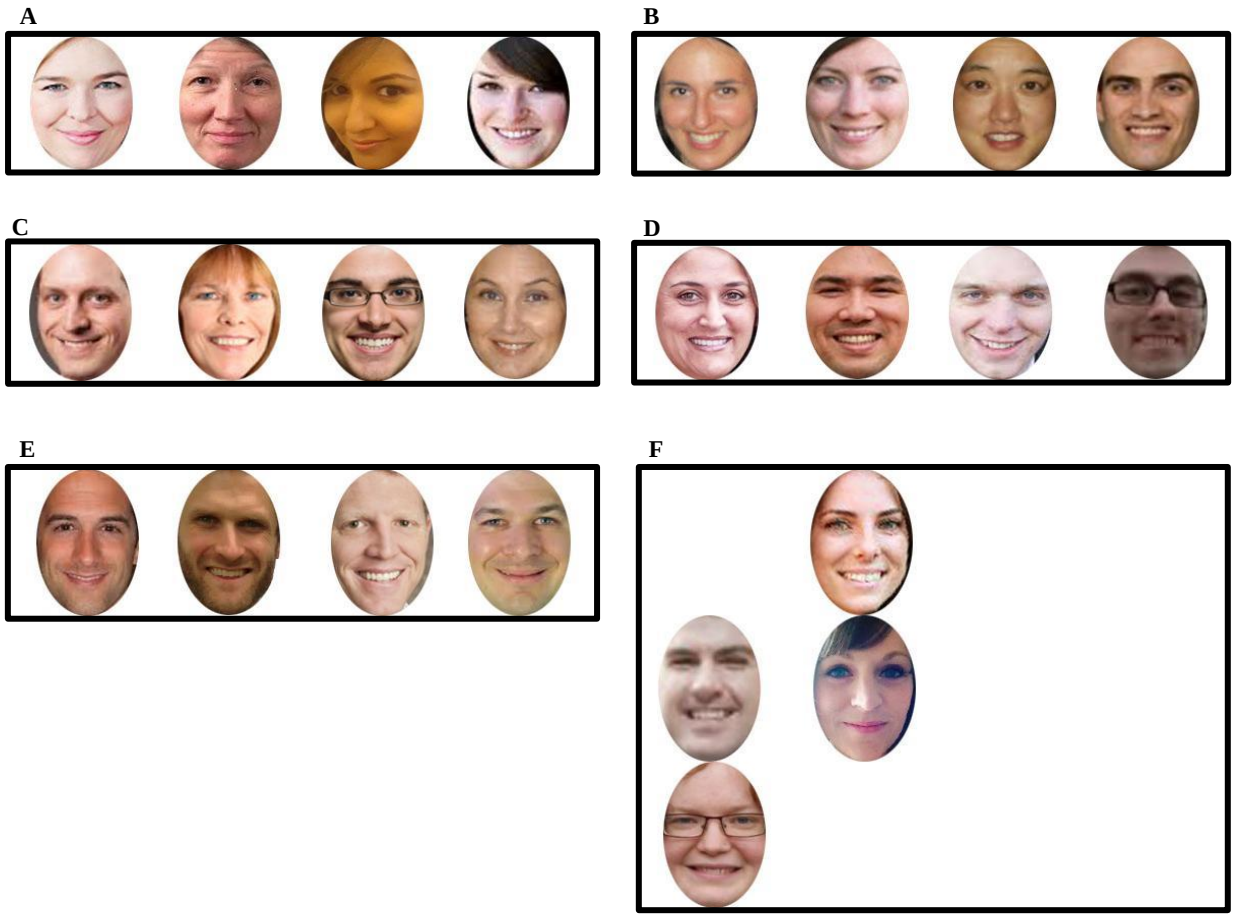


Figure 8. Sample stimuli used in Studies 4a and 5, in which the ensemble was comprised of all-women (panel A), 1 man: 3 women (panel B), 2 men: 2 women (panel C), 3 men: 1 woman (panel D), or all-men (panel E). Panel F shows a sample stimulus generated by the Python script in Study 4b, with individual target faces randomly placed in a 3x4 array.

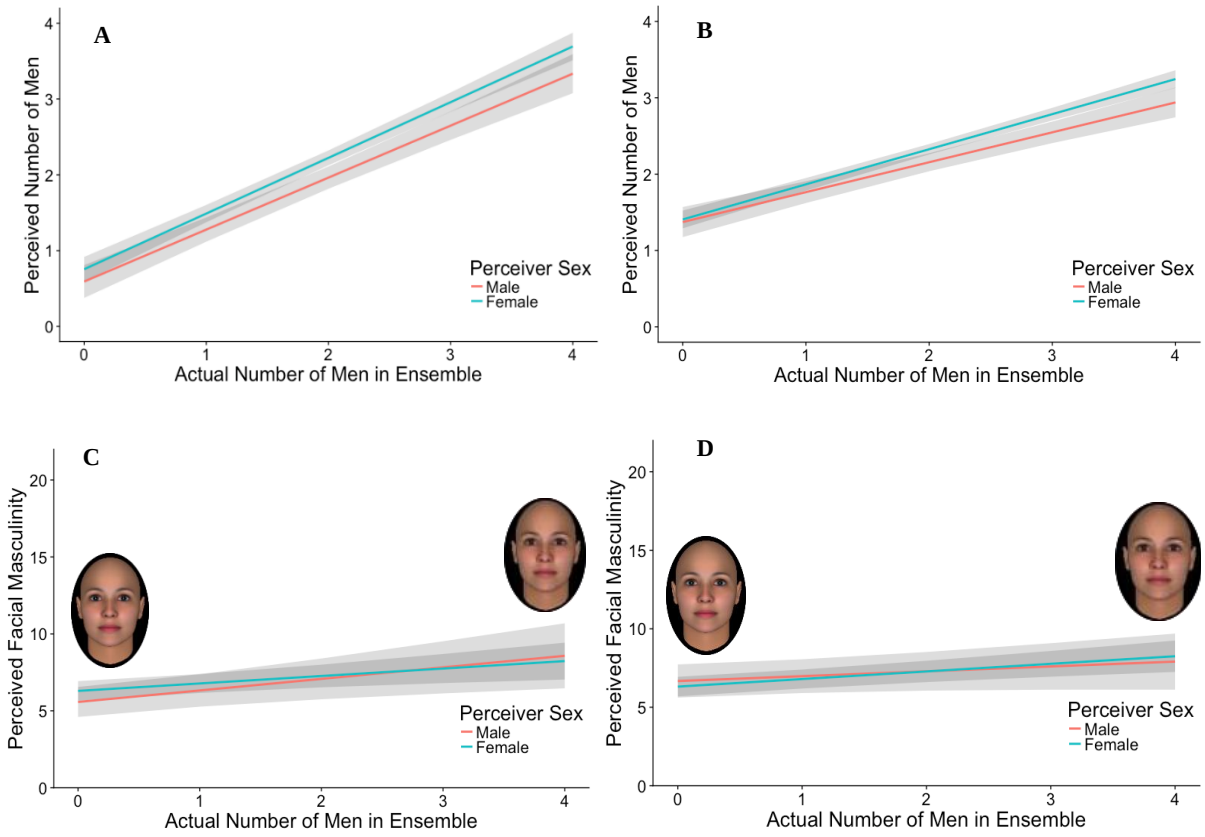


Figure 9. Perceived Sex Ratio and Perceived Facial Masculinity in Studies 4a and 4b. After 500 ms exposure, perceivers accurately report a group's sex ratio in Studies 4a (panel A) and 4b (panel B). Perceivers chose a more masculine-gendered face to represent the average group member in Studies 4a (panel C) and 4b (panel D) as the number of men in the group increased. The median face corresponding to participants' ratings of all-women or all-men ensembles are shown above their respective conditions.

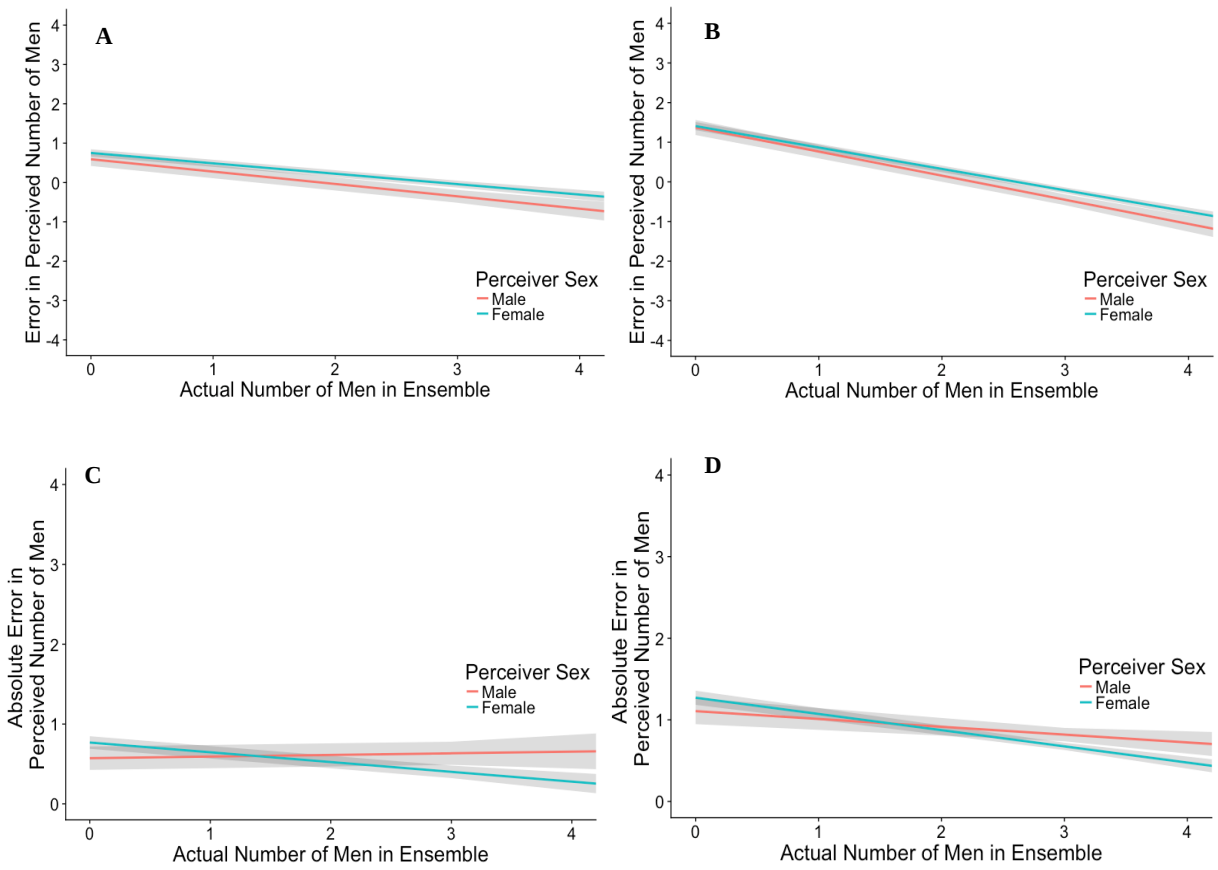


Figure 10. Error in Perceived Sex Ratio in Studies 4a and 4b. Perceivers overestimate the number of men in an ensemble less as the ratio of men to women increased in Studies 4a (panel A) and 4b (panel B). Male perceivers made similar number of errors across conditions when estimating the Perceived Sex Ratio. Female perceivers, however, made significantly fewer estimation errors as the ensemble sex ratio increased in Studies 4a (panel C) and 4b (panel D).

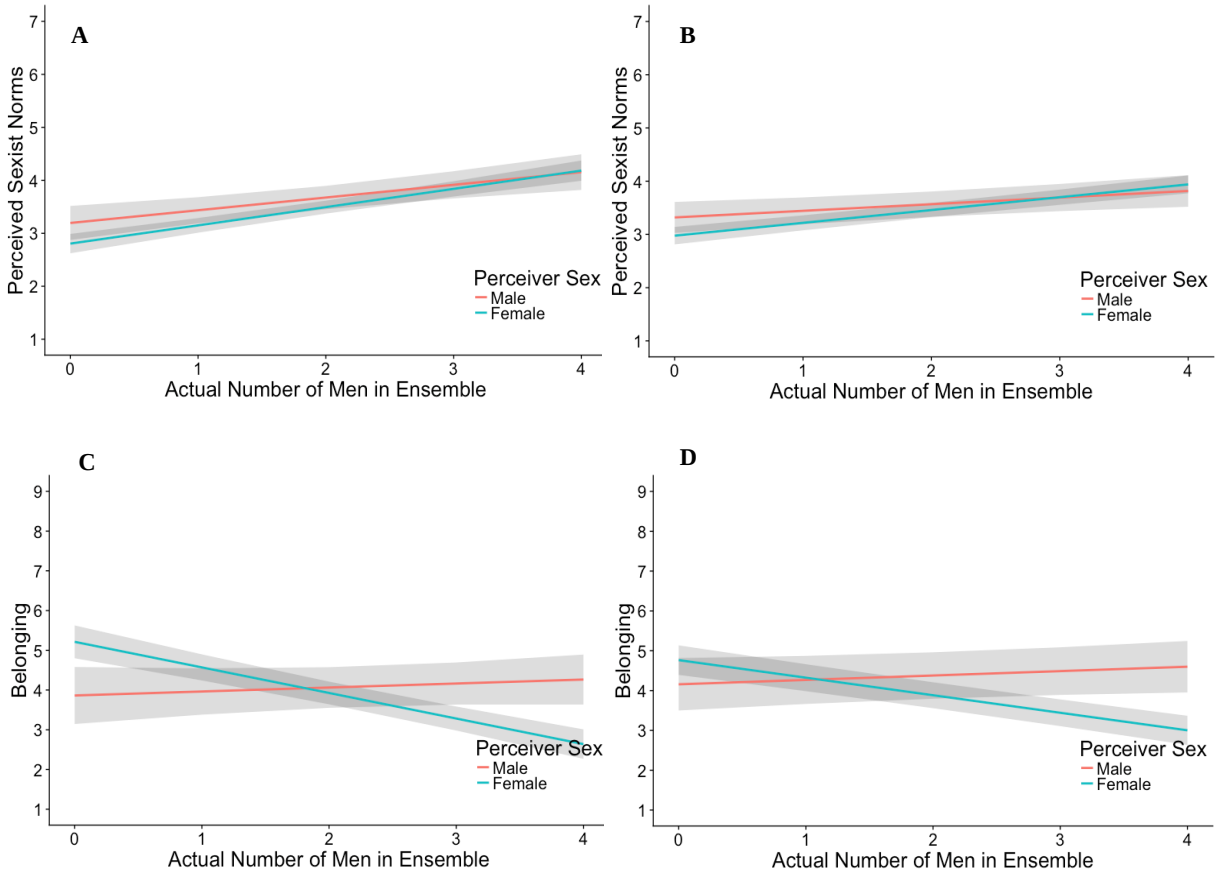


Figure 11. Perceived Sexist Norms and Belonging in Studies 4a and 4b. Participants view male-dominated groups as having more sexist norms in Studies 4a (panel A) and 4b (panel B). Their feelings of Belonging depended on the number of same-sex others in the group in Studies 4a (panel C) and 4b (panel D).

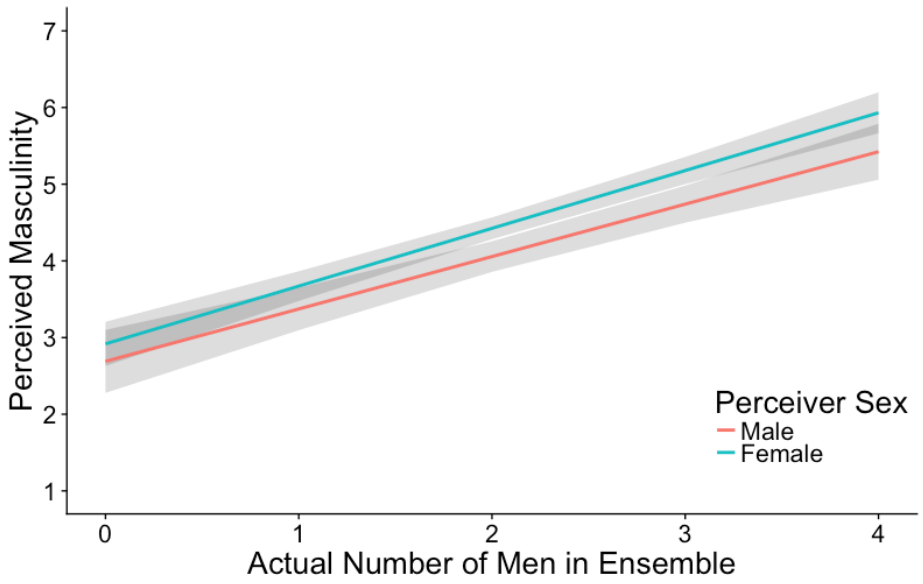


Figure 12. Perceived Masculinity in Study 5. Participants in Study 5 rated the average group member as significantly more masculine as the ratio of men to women in the ensemble increased.

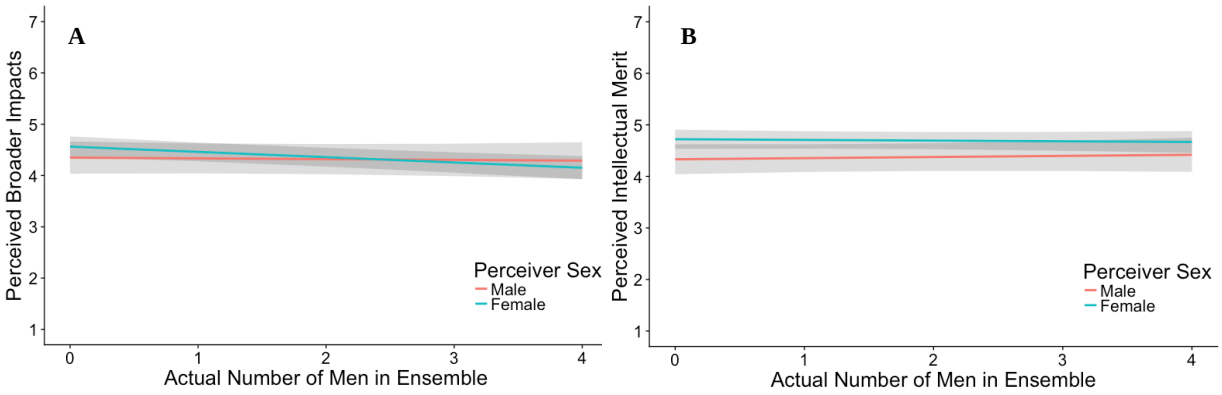


Figure 13. Perceived Broader Impacts and Perceived Intellectual Merit in Study 5. In Study 5, participants view male-dominated groups as contributing work with fewer broader impacts (panel A). Belief in the Perceived Intellectual Merit depended on the sex of the perceiver, with women rating all ensembles as significantly higher in intellectual merit (panel B).

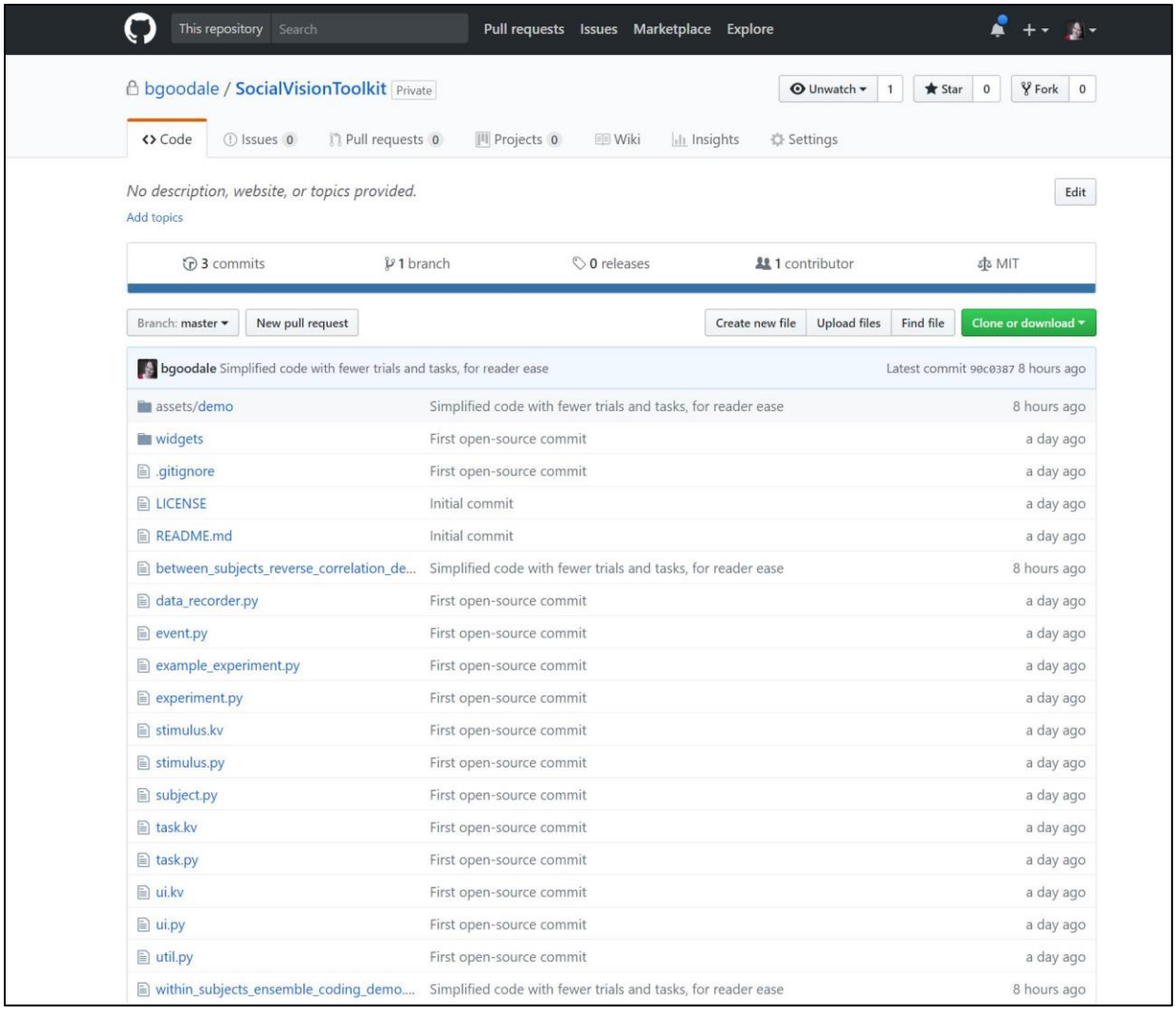


Figure 14. Social Vision Toolkit on GitHub. A screenshot of the GitHub repository where readers can access the Social Vision Toolkit.

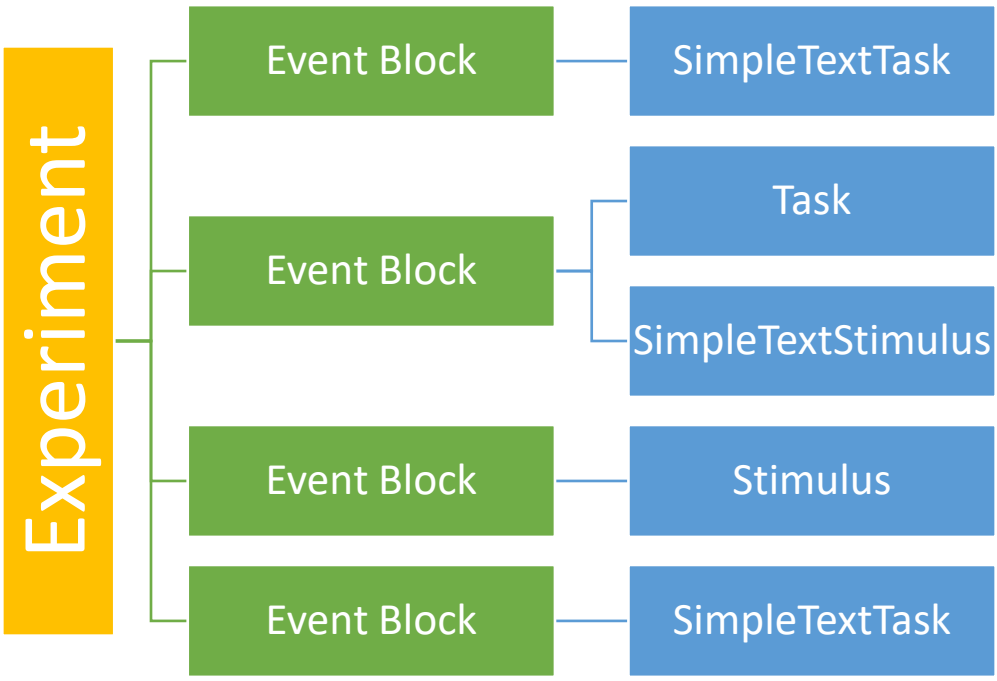


Figure 15. Overall hierarchical structure of object classes used in the Social Vision Toolkit. The Experiment class encompasses all subclasses and indicates the order in which subsequent Event Blocks and Tasks should be executed. Event Blocks constitute the second level of classes; users may specify whether the Event Blocks are randomized or follow a specific order within the program. Finally, individual Tasks constitute the third level. These are the images, text, and/or stimuli that are shown to the participant. Each Event Block must have one or more task within it. The above figure shows an experiment with four Event Blocks organized in sequential order; the first, third, and last Event Block all have only one task within them while the second Event Block has two.

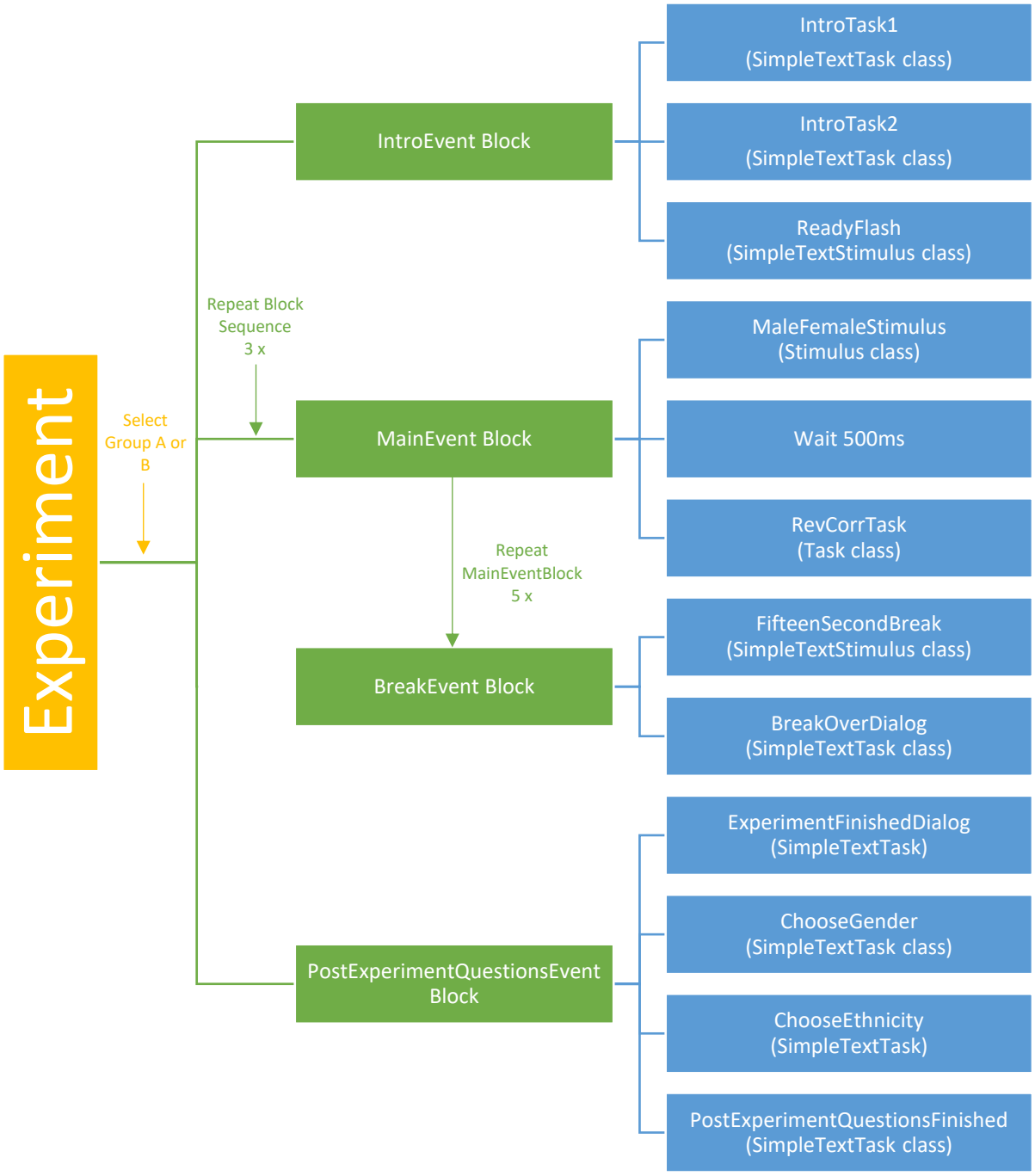


Figure 16. Class structure and design specification for the Between Subjects, Reverse Correlation demo included in Social Vision Toolkit. Experimenters select the condition (A or B) on the opening screen, which determines the ratio of men to women (3 men: 9 women v. 9 men: 3 women) seen by participants during the study

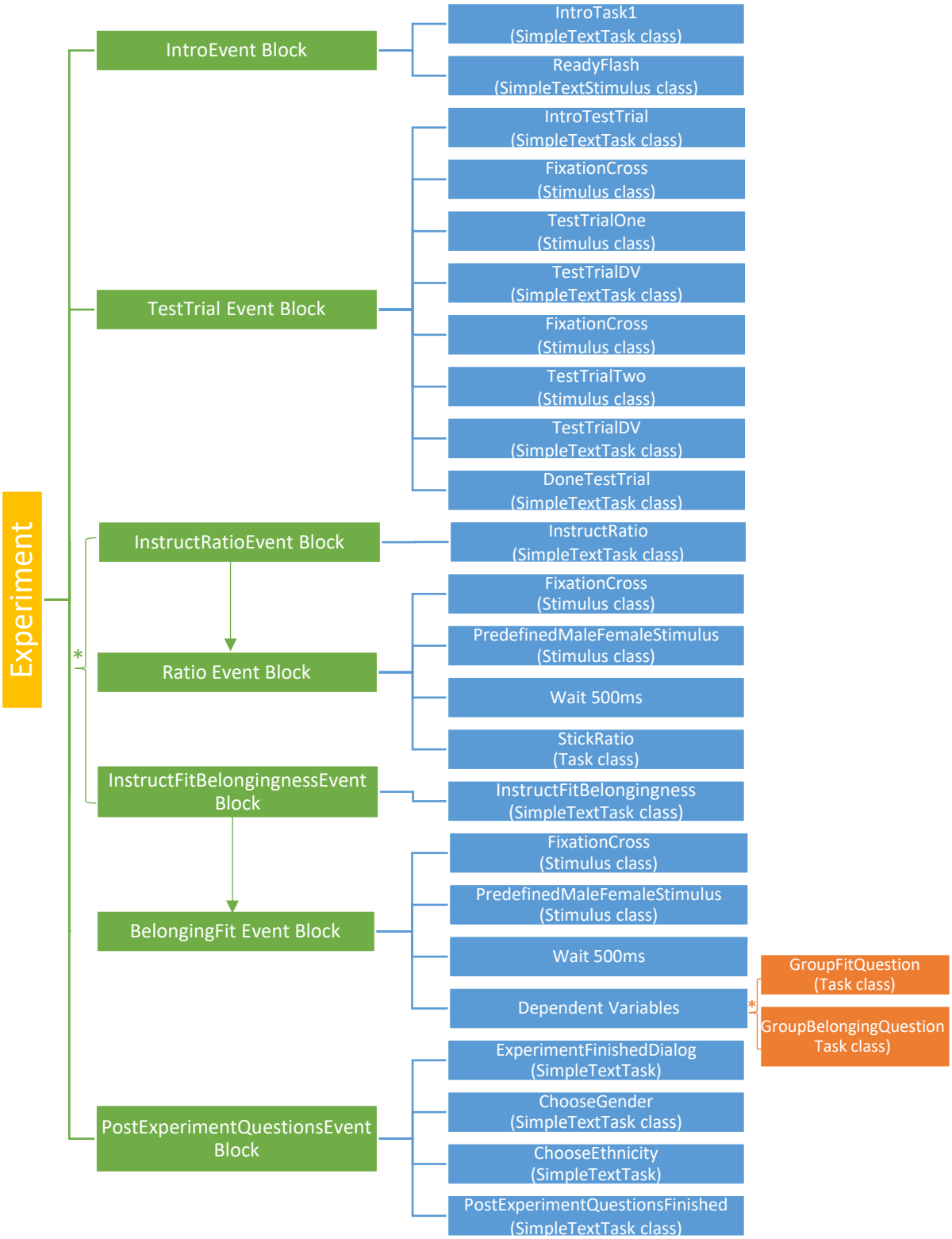


Figure 17. Class structure and design specification for the Within Subjects, Ensemble Coding demo included in Social Vision Toolkit. *Note: * indicates block or task order randomization.*

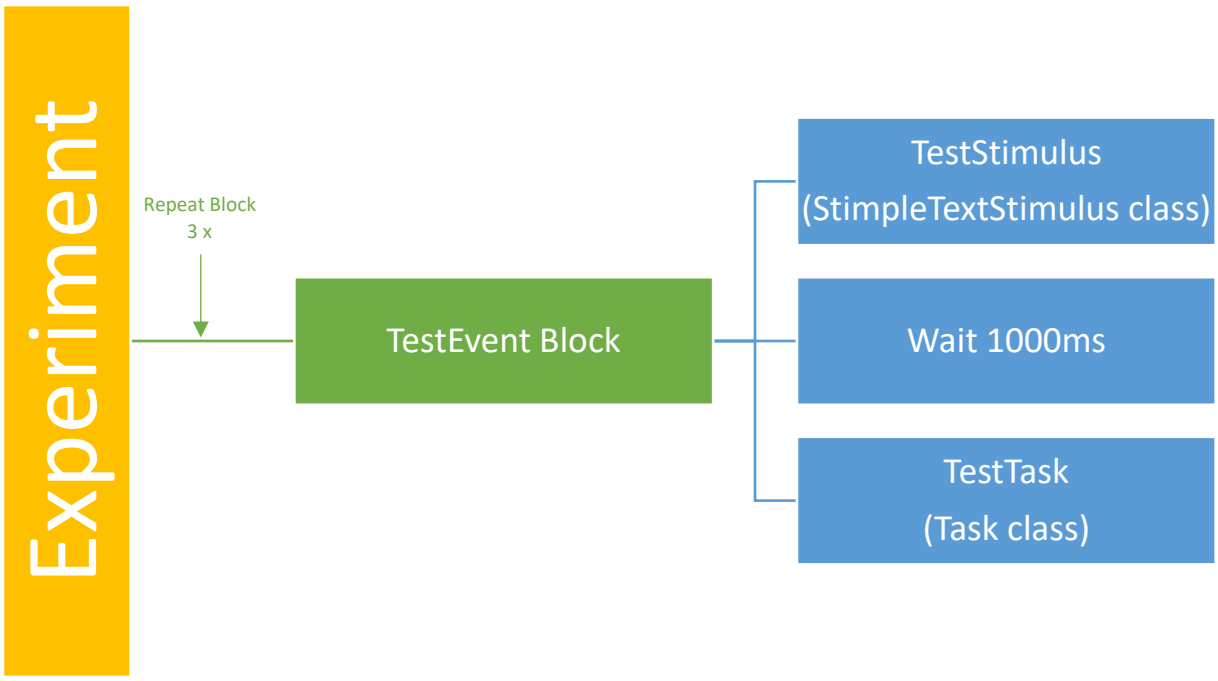


Figure 18. Class structure and design specification for the sample Example Experiment demo included in Social Vision Toolkit. Within the Experiment Class, one Event Block is repeated three times; within the Event Block, the three tasks are executed in sequential order.

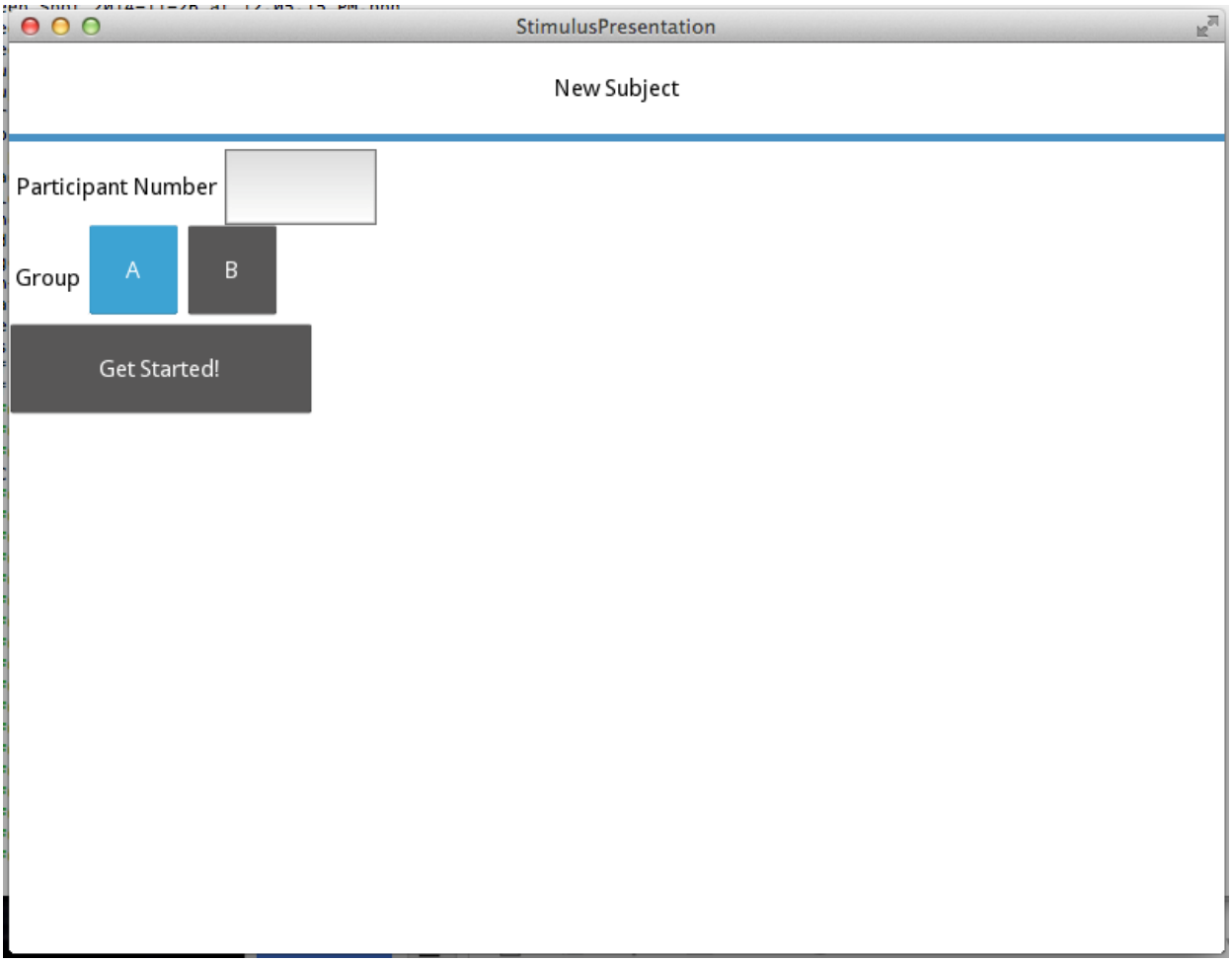


Figure 19. Social Vision Toolkit starting page. The starting page that appears after opening any experiment script developed using the Social Vision Toolkit. Researchers should type in the participant's number and select a condition group (if applicable) before clicking "Get Started!" Once the "Get Started!" button has been clicked, the experiment begins.

Appendix A: Study Set One Measures

Perceived Sexist Norms Scale

"Norms are a group's spoken or understood rules about member behavior; they explain how members ought to act. For example, in many classrooms, students understand the norm that they must raise their hands before they speak. After reading the story, please answer the following questions about the group's potential norms.

How important, if at all, are the following norms to the majority of the group?

	Extremely Unimportant To the Group		Unimportant to the Group		Important to the Group		Extremely Important to the Group
Members should treat women like they're good at math.*	1	2	3	4	5	6	7
Members should treat men differently than women.	1	2	3	4	5	6	7
It's acceptable to exclude some people.	1	2	3	4	5	6	7
Women have a lot to contribute to the group.*	1	2	3	4	5	6	7
No one should be made fun of.*	1	2	3	4	5	6	7
Group members should be friendly towards each other.*	1	2	3	4	5	6	7
Women should defer to men.	1	2	3	4	5	6	7
It is okay to ask for help.*	1	2	3	4	5	6	7

*Note: Reverse coded items are marked with **

Appendix B: Study Set Two Measures

Broader Impacts Scale

How likely is work by members of this group to improve science, technology, engineering, and mathematics education through mentoring their own subordinates or recruiting and training K-12 teachers?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is work by members of this group to improve the well-being of individuals in society?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is work by members of this group to encourage underrepresented individuals like racial minorities, women, and persons with disabilities to participate in science?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is work by members of this group to increase the economic competitiveness of the United States and contribute to a diverse, globally competitive science, technology, engineering, and mathematics workforce?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is work by members of this group to improve national security?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is work by members of this group to increase public engagement with science and technology, including scientific literacy?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely is the work by members of this group to be shared with members of Congress or used to inform public policy?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

Intellectual Merit Scale

How likely are members of this group to advance knowledge and understanding *within their own* field/industry?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely are members of this group to advance knowledge and understanding *across different* fields/industries?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely are members of this group to develop creative, original or potentially transformative concepts?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How well qualified do members of this team seem to do their work?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How well qualified do members of this team seem to present their work to others?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How likely are members of this group to transform the frontiers of knowledge?

1	2	3	4	5	6	7
Extremely unlikely						Extremely likely

How rigorous is the science done by members of this group?

1	2	3	4	5	6	7
Not at all rigorous						Extremely rigorous

Appendix C: Demo Code Included in the Social Vision Toolkit

Code for Example Experiment

```
from stimulus import Stimulus
from task import Task
from event import Event, wait
from experiment import Experiment
from ui import run_app

from kivy.uix.button import Button

class TestStimulus(SimpleTextStimulus):
 stimulus_id = "test_stimulus" # Identifies this stimulus in the results.
 duration = 1000    # Show this stimulus for 1000 ms.
 text = "Test"

class TestTask(Task):
 task_id = "test_task" # Identifies this task in the results.

 def render(self):
  self.layout.add_widget(Button(pos=self.layout.pos, text="I am a task. Press
any key to continue."))

class TestEvent(Event):
 event_id = "test_event" # Identifies this event in the results.
 stages = [TestStimulus, wait(1000), TestTask]

class TestExperiment(Experiment):
 events = [TestEvent] * 3

if __name__ == '__main__':
 run_app(TestExperiment)
```

Code for Between Subjects, Reverse Correlation Demo

```
from stimulus import Stimulus
from task import Task
from event import Event, wait
from experiment import Experiment
from ui import run_app
import util
from widgets.image_grids import RandomImageGrid, SequentialPairedImageGrid,
get_sequential_paired_image_grid_iterator
from widgets.simple_tasks import SimpleTextTask
from widgets.simple_stimuli import SimpleTextStimulus

import os

class MaleFemaleStimulus(Stimulus):
 stimulus_id = "image_grid_stimulus" # Identifies this stimulus in the
results.
 duration = 500     # Show this stimulus for 500 ms.
 def render(self):
  if self.subject.group == 'A':
   directories = {os.path.join(os.path.dirname(os.path.realpath(__file__)),
          "assets", "demo", "men"): 3,
       os.path.join(os.path.dirname(os.path.realpath(__file__)),
          "assets", "demo", "women"): 9}
  elif self.subject.group == 'B':
   directories = {os.path.join(os.path.dirname(os.path.realpath(__file__)),
          "assets", "demo", "men"): 9,
       os.path.join(os.path.dirname(os.path.realpath(__file__)),
          "assets", "demo", "women"): 3}
  else:
   raise ValueError("Expected subject.group to be one of A or B.")

  self.layout.add_widget(RandomImageGrid(directories,
            pos=self.layout.pos,
            size=self.layout.size,
            rows=3,
            cols=4))

# Read all the image pairs into a list, so that we can pass them to
# SequentialPairedImageGrid. This is necessary to do sampling without
# replacement.
image_pair_directory =
os.path.join(os.path.dirname(os.path.realpath(__file__)),
        "assets", "demo", "base_faces")
```

```
image_pair_list = get_sequential_paired_image_grid_iterator(
 ["_ori.jpg", "_inv.jpg"], image_pair_directory)
assert len(image_pair_list) == 15 # this must match number of paired images
in base_faces
image_pair_iterator = iter(image_pair_list)

class RevCorrTask(Task):
 task_id = "rev_corr"
 acceptable_keys = ['s', 'l']
 def render(self):
  self.image_grid = SequentialPairedImageGrid(image_pair_iterator,
             pos=self.layout.pos,
             size=self.layout.size)
  self.layout.add_widget(self.image_grid)
 def teardown(self):
  last_keypress = util.last_keypress(self.data)
  print "last keypress:", last_keypress
  if last_keypress == 's':
   self.data['filename'] = self.image_grid.image_filenames[0]
  elif last_keypress == 'l':
   self.data['filename'] = self.image_grid.image_filenames[1]

class IntroTask1(SimpleTextTask):
 task_id = "intro_text1"
 text = """
.. class:: center

Welcome to the lab!

For this study you will first see an array of faces. Please try to imagine
joining this group and how much you would belong there. After the array
image, you will be presented with two different faces, side-by-side.

Your task is to pick the face (out of the two) that best characterizes or
represents the average or mean face you saw in the array. That is, pick the
face that best represents all the faces from the array combined.

There will be many trials so please be patient and do your best to remain
alert. There will be a break every 5 trials.

Press any key to continue."""

class IntroTask2(SimpleTextTask):
 task_id = "intro_text2"
 text = """
```

```
You will use the S key to indicate the face on the left is the best
representation and the L key to indicate that the face on the right is the
best representation.

Please place your fingers on these keys now so that you can be ready to make
your judgments as soon as the task begins.

Press any key to begin the study."""

class ReadyFlash(SimpleTextStimulus):
 duration = 500
 stimulus_id = "ready_flash"
 text = "Ready?"

class FifteenSecondBreak(SimpleTextStimulus):
 duration = 15000
 stimulus_id = "fifteen_second_break"
 text = "This is a 15 second break. Please rest your eyes or stretch!"

class BreakOverDialog(SimpleTextTask):
 task_id = "break_over"
 text = """
The break has concluded.

Remember: you will use the S key to indicate the face on the left is the best
representation and the L key to indicate that the face on the right is the
best representation.

Please place your fingers on these keys now so that you can be ready to make
your judgments as soon as the task begins.

Press any key to continue."""

class ExperimentFinishedDialog(SimpleTextTask):
 task_id = "experiment_finished_dialog"
 text = """
Thank you! You have finished all of the judgment trials in this study. Before
you go, please tell us a bit about yourself.

Press any key to continue."""


class ChooseGender(SimpleTextTask):
 task_id = "gender"
 text = """Are you a man or woman?
```

```
Use the S (man) or L (woman) key to record your response.
"""
 acceptable_keys = ['s', 'l']

class ChooseEthnicity(SimpleTextTask):
 task_id = "ethnicity"
 text = """What is your race/ethnicity?
1 = Asian
2 = Black
3 = Latino
4 = White
5 = Multiracial / Other
"""
 acceptable_keys = ['1', '2', '3', '4', '5']

class PostExperimentQuestionsFinished(SimpleTextTask):
 task_id = "finished"
 text = """You are now finished with the experiment.
Please find your experimenter. Do not close this window.
"""
 acceptable_keys = ['p']

class IntroEvent(Event):
 event_id = "intro"
 stages = [IntroTask1, IntroTask2, ReadyFlash]

class MainEvent(Event):
 event_id = "main" # Identifies this event in the results.
 stages = [MaleFemaleStimulus, wait(500), RevCorrTask]

class BreakEvent(Event):
 event_id = "break"
 stages = [FifteenSecondBreak, BreakOverDialog]

class PostExperimentQuestionsEvent(Event):
 event_id = "questionnaire"
 stages = [ExperimentFinishedDialog, ChooseGender,
    ChooseEthnicity, PostExperimentQuestionsFinished]

class TestExperiment(Experiment):
 events = ([IntroEvent] +
    [MainEvent] * 5 + [BreakEvent] +
    [MainEvent] * 5 + [BreakEvent] +
    [MainEvent] * 5 + [BreakEvent]+ [PostExperimentQuestionsEvent])
```

```
 # For longer experiment with more trials, uncomment the below for faster
testing
 # events = ([IntroEvent] +
 #   [MainEvent] * 2 + [BreakEvent] +
 #   [MainEvent] * 2 + [PostExperimentQuestionsEvent])

if __name__ == '__main__':
 run_app(TestExperiment)
```

Code for Within Subjects, Ensemble Coding Demo

```
from stimulus import Stimulus
from task import Task
from event import Event, wait
from experiment import Experiment
from ui import run_app
import util
from widgets.image_grids import RandomImageGrid, CaptionedImage,
CenteredImage
from widgets.simple_tasks import SimpleTextTask
from widgets.simple_stimuli import SimpleTextStimulus
from functools import partial

import os
import random

BASE_DIR = os.path.dirname(os.path.realpath(__file__))

class FixationCross(Stimulus):
 stimulus_id = "fixation_cross"
 duration = 500
 def render(self):
  self.content = CenteredImage(os.path.join(BASE_DIR, 'assets', 'demo',
'fixation_cross.jpg'),
          pos=self.layout.pos,
          size=self.layout.size)
  self.layout.add_widget(self.content)

class TestTrialOne(Stimulus):
 stimulus_id = "TestTrialOne"
 duration = 500
 def render(self):
  self.content = CenteredImage(os.path.join(BASE_DIR, 'assets', 'demo',
'TestTrialOne.jpg'),
          pos=self.layout.pos,
          size=self.layout.size)
  self.layout.add_widget(self.content)

class TestTrialTwo(Stimulus):
 stimulus_id = "TestTrialTwo"
 duration = 500
 def render(self):
  self.content = CenteredImage(os.path.join(BASE_DIR, 'assets', 'demo',
'TestTrialTwo.jpg'),
```

```
            pos=self.layout.pos,
            size=self.layout.size)
  self.layout.add_widget(self.content)

class TestTrialDV(SimpleTextTask):
 task_id = "test_trial_dv"
 text = """
During a real trial, actual questions will be presented here.

Please press any key to continue.
"""

class PredefinedMaleFemaleStimulus(Stimulus):
 stimulus_id = "image_grid_stimulus" # Identifies this stimulus in the
results.
 duration = 500       # Show this stimulus for 500 ms.

 def __init__(self, num_male, num_female, subject, **kwargs):
  assert num_male + num_female == 12
  self.num_male = num_male
  self.num_female = num_female
  super(PredefinedMaleFemaleStimulus, self).__init__(subject, **kwargs)

 def render(self):
  directories = {os.path.join(BASE_DIR, "assets", "demo", "men"):
self.num_male,
      os.path.join(BASE_DIR, "assets", "demo", "women"): self.num_female}

  self.data['num_men'] = self.num_male
  self.layout.add_widget(RandomImageGrid(directories,
            pos=self.layout.pos,
            size=self.layout.size,
            rows=3,
            cols=4))

class StickRatio(Task):
 task_id = "stick_corr"
 acceptable_keys = ['1', '2', '3', '4', '5', '6', '7', '8', '9', '0', 'q',
'w', 'e']

 def render(self):
  self.content = CaptionedImage(os.path.join(BASE_DIR, 'assets', 'demo',
'ratio_measure.jpg'),
          'Which **PICTURE** best represents the ratio of men to women you
just saw?',
```

```
            pos=self.layout.pos,
            size=self.layout.size)
    self.layout.add_widget(self.content)

  def teardown(self):
    pass

class IntroTask1(SimpleTextTask):
  task_id = "intro_text1"
  text = """
.. class:: center

Welcome to the lab!

For this study you will see a series of pictures of groups of people.

**You will be shown each image VERY BRIEFLY.**

Please answer the questions following each picture as best you can.

Press any key to continue."""

class InstructFitBelongingness(SimpleTextTask):
  task_id = "intro_fit_belongingness"
  text = """
DIRECTIONS: In this block, you will be asked **how much you think you fit in
with the group** shown in each picture.

Please answer the questions following each picture as best you can.

Press any key to begin this block."""

class InstructFitBelongingnessEvent(Event):
  event_id = "InstructFitBelongingnessEvent"
  stages = [InstructFitBelongingness]

class InstructRatio(SimpleTextTask):
  task_id = "intro_ratio"
  text = """
DIRECTIONS: In this block, after each picture, you will be asked to choose
the stick figure drawing that **best represents the number of men and women
in the group just shown,** with men represented as blue stick figures and
women represented as pink stick figures.

Please answer the questions following each picture as best you can.
```

```
Press any key to begin this block."""

class InstructRatioEvent(Event):
 event_id = "InstructRatioEvent"
 stages = [InstructRatio]

class ReadyFlash(SimpleTextStimulus):
 duration = 500
 stimulus_id = "ready_flash"
 text = "Ready?"

class ExperimentFinishedDialog(SimpleTextTask):
 task_id = "experiment_finished_dialog"
 text = """
Thank you! You have finished all of the judgment trials in this study. Before
you go, please tell us a bit about yourself.

Press any key to continue."""

class ChooseGender(SimpleTextTask):
 task_id = "gender"
 text = """What is your gender?
Use the 1 (man), 2 (woman), or 3 (non-binary or other identification) key to
record your response.
"""
 acceptable_keys = ['1', '2', '3']

class ChooseEthnicity(SimpleTextTask):
 task_id = "ethnicity"
 text = """Which ethnicity category best describes you?
1 = Asian/Asian American
2 = Black/African/African American
3 = Latino/Hispanic American
4 = White/European American
5 = Native American
6 = Pacific Islander
7 = Multiracial
8 = Other
"""
 acceptable_keys = ['1', '2', '3', '4', '5', '6', '7', '8']



class PostExperimentQuestionsFinished(SimpleTextTask):
```

```
 task_id = "finished"
 text = """You are now finished with the experiment.

Please find your experimenter. Do not close this window.
"""
 acceptable_keys = ['p']

class GroupFitQuestion(Task):
 task_id = "group_fit"
 acceptable_keys = ['1', '2', '3', '4', '5', '6', '7', '8', '9']
 def render(self):
  self.content = CaptionedImage(os.path.join(BASE_DIR, 'assets', 'demo',
'FitDV.jpg'),
           ' ',
           pos=self.layout.pos,
           size=self.layout.size)
  self.layout.add_widget(self.content)

 def teardown(self):
  # TODO recode 'a' -> 1, etc.
  pass

class GroupBelongingQuestion(Task):
 task_id = "group_belonging"
 acceptable_keys = ['1', '2', '3', '4', '5', '6', '7', '8', '9']
 def render(self):
  self.content = CaptionedImage(os.path.join(BASE_DIR, 'assets', 'demo',
'BelongingDV.jpg'),
           ' ',
           pos=self.layout.pos,
           size=self.layout.size)
  self.layout.add_widget(self.content)

 def teardown(self):
  # TODO recode 'a' -> 1, etc.
  pass

class IntroEvent(Event):
 event_id = "intro"
 stages = [IntroTask1, ReadyFlash]




class RatioEvent(Event):
```

```
 event_id = "ratio" # Identifies this event in the results.

 def __init__(self, num_men, *args, **kwargs):
  self.stages = [FixationCross, partial(PredefinedMaleFemaleStimulus,
num_men, 12-num_men),
      wait(500), StickRatio]
  super(RatioEvent, self).__init__(*args, **kwargs)

class BelongingFitEvent(Event):
 event_id = "belonging_fit_event" # Identifies this event in the results.

 def __init__(self, num_men, *args, **kwargs):
  dvs = [GroupFitQuestion, GroupBelongingQuestion]
  random.shuffle(dvs)
  self.stages = [FixationCross, partial(PredefinedMaleFemaleStimulus,
num_men, 12-num_men),
      wait(500)] + dvs
  super(BelongingFitEvent, self).__init__(*args, **kwargs)

class PostExperimentQuestionsEvent(Event):
 event_id = "demographics"
 stages = [ExperimentFinishedDialog, ChooseGender,
    ChooseEthnicity, PostExperimentQuestionsFinished]

class IntroTestTrial(SimpleTextTask):
 task_id = "intro_test_trials"
 text = """
To get you prepared for the actual experiment, you will be completing three
test trials.
These are just to get you familiar with the keyboard and prompts. You will
see an image flash on screen before being asked a question about it.

Press any key to continue.
"""

class DoneTestTrial(SimpleTextTask):
 task_id = "completed_test_trials"
 text = '''
You have successfully completed the test trials. The actual experiment will
begin now.

Press any key to continue when you are ready.
 '''

class TestTrial(Event):
```

```
 event_id = "test_trial"
 stages = [IntroTestTrial, FixationCross, TestTrialOne, TestTrialDV,
FixationCross, TestTrialTwo, TestTrialDV, DoneTestTrial]

def make_events(intro, event, trials_by_num_men):
 """Takes a dictionary mapping number of men to number of trials.
 (e.g. {0: 40, 3: 40, 6: 40, 9: 40, 12: 40}. Returns a sequence of events
 that will contain one of each of the trials requested, in random order.
 For example, the dictionary above would result in 200 trials evenly split
 among 0, 3, 6, 9, and 12 men."""
 events = []
 for num_men, num_trials in trials_by_num_men.iteritems():
  events += [partial(event, num_men) for i in range(num_trials)]
 random.shuffle(events)
 return [intro] + events

class TestExperiment(Experiment):
 def __init__(self, *args, **kwargs):
  blocks = [make_events(InstructRatioEvent, RatioEvent, {0: 10, 3: 10, 6: 10,
9: 10, 12: 10}),
    make_events(InstructFitBelongingnessEvent, BelongingFitEvent, {0: 10, 3:
10, 6: 10, 9: 10, 12: 10}),
  random.shuffle(blocks)
  self.events = [IntroEvent] + [TestTrial] + blocks[0] + blocks[1] +
[PostExperimentQuestionsEvent]
  super(TestExperiment, self).__init__(*args, **kwargs)

if __name__ == '__main__':
 run_app(TestExperiment)
```

References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*(6), 1490–1528. https://doi.org/10.1177/0149206313478188

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world extracting statistical summary representations over time. *Psychological Science*, *21*(4), 560–567. https://doi.org/10.1177/0956797610363543

Alt, N. P., Goodale, B. M., Lick, D. J., & Johnson, K. L. (2017). Threat in the company of men: Ensemble perception and threat evaluations of groups varying in sex ratio. *Social Psychological and Personality Science*, 1–8. https://doi.org/10.1177/1948550617731498

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences*, *15*(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science*, *19*(4), 392–398. https://doi.org/10.1111/j.1467-9280.2008.02098.x.The

Alvarez, G. A., & Oliva, A. (2010). The representation of ensemble visual features outside the focus of attention. *Journal of Vision*, *7*(9), 129–129. https://doi.org/10.1167/7.9.129

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162. https://doi.org/10.1111/1467-9280.00327

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193. https://doi.org/10.1037/h0054663

Bai, Y., Leib, A. Y., Puri, A. M., Whitney, D., & Peng, K. (2015). Gender differences in crowd perception. *Frontiers in Psychology*, *6*(1300), 1–12.

https://doi.org/10.3389/fpsyg.2015.01300

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). Cambridge, MA: MIT Press. Retrieved from citeulike-article-id:838940

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/doi:10.18637/jss.v067.i01.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletinulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Baumeister, R., & Tice, D. (1990). Point-counterpoints: Anxiety and social exclusion. *Journal of Social and Clinical Psychology*, *9*(2), 165–195. https://doi.org/10.1521/jscp.1990.9.2.165

Beede, D., Julian, T., Langdon, D., McKittrick, G., Khan, B., & Doms, M. E. (2011). Women in STEM : A gender gap to innovation. *U.S. Department of Commerce, Economics and Statistics Administration*, 1–11. https://doi.org/10.2139/ssrn.1964782

Biederman, I., Glass, A. L., & Stacy, Jr., E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*(1), 22–27.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597.

Blischak, J. D., Davenport, E. R., & Wilson, G. (2016). A Quick Introduction to Version Control with Git and GitHub. *PLoS Computational Biology*, *12*(1), 1–18. https://doi.org/10.1371/journal.pcbi.1004668

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392. https://doi.org/10.1177/0956797610397956

Breiter, H. C., Etcoff, N. L., Whalen, P. J., Kennedy, W. A., Rauch, S. L., Buckner, R. L., … Rosen, B. R. (1996). Response and habituation of the human amygdala during visual processing of facial expression. *Neuron*, *17*, 875–887. https://doi.org/10.1016/S0896-6273(00)80219-6

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*. US: American Psychological Association. https://doi.org/10.1037/0033-2909.86.2.307

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology*, *41*(4), 656–670. https://doi.org/10.1037//0022-3514.41.4.656

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F., & Rosenkrantz, P. S. (1972). Sex-role stereotypes: A current appraisal. *Journal of Social Issues*, *28*(2), 59–79. https://doi.org/10.1111/j.1540-4560.1972.tb00018.x

Brown, V., Huey, D., & Findlay, J. M. (1997). No TitleFace detection in peripheral vision: Do faces pop out? *Perception*, *26*(12), 1555–1570.

Burr, D., & Ross, J. (2008). A visual sense of number. *Current Biology*, *18*(6), 425–428.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., &

Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cacioppo, J. T., & Hawkley, L. C. (2003). Social isolation and health, with an emphasis on underlying mechanisms. *Perspectives in Biology and Medicine*, *46*(3), S39–S52. https://doi.org/10.1353/pbm.2003.0063

Cacioppo, J. T., Hawkley, L. C., Crawford, L. E., Ernst, J. M., Burleson, M. H., Kowalewski, R. B., … Berntson, G. G. (2002). Loneliness and health: Potential mechanisms. *Psychosomatic Medicine*, *64*(3), 407–417. https://doi.org/10.1097/00006842-200205000-00005

Carpinella, C. M., Chen, J. M., Hamilton, D. L., & Johnson, K. L. (2015). Gendered facial cues influence race categorizations. *Personality and Social Psychology Bulletin*, *41*(3), 405–419. https://doi.org/10.1177/0146167214567153

Carretié, L., Hinojosa, J. A., & Mercado, F. (2003). Cerebral patterns of attentional habituation to emotional visual stimuli. *Psychophysiology*, *40*(3), 381–388. https://doi.org/10.1111/1469-8986.00041

Cedrus Corporation. (2017). Shop SuperLab Licenses. Retrieved January 15, 2018, from https://cedrus.com/store/superlab.html

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: how stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045–1060. https://doi.org/10.1037/a0016239

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual

displays. *Perception & Psychophysics*, *67*(1), 1–13. https://doi.org/10.3758/BF03195009

Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*, 891–900. https://doi.org/10.1016/j.visres.2004.10.004

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307–1310. https://doi.org/10.1126/science.1128317

Conte, H. R., Weiner, M. B., & Plutchik, R. (1982). Measuring death anxiety: Conceptual, psychometric, and factor-analytic aspects. *Journal of Personality and Social Psychology*, *43*(4), 775–785. https://doi.org/10.1037/0022-3514.43.4.775

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

de Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*(5), 789–794. https://doi.org/10.3758/PP.70.5.789

de Fockert, J. W., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, *62*(9), 1716–1722. https://doi.org/10.1080/17470210902811249

DeLongis, A., Folkman, S., & Lazarus, R. S. (1988). The impact of daily stress on health and mood: Psychological and social resources as mediators. *Journal of Personality and Social Psychology*, *54*(3), 486–495. https://doi.org/10.1037/0022-3514.54.3.486

Dotsch, R. (2015). rcicr: Reverse correlation classification toolbox.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571. https://doi.org/10.1177/1948550611430272

Dotsch, R., Wigboldus, D. H. J., Langner, O., & Van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, *19*(10), 978–980. https://doi.org/10.1111/j.1467-9280.2008.02186.x

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498. https://doi.org/10.1037/0033-295X.105.3.482

Feldman, D. C. (1984). The development and enforcement of group norms. *The Academy of Management Review*, *9*(1), 47–53.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Forbach, G. B., Stanners, R. F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition*, *2*(2), 337–339. https://doi.org/10.3758/BF03209005

Forsman, J. A., & Barth, J. M. (2017). The effect of occupational gender stereotypes on men's interest in female-dominated occupations. *Sex Roles*, *76*, 460–472. https://doi.org/10.1007/s11199-016-0673-3

Garner, W. (1978). Aspects of a stimulus: Features, dimensions, and configurations. In E. Rosch

& B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 99–133). Hillsdale, NJ: Erlbaum.

Garstka, T. A., Schmitt, M. T., Branscombe, N. R., & Hummert, M. L. (2004). How young and older adults differ in their responses to perceived age discrimination. *Psychology and Aging*, *19*(2), 326–335. https://doi.org/10.1037/0882-7974.19.2.326

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Ghavami, N., & Peplau, L. A. (2012). An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, *37*(1), 113–127. https://doi.org/10.1177/0361684312464203

Gombrich, E. (1950). *The Story of Art*. London, UK: Phaidon.

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*(4), 700–17. https://doi.org/10.1037/a0026659

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology. General*, *144*(2), 432–446. https://doi.org/10.1037/xge0000053

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), 751–753. https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009a). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology.* https://doi.org/10.1016/j.cub.2007.06.039

Haberman, J., & Whitney, D. (2009b). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(3), 718–34.

https://doi.org/10.1037/a0013899

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception & Psychophysics*, *72*(7), 1825–1838. https://doi.org/10.3758/APP.72.7.1825.The

Hackman, J. R. (1992). Group influences on individuals in organizations. In M. D. D. L. M. Hough (Ed.), *Handbook of industrial and organizational psychology, Vol. 3, 2nd ed* (pp. 199–267). Palo Alto, CA, US: Consulting Psychologists Press.

Hagerty, B. M., & Williams, A. (1999). The effects of sense of belonging, social support, conflict, and loneliness on depression. *Nursing Research*, *48*(4), 215–219. Retrieved from http://journals.lww.com/nursingresearchonline/Fulltext/1999/07000/The_Effects_of_Sense_of_Belonging,_Social_Support,.4.aspx

Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science*, *17*(7), 572–576.

Hamilton, D. L., Chen, J. M., Ko, D. M., Winczewski, L., Banerji, I., & Thurston, J. A. (2015). Sowing the seeds of stereotypes: Spontaneous inferences about groups. *Journal of Personality and Social Psychology*, *109*(4), 569–588. https://doi.org/10.1037/pspa0000034

Hawkley, L. C., & Cacioppo, J. T. (2003). Loneliness and pathways to disease. In *Brain, Behavior, and Immunity* (Vol. 17). https://doi.org/10.1016/S0889-1591(02)00073-9

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. *Quantitative methodology series* (Vol. 98). https://doi.org/10.1198/jasa.2003.s281

Hugenberg, K., & Wilson, J. P. (2013). Faces are central to social cognition. In D. E. Carlston (Ed.), *The Oxford Handbook of Social Cognition* (pp. 167–193). New York, NY: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199730018.013.0009

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception & Psychophysics*, *75*(2), 278–86. https://doi.org/10.3758/s13414-012-0399-4

Isaacowitz, D. M. (2006). Motivated gaze: The view from the gazer. *Current Directions in Psychological Science*, *15*(2), 68–72.

Itier, R. J., & Taylor, M. J. (2004). Effects of repetition learning on upright, inverted and contrast-reversed face processing using ERPs. *NeuroImage*, *21*(4), 1518–1532. https://doi.org/10.1016/j.neuroimage.2003.12.016

Johnson, K. L., Freeman, J. B., & Pauker, K. (2011). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*(1), 116–31. https://doi.org/10.1037/a0025335

Johnson, K. L., Iida, M., & Tassinary, L. G. (2012). Person (mis)perception: functionally biased sex categorization of bodies. *Proceedings of the Royal Society B: Biological Sciences*, *279*(October), 4982–4989. https://doi.org/10.1098/rspb.2012.2060

Jones, Z. M. (2013). Git/GitHub , transparency , and legitimacy in quantitative research. *The Political Methodologist*, 1–2.

Joo, S. J., Shin, K., Chong, S. C., & Blake, R. (2009). On the nature of the stimulus information necessary for estimating mean size of visual arrays. *Journal of Vision*, *9*(9), 1–12. https://doi.org/10.1167/9.9.7

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *17*(11), 4302–11. https://doi.org/10.1098/Rstb.2006.1934

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*, 498–525. https://doi.org/10.2307/1418556

Kenny, D. A., Kashy, D., & Bolger, N. (1998). Data analysis in social psychology. *Handbook of Social Psychology*, 233–265. https://doi.org/10.1002/pits.10035

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115–128. https://doi.org/10.1037/1082-989X.8.2.115

Kiecolt-Glaser, J. K., Fisher, L. D., Ogrocki, P., Stout, J. C., Speicher, C. E., & Glaser, R. (1987). Marital quality, marital disruption, and immune function. *Psychosomatic Medicine*, *49*(1), 13–34. https://doi.org/10.1097/00006842-198701000-00002

Kiecolt-Glaser, J. K., Glaser, R., Cacioppo, J. T., & Malarkey, W. B. (1998). Marital stress: Immunologic, neuroendocrine, and autonomic correlates. In *Annals of the New York Academy of Sciences* (Vol. 840, pp. 656–663). https://doi.org/10.1111/j.1749-6632.1998.tb09604.x

Kiecolt-Glaser, J. K., Ricker, D., George, J., Messick, G., Speicher, C. E., Garner, W., & Glaser, R. (1984). Urinary cortisol levels, cellular immunocompetency, and loneliness in psychiatric inpatients. *Psychosomatic Medicine*, *46*(1), 15–23.

Kimchi, R. (1983). *Perceptual organization of visual patterns*. University of California, Berkeley, 1982.

Kimchi, R. (1988). Selective attention to global and local eveles in teh comparison of hierarchical patterns. *Perception and Psychophysics*, *43*, 189–198.

Kimchi, R. (1990). Children's perceptual organization of hierarchical patterns. *European Journal of Cognitive Psychology*, *2*, 133–149.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological Bulletin*, *112*(1), 24–38. https://doi.org/10.1037/0033-2909.112.1.24

Kimchi, R., & Merhav, I. (1991). Hemispheric processing of global form, long form, and texture. *Acta Psychologica*, *76*, 133–147.

Kimchi, R., & Palmer, S. E. (1982). Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 521–535.

Kimchi, R., & Palmer, S. E. (1985). Separability and integrality of global and local levels of hierarchical patterns. *Journal of Experimental Psychology: Human Perception and Performance1*, *11*, 673–688.

Kirjakovski, A., & Matsumoto, E. (2016). Numerosity underestimation in sets with illusory contours. *Vision Research*, *122*, 34–42. https://doi.org/10.1016/j.visres.2016.03.005

Kivy. (2017). Changelog. Retrieved from https://kivy.org/#changelog

Kivy Organization. (2012). Kivy Documentation. Retrieved from https://media.readthedocs.org/pdf/kivy/latest/kivy.pdf

Kramer, R. S. S., Ritchie, K. L., & Burton, M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, *15*, 1–9. https://doi.org/10.1167/15.4.1.doi

Kreft, I., & de Leeuw, J. (1998). *Introducing multivlevel modeling*. (D. B. Wright, Ed.). London, UK: SAGE Publications.

Krogh, G. von, & Hippel, E. von. (2006). The promise of research on open source software. *Management Science*, *52*(7), 975–983. https://doi.org/10.1287/mnsc.1110.1374

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*(4), 418–444. https://doi.org/10.1177/0193841X9902300404

Krull, J. L., & MacKinnon, D. P. (2001). Testing predictive developmental hypotheses. *Multivariate Behavioral Research*, *36*(2), 227–248. https://doi.org/10.1207/S15327906MBR3602

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.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models. Retrieved from https://cran.r-project.org/package=lmerTest

Lahuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, *17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Leib, A. Y., Fischer, J., Liu, Y., Qiu, S., Robertson, L., & Whitney, D. (2014). Ensemble crowd perception : A viewpoint-invariant mechanism to represent average crowd identity. *Journal of Vision*, *14*(8), 1–13. https://doi.org/10.1167/14.8.26.doi

Levy, B. (1996). Improving memory in old age through implicit self-stereotyping. *Journal of Personality and Social Psychology*, *71*(6), 1092–1107. https://doi.org/10.1037/0022-3514.71.6.1092

Levy, B. (2009). Stereotype embodiment: A psychosocial approach to aging. *Current Directions in Psychological Science*, *18*(6), 332–336. https://doi.org/10.1111/j.1467-8721.2009.01662.x

Levy, B., Ashman, O., & Dror, I. (2000). To be or not to be: The effecs of age stereotypes on the will to live. *Omega - Journal of Death and Dying*, *40*(3), 409–420. https://doi.org/10.2190/Y2GE-BVYQ-NF0E-83VR

Lofland, L. H. (1982). Loss and human connection: An exploration into the nature of the social bond. In W. Ickes & E. S. Knowles (Eds.), *Personality, Roles, and Social Behavior* (pp. 219–242). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4613-9469-3_8

Los Angeles County Registrar. (2016). *Los Angeles County Election Results*. Los Angeles, CA. Retrieved from https://www.lavote.net/ElectionResults/Text/3496

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135. https://doi.org/10.3758/s13428-014-0532-5

Macrae, C. N., & Quadflieg, S. (2010). Perceiving people. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 428–463). New York, NY: Wiley.

Maher, B. A. (1992). A reader's, writer's, and reviewer's guide to assessing research reports in clinical psychology. *Methodological Issues & Strategies in Clinical Research.* Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10109-047

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*(1), 1–22.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209–226. https://doi.org/10.1016/j.cogsci.2003.11.004

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

McKone, E. (2004). Isolating the special component of face recognition: peripheral identification and a Mooney face. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 181.

Miller, L., & Budd, J. (1999). The development of occupational sex-role stereotypes, occupational preferences and academic subject preferences in children at ages 8, 12 and 16. *Educational Psychology*. United Kingdom: Taylor & Francis. https://doi.org/10.1080/0144341990190102

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience*, *9*(5), 555–604.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *1009*(41), 16474–16479.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception and Psychophysics*, *70*(5), 772–788.

Nagy, K., Zimmer, M., Greenlee, M. W., & Kovács, G. (2012). Neural correlates of after-effects caused by adaptation to multiple face displays. *Experimental Brain Research*, *220*(3–4), 261–275. https://doi.org/10.1007/s00221-012-3135-3

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*(3), 217–243. https://doi.org/10.1080/1047840X.2012.692215

Nothdurft, H. C. (1993). Faces and facial expressions do not pop out. *Perception*, *22*(11), 1287–1298. https://doi.org/10.1068/p221287

Olhausen, B. A., & Field, D. J. (1997). Sparse coding with an incomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, *41*, 176–210. https://doi.org/10.1006/cogp.1999.0728

Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of

the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *TRENDS in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Olshausen, B. A. (2003). Principles of image representation in visual cortex. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1603–1615). Cambridge, MA: MIT Press. Retrieved from https://groups.oist.jp/sites/default/files/img/ocnc/2004/Olshausen.pdf

Osborne, J. M., Bernabeu, M. O., Bruna, M., Calderhead, B., Cooper, J., Dalchau, N., … Deane, C. (2014). Ten simple rules for effective computational research. *PLoS Computational Biology*, *10*(3), 10–12. https://doi.org/10.1371/journal.pcbi.1003506

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744. https://doi.org/10.1038/89532

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*(January), 1–8. https://doi.org/10.3389/neuro.11.010.2008

Perez-Riverol, Y., Gatto, L., Wang, R., Sachsenberg, T., Uszkoreit, J., Leprevost, F. da V., … Vizcaíno, J. A. (2016). Ten simple rules for taking advantage of Git and GitHub. *PLoS Computational Biology*, *12*(7), 1–11. https://doi.org/10.1371/journal.pcbi.1004947

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Pfungst, O. (1911). *Clever Hans (the horse of Mr. Von Osten): A contribution to experimental*

animal and human psychology. New York, NY: Henry Holt and Company.

Phillips, L. T., Weisbuch, M., & Ambady, N. (2014). People perception: Social vision of groups and consequences for organizing and interacting. *Research in Organizational Behavior*, *34*, 101–127. https://doi.org/10.1016/j.riob.2014.10.001

Piazza, M., Mechelli, A., Butterworth, B., & Price, C. J. (2002). Are subitizing and counting implemented as separate or functionally overlapping processes? *NeuorImage*, *15*, 435–446. https://doi.org/10.1006/nimg.2001.0980

Pomerantz, J. R., & Pristach, E. A. (1989). Emergent features, attention, and perceptual glue in visual form perception. *Journal of Experimental Psychology: General*, *112*, 511–535.

Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to Be: The contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, *26*, 269–281. https://doi.org/10.1111/1471-6402.t01-1-00066

Pronin, E., Steele, C. M., & Ross, L. (2004). Identity bifurcation in response to stereotype threat: Women and mathematics. *Journal of Experimental Social Psychology*, *40*(2), 152–168. https://doi.org/10.1016/S0022-1031(03)00088-X

Psychology Software Tools. (2018). Licensing & Pricing. Retrieved January 15, 2018, from https://pstnet.com/products/e-prime/

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*(4), 615–630. https://doi.org/10.1037/0022-3514.94.4.615

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science.

Source Code for Biology and Medicine, *8*, 1–8. https://doi.org/10.1186/1751-0473-8-7

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, *106*(6), 897–911. https://doi.org/10.1037/a0036498

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods: Applications and Data Analysis Methods* (Second Ed., Vol. 1). Thousand Oaks, California: SAGE Publications.

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *38*(1972), 119–126.

Robbins, R., & McKone, E. (2003). Can holistic processing be learned for inverted faces? *Cognition*, *88*(1), 79–107.

Rock, I. (1986). The description and analysis of object and event perception. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 1–71). New York, NY: Wiley.

Romero, R., & Polich, J. (1996). P3(00) habituation from auditory and visual stimuli. *Physiology and Behavior*, *59*(3), 517–522. https://doi.org/10.1016/0031-9384(95)02099-3

Ross, J., & Burr, D. (2010). Vision senses number directly. *Journal of Vision*, *10*(2), 10.1-8. https://doi.org/10.1167/10.2.10

Schweinberger, S. R., & Neumann, M. F. (2016). Repetition effects in human ERPs to faces. *Cortex*, *80*, 141–153. https://doi.org/10.1016/j.cortex.2015.11.001

Seger, C. R., Smith, E. R., Kinias, Z., & Mackie, D. M. (2009). Knowing how they feel: Perceiving emotions felt by outgroups. *Journal of Experimental Social Psychology*, *45*(1),

80–89. https://doi.org/10.1016/j.jesp.2008.08.019

Sekuler, A. B., Gaspar, C. M., Gold, J. M., & Bennett, P. J. (2004). Inversion leads to quantitative, not qualitative, changes in face processing. *Current Biology*, *14*, 391–396. https://doi.org/10.1016/j.cub.2004.02.028

Shaw, M. E. (1981). *Group dynamics: The psychology of small group behavior*. (R. Robbin & J. R. Belser, Eds.) (3rd ed.). New York, NY: McGraw-Hill.

Shen-Miller, D., & Smiler, A. P. (2015). Men in female-dominated vocations: A rationale for academic study and introduction to the special issue. *Sex Roles*, *72*(7–8), 269–276. https://doi.org/10.1007/s11199-015-0471-3

Shnabel, N., Purdie-Vaughns, V., Cook, J. E., Garcia, J., & Cohen, G. L. (2013). Demystifying values-affirmation interventions: Writing about social belonging is a key to buffering against identity threat. *Personality and Social Psychology Bulletin*, *39*(5), 663–676. https://doi.org/10.1177/0146167213480816

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin and Review*, *5*(4), 644–649.

Snijders, T. a. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1570–1573). Wiley. https://doi.org/10.1002/0470013192.bsa492

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7–8), 509–528.

Stout, J. G., & Dasgupta, N. (2011). When he doesn't mean you : Gender- exclusive language as ostracism. *Personality and Social Psychology Bulletin*, *36*(6), 757–769.

https://doi.org/10.1177/0146167211406434

Sweeny, T. D., Haroz, S., & Whitney, D. (2013). Perceiving group behavior: Sensitive ensemble coding mechanisms for biological motion of human crowds. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 329–337. https://doi.org/10.1037/a0028712

Sweeny, T. D., & Whitney, D. (2014). Perceiving crowd attention: ensemble perception of a crowd's gaze. *Psychological Science*, *25*(10), 1903–13. https://doi.org/10.1177/0956797614544510

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology.*, *46*(2), 225–245. https://doi.org/10.1080/14640749308401045

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*(5), 583–592.

The National Science Foundation. (2016). Proposal and award policies and procedures guide. Retrieved from https://www.nsf.gov/pubs/policydocs/pappguide/nsf16001/gpg_print.pdf

Thornton, I. M., Srismith, D., Oxner, M., & Hayward, W. G. (2014). Estimating the racial composition of groups of faces : An ensemble other-race effect. In *TSPC2014* (pp. 76–78).

Thornton, I. M., & Vuong, Q. C. (2004). Incidental processing of biological motion. *Current Biology*, *14*(12), 1084–1089. https://doi.org/10.1016/j.cub.2004.06.025

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Torralba, A. (2003). Contextual priming for object detection. *International Journal of Computer*

Vision, *53*(2), 169–191.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412. https://doi.org/10.1088/0954-898X/14/3/302

Treisman, A. (1986). Properties, parts, and objects. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (p. 35: 1-70). New York, NY: Wiley.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 331–351. https://doi.org/10.1037/0096-1523.19.2.331

U.S. Department of Education. (2017). Bachelor's, master's, and doctor's degrees conferred by postsecondary institutions, by sex of student and discipline division: 2014-2015. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/programs/digest/d16/tables/dt16_318.30.asp

Uchino, B. N., Cacioppo, J. T., & Kiecolt-Glaser, J. K. (1996). The relationship between social support and physiological processes: A review with emphasis on underlying mechanisms and implications for health. *Psychological Bulletin*, *119*(3), 488–531. https://doi.org/10.1037/0033-2909.119.3.488

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2014). Bias correction for standardized effect size estimates used with single-subject experimental designs. *Journal of Experimental Education*, *82*(3), 358–374.

Vihinen, M. (2015). No more hidden solutions in bioinformatics. *Nature*, *521*(7552), 261. https://doi.org/10.1038/521261a

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*, 1447–1451.

Walton, G. M., Cohen, G. L., Cwir, D., & Spencer, S. J. (2012). Mere belonging: The power of social connections. *Journal of Personality and Social Psychology*, *102*(3), 513–532. https://doi.org/10.1037/a0025731

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Williams, D. W., & Sekuler, R. (1984). Coherent global motion percepts from stochastic local motions. *Vision Research*, *24*(1), 55–62. https://doi.org/10.1016/0042-6989(84)90144-5

Wood, W., & Eagly, A. H. (2010). Gender. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed., pp. 629–667). New York, NY: Wiley.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101. https://doi.org/10.1111/j.1467-7687.2005.00395.x