Artificial Aging and Face Morphing: Testing the Double Standard of Aging Hypothesis*

Eric Wilken† Mary Kite‡

June 21, 2026

Abstract

We introduce a method using artificial aging and face-morphing to test whether the perceived threshold between young and old differs by gender and race. Undergraduates ($N = 172$) viewed automatically advancing videos which morphed a young face to an artificially aged face by advancing smoothly through a discrete set of slides. Participants were asked to stop the video when the face appeared to cross from young to old. The faces featured four racial groups (White, Asian, Black, Latino) with both men and women. We summarize the results with two primary outcomes: the slide at which participants stopped the sequence and a participant-specific estimated transition age obtained by linearly interpolating between each participant's own age ratings of the young and old anchors. The two measures produce effects in opposing directions. Participants stopped the slide progression later for female faces across all racial groups, a pattern inconsistent with the double standard of aging. Estimated transition age, by contrast, was lower for female than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest, consistent with the double standard. The divergence traces to systematic differences in anchor ratings: young anchors of female morphs were rated as older than young anchors of male morphs. The gender effect on estimated age was not attenuated by adjustment for rated attractiveness of the aged face.

Keywords: ageism, age stereotypes, double standard of aging, face perception, AI

*The research design, hypotheses, outcome measures, and analytic plan were preregistered on the AsPredicted platform (AsPredicted #213,730). This study was approved under the exempt category by the Ball State University Institutional Review Board Protocol #2156654.

†Department of Economics, University of Oregon. Email: ewilken@uoregon.edu

‡Department of Counseling Psychology, Social Psychology, and Counseling, Ball State University

1 Introduction

Older adults are often perceived as warm but incompetent, a pattern that holds across cultures (Cuddy, Norton, and Fiske 2005; Chan et al. 2012). At the same time, aging is strongly associated with declines in perceived physical attractiveness, with older faces described using explicitly negative language (Henss 1991; Kite, Deaux, and Miele 1991). These perceptions matter. A large literature shows that physical attractiveness shapes social and economic outcomes, influencing everything from earnings to perceived competence (Hamermesh 2011; Mobius and Rosenblat 2006). Taken together, this suggests that age-based declines in attractiveness are not merely aesthetic, but may translate into meaningful social and economic disadvantages.

Aging is universally experienced, yet its consequences are profoundly influenced by gender. Older women experience more disadvantages than older men, both economically and socially. For example, they face higher rates of poverty and financial insecurity, reflecting the accumulation of gender disparities in earnings and career advancement (Calasanti and Slevin 2003). In addition, evidence from hiring suggests that older women are more likely to face discrimination, particularly as they approach retirement age, ostensibly because they are perceived as less competent or less desirable by employers (Neumark, Burn, and Button 2017). These disparities are not confined to the labor market. Older women also face disadvantages in marriage markets (England and McClintock 2009; Fileborn et al. 2015), are underrepresented or narrowly portrayed in media (Lemish and Muhlbauer 2012), and encounter biases in healthcare settings (Chrisler, Barney, and Palatino 2016). Sontag (1972) coined the phrase "double standard of aging" to describe these observations suggesting that men gain status and dignity with age, whereas women become perceived as less attractive and are consequently devalued.

Evidence for the double standard of aging is most consistent when individuals directly evaluate physical appearance. Across a range of studies, signs of aging such as wrinkles and gray hair are viewed as less attractive in women than in men (Berman, O'Nan, and Floyd

1981; Deutsch, Zalenski, and Clark 1986; Henss 1991; Harris 1994; Hummert, Garstka, and Shaner 1997). These differences extend beyond external evaluations. Women also report experiencing aging differently, frequently describing feelings of social invisibility and increased anxiety tied to societal standards of youthful appearance (Grabe, Ward, and Hyde 2008; Clarke and Korotchenko 2010; Ward and Holland 2011; Wilcox 1997). In response, many engage in appearance management, such as dyeing their hair, even when doing so conflicts with their sense of authenticity (Cecil et al. 2022). Taken together, these patterns suggest that the consequences of aging are shaped not only by age itself, but by when individuals come to be perceived as old. To the extent that this transition occurs earlier for women, the social and economic costs associated with aging are likely to be disproportionately borne by them.

We study the transition between adulthood and elderhood to offer new insights into the double standard of aging. Existing work shows that women are perceived as reaching middle and elder age sooner than men (Kogan 1979) and are thought to enter the prime of life at younger ages compared to men (Zepelin, Sills, and Heath 1986; Seccombe and Ishii-Kuntz 1991; Harris 1994). While researchers have tried to pin down a typical age at which an individual perceived as old, this goal is fraught with complications. For example, Wettstein et al. (2024) find that older adults place the onset of old age later in life, while younger individuals report earlier thresholds. Researchers that have used facial features to establish the elderhood threshold have discovered that raters are mixed in their accuracy when assessing age (Burt and Perrett 1995; Willner and Rowe 2001; Dehon and Brédart 2001; Christensen 2009; Norja 2021). Specifically, raters are most accurate for middle-aged faces, tend to underestimate older ages, and overestimate the age of young women. These studies use static photographs or surveys to estimate when individuals are perceived as old. They also typically used sets of photographs the researchers selected or had created for their study and so reflect the investigators' ideas about what representative older and younger people look like (Deutsch, Zalenski, and Clark 1991; Hummert, Garstka, and Shaner

1997). To address these limitations, we developed a new methodology; specifically, we take a more active approach by requiring participants to stop an automatically advancing video to identify when a target face "looks old." We call this approach "face morphing", and use it to compare how the perceived transition to elderhood occurs differentially by race and gender.

We find that when participants start with a young face ("target") they stop the video later for female targets versus male targets. Considered in isolation, this result is inconsistent with the double standard of aging. However, participants' ratings of the morph anchors (the static young and old starting points of each morph sequence) varied systematically by target gender and race: young anchors of female morphs were rated as older than young anchors of male morphs, and old-anchor ratings differed by gender and varied across racial groups. The morph sequences therefore did not span an equivalent perceived age range across faces within participants, and thus stop time cannot be interpreted as a direct measure of perceived age at the point of stopping.[1]

We use these anchor ages to interpolate a participant-specific estimated age for each face at every point in time during a morph sequence. Using this measure, female faces were judged old at a younger estimated age than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest. This pattern is consistent with the double standard of aging. The gender effect on estimated age was not attenuated by adjustment for rated attractiveness of the aged face, self-reported tiredness, or self-reported focus, in contrast to suggested pathways for the double-standard of aging in the literature. We report results for both stopping time and estimated transition age, and discuss the implications of their divergence for the interpretation of morph-based age-perception paradigms.

[1]While the heterogeneity here is less systematic, this also means that the morph sequences did not necessarily span an equivalent perceived age range across participants within faces, highlighting the value of a methodology where the perceived ages of each anchor are measured.

2 Background

People's beliefs about older adults reflect a combination of positive and negative characteristics; older adults are seen as warm, likable, and experienced, but they are also thought to be incompetent, rigid, and complaining (Cuddy and Fiske 2002; Rupp, Vodanovich, and Credé 2005; Narayan 2008). The view that older adults are "weak but wonderful" (Bugental and Hehman 2007, p. 184) emerges across a number of cultures, including both collectivist and individualist societies. For example, respondents from 26 countries reported that older adults are warmer, but less active, than younger people (Chan et al. 2012), and cross-cultural research shows that older adults are consistently viewed as warm but incompetent (Cuddy, Norton, and Fiske 2005). When people evaluate older adults' physical appearance, they are less generous. Older adults are generally rated as less attractive than younger adults (Canetto, Kaminski, and Felicio 1995; Henss 1991; Kissler and Bäuml 2000), and their appearance is described by terms such as wrinkled, hard of hearing, slow, and in decline (Kite, Deaux, and Miele 1991). Terms commonly associated with old age reinforce this perceived unattractiveness, including crone, fossil, withered, wizened, and wrinkled (Palmore 1999).

The negative consequences of being perceived as physically unattractive are well documented. A meta-analysis of 76 studies on the "what is beautiful is good" stereotype showed that, on average, physically unattractive people are viewed as less socially competent than physically attractive people (Eagly et al. 1991). Physically unattractive people also report fewer friends, poorer social skills, and a less active sex life (Rhodes, Simmons, and Peters 2005). In the labor market, attractive workers' performance is more highly rewarded than equivalent performance by unattractive workers (Mobius and Rosenblat 2006), and attractive workers earn a measurable wage premium over their less attractive counterparts (Hamermesh 2011; Scholz and Sicinski 2014). Notably, attractiveness is interconnected with age: unattractive faces are judged to be older than attractive ones (Wernick and Manaster 1984), and unattractive people in their 60s and beyond are judged to have less favorable personalities, fewer successful life experiences, and fewer occupational achievements than more attractive people of similar ages (Hummert, Garstka, and Shaner 1997; Johnson and Pittenger 1984). Because perceived attractiveness and perceived age are tightly coupled in raters' own judgments, any study of how people categorize faces as old has to take seriously the visual information on which that categorization is based.

2.1 The Double Standard of Aging

The cost of aging is not borne equally. Physical attractiveness is perceived to be more important for women than for men (Chrisler, Rossini, and Newton 2015), and Hamermesh (2011) found that judgments of women's appearance were more polarized than judgments of men's appearance. Women, more so than men, are expected to attend to their appearance by doing things such as wearing make-up and staying thin (Saucier 2004). If these expectations are met, women are seen as more attractive, competent, and likable than otherwise (Etcoff et al. 2011). Harris (1994) found that most visible signs of aging, such as wrinkles and gray hair, were viewed more negatively for women than men, while baldness was viewed as less attractive for men. These findings suggest that women who age might be judged particularly harshly. Sontag (1972) coined the phrase "double standard of aging" to refer to this asymmetry, arguing that women's success is measured by how they look while men's is measured by what they do, so that men gain status and dignity with age whereas women are devalued.

When people are asked to name the specific ages at which middle- and old-age begin, their responses line up with Sontag's account. Women are perceived to attain middle age and elder status sooner than men (Kogan 1979) and to reach the prime of life earlier (Harris 1994; Seccombe and Ishii-Kuntz 1991; Zepelin, Sills, and Heath 1986). A U.S. national survey found women were categorized as old at 55 compared to 60 for men (Seccombe and Ishii-Kuntz 1991). A more recent survey of older U.S. adults showed that male and female respondents agreed men were old at 70, but disagreed about the corresponding threshold for women. Men placed it at 68, a double standard favoring men, while women placed it at 75, a double standard favoring women (AARP 2014). The divergence between respondents is itself a reminder that these estimates of when elderhood begins are self-reported judgments, and may reflect the respondent's position as much as the target's.

Actual behavior is more consistent. In a Canadian sample, 67 percent of women reported using anti-aging products to maintain a youthful appearance even though most also reported that doing so was probably ineffective (Muise and Desmarais 2010), and in a qualitative study most of the 50- to 70-year-old Canadian women interviewed engaged in beauty work, such as make-up, hair color, and surgery, to push back against society's views of older women (Clark and Griffin 2008). Harris (1994) found that older women in the U.S. were much more likely than older men to engage in such behaviors. These are non-trivial investments of time and money, and they are consistent with women facing a steeper appearance-related penalty as they age. Evidence from media representation points the same direction. Older women are underrepresented in U.S. movies relative to older men (Bazzini et al. 1997; Lauzen and Dozier 2005), and when they do appear the roles skew white, high-income, and attractive (Lemish and Muhlbauer 2012). Older male actors appear more frequently, including in romantic roles, and are routinely paired with much younger actresses (Saucier 2004). Older women are likewise underrepresented in advertisements in Japan (Prieler et al. 2011), Great Britain (Kay and Furnham 2013), India (Raman et al. 2008), and the United States (Lee, Carpenter, and Meyers 2007; Raman et al. 2008).

2.2 Direct Tests of the Double Standard Hypothesis

Despite this indirect support, direct tests of the double standard have produced mixed results, and the pattern of that mix is informative. When photographs of actual faces are assessed, older women are usually judged more harshly than older men (Berman, O'Nan, and Floyd 1981; Deutsch, Zalenski, and Clark 1986; Foos and Clark 2011; Henss 1991; Hummert, Garstka, and Shaner 1997), though some studies find the effect only among younger raters (Foos and Clark 2011) or among men (Berman, O'Nan, and Floyd 1981; Henss 1991). When ratings are instead based on a verbal target label such as "75-year-old woman," rather than on a photograph, the double standard favoring men disappears and people rate women's and men's physical appearance similarly (Canetto, Kaminski, and Felicio 1995; Sherman 1986). In one case, the ratings even tilt in women's favor (Kite, Deaux, and Miele 1991). A similar visual-versus-verbal pattern appears on gender-associated traits. Photographs of older women are rated as less feminine than photographs of younger women while men's masculinity does not vary with perceived age (Deutsch, Zalenski, and Clark 1986), but when age is conveyed by label the gender differences attenuate or disappear (Canetto, Kaminski, and Felicio 1995; Kite, Deaux, and Miele 1991; Sherman 1986).

This gap between photograph-based and label-based findings is important for interpretation. It implies that the double standard is carried by the visual features of an aging face rather than by abstract category knowledge about older women and men, and it means that any credible test of whether women are judged old sooner than men requires visual rather than verbal stimuli. The existing photograph-based literature either compares different younger and older targets or uses personal photographs over time of the person. This introduces possible confounds. Specifically, the measured age effect could be due to between-target differences in identity, expression, and imaging conditions or it could be due to other cultural markers of age such as choice of clothing or hairstyle. Separating the effect of age from those other sources of variation requires stimuli in which a single face is observed across a range of apparent ages, holding everything else constant. Additionally, past studies have used photographs that represent distinct periods in a target's life (20s, 40s, 60s) which does not as accurately represent the continuous measure that is age.

We develop a method, described in Section 3, to address both of these concerns: artificial aging and face morphing. Participants view a continuous young-to-old morph of the same identity and stop the sequence at the point at which they judge the face to have crossed from young to old, permitting a within-identity estimate of the perceptual threshold for elderhood and a direct test of whether that threshold differs by target gender and across White, Asian, Black, and Latino faces.

3 Methods

3.1 Participants

Participants were a convenience sample of 172 students at a midsize Midwestern public university who took part in the research in partial fulfillment of an introductory psychology course requirement. The average age of participants was 20.9 years (SD = 2.58) and the sample was predominantly White, with 69.6% of participants identifying as such. In terms of gender, 75.6% of participants self-identified as woman, 11.6% as men, and 10.5% as non-binary, with 2.3% of participants not reporting their gender. On average, participants reported moderate levels of tiredness (M = 3.73, SD = 1.61) but with an ability to stay focused throughout the task (M = 3.46, SD = 1.37).

Table 1: Participant Characteristics by Condition

	Asian	Black	Latino	White
N	43	46	35	48
Age	20.5	21.5	20.3	21.0
	(1.04)	(4.42)	(1.19)	(1.73)
% White	0.674	0.682	0.735	0.702
% Man	0.093	0.130	0.143	0.104
% Woman	0.767	0.717	0.771	0.771
% Non-Binary	0.140	0.109	0.057	0.104
Tired	3.23	3.77	3.82	4.06
	(1.59)	(1.48)	(1.53)	(1.75)
Focus	3.60	3.30	3.50	3.45
	(1.50)	(1.32)	(1.31)	(1.38)

Notes: Standard deviations in parentheses.

3.2 Procedure

Data were collected in person in the laboratory. Upon arrival, participants were randomly assigned to a target race condition (White, Asian, Black, or Latino) and to a starting-point condition (young-to-old or old-to-young), determining which faces they would view and the direction in which the morph sequences would advance. Before beginning the experimental trials, participants completed a practice trial using a cat-to-dog morph sequence to familiarize themselves with the task of stopping a sequence at a predetermined threshold.

Participants then completed six morph sequences — three male and three female faces from their assigned racial group — presented in a random order. Each sequence advanced automatically at one frame per second. A researcher recorded the slide number at which the participant stopped each sequence. Following the morph task, participants completed a questionnaire containing the endpoint rating items, task experience items, and demographic questions described in the Measures section.

3.3 Measures

Morph Task

The primary outcome measures were collected via the face morphing task. Each morph sequence consisted of 50 images created by linearly interpolating between a young face and an artificially aged version of the same face, with each step incrementing the blend by 2%, such that the sequence spanned the full range from young to old appearance over 50 frames advancing at one frame per second. Young faces were drawn from the Chicago Face Database (Ma, Correll, and Wittenbrink 2015). Artificial aging was applied using the FaceApp application, which transforms a face to approximate an aged appearance; because FaceApp provides no numeric age metric of its own, the aged endpoint faces were pretested to confirm that participants perceived them as aged and to verify that attractiveness ratings were equivalent across young and old versions of each face. Morphing of the young and artificially aged endpoints was performed using the MorphMan application. Examples of the young and aged endpoints used in the task are shown in Figure 1.

Participants were shown six morph sequences — three male and three female faces, all of the same racial group to which they were assigned — presented in a random order. Half of participants were shown sequences that began with the young face and progressed toward the aged face; the remaining half were shown sequences beginning with the aged face and progressed toward the young face. Participants were instructed to stop the sequence at the point at which they judged the face to have crossed the threshold from young to old (or, in the reverse condition, from old to young). Prior to the experimental trials, participants completed a practice trial using a cat-to-dog morph sequence to familiarize themselves with the stopping procedure. A researcher recorded the slide number at which each participant stopped the sequence for each face. The slide number at which the participant stopped the sequence constitutes the first primary outcome measure, referred to throughout as *stop time*.

(a): Example of Young Face

(b): Example of "Old" Face

Figure 1: Examples of Face Morph End Points

Endpoint Age and Attractiveness Ratings

Following the morph task, participants were shown a static image of each morph endpoint — that is, the young anchor face and the aged anchor face for each sequence they had viewed — and asked to provide age and attractiveness ratings for each. Age was assessed with a single open-ended item: *"Estimate the exact age of this person (in years)."* Participants also indicated the perceived race/ethnicity and gender of each face. Attractiveness was assessed by asking participants to rate the person relative to others of the same race, age, and gender in the United States (e.g., relative to other 40-year-old males of a given ethnicity), providing a within-category comparison that controls for the confound between age and attractiveness that would arise from an absolute rating. The age ratings assigned to the young and old endpoints were subsequently used to construct a participant-specific estimate of the perceived age of the face at the point of stopping, via linear interpolation between each participant's own endpoint ratings (see Equation 2).

Task Experience

Two items assessed participants' subjective experience during the morph task. Focus was measured with the item *"How easy or difficult was it for you to focus on the faces as they changed during the study?"* and tiredness was measured with the item *"How tired did you feel during the study?"* Both items were included as covariates in sensitivity analyses to account for the possibility that variation in task engagement influenced stop time or perceived age judgments (see Section 4.4).[2]

Demographics

Following the task, participants completed a brief demographic questionnaire. Race/ethnicity was assessed with a multi-select item allowing participants to select all applicable categories or to self-describe. Gender was assessed with a multi-select item. Age was assessed with a single open-ended numeric item.

[2]Participants were also asked an open-ended item regarding their decision strategy: *"How did you determine when the faces became [old/young]?"* where the bracketed term reflected the participant's assigned condition. Responses to this item were not analyzed quantitatively but provide qualitative context for interpreting individual differences in stopping behavior.

4 Results

The experimental design comprised one main randomized between-subjects factor — the racial group of the target face (White, Asian, Black, Latino) — and one within-subjects factor — target gender (male, female). Young-to-old vs. old-to-young progression order was also randomized between-subjects for robustness. This combination of between- and within-subjects variation makes the study well-suited for mixed-design analysis of variance modeling (also termed a "split-plot ANOVA"), in which between- and within-subjects variance components are partitioned and allowed to have separate error terms.

All models were estimated as linear mixed-effects models via restricted maximum likelihood (REML) using the `lme4` package in R, with F-ratios and p-values derived from Satterthwaite-approximated degrees of freedom (Bates et al. 2015; Luke 2017). This specification is formally equivalent to a traditional mixed-design ANOVA while accommodating unbalanced cell sizes and partially missing data.[3] All models included two random effects: a participant random intercept and a participant-by-target gender random intercept-shifter.[4] Each experimental factor — target race, target gender, and their interaction — was tested using Type III sums of squares, such that each factor is evaluated after accounting for all remaining factors in the model. The effect sizes of these experimental conditions are reported as partial eta-squared $((\eta_p^2))$, which expresses the variance attributable to each factor as a proportion of the variance remaining after all other factors are accounted for, placing the between-subjects factor (target race) and within-subjects factor (target gender) on a comparable scale.

The decision to frame the analysis as a mixed-design ANOVA rather than as a regression model reflects the structure of the research question and norms of the existing literature on this topic rather than any substantive difference in the underlying model assumptions. The

[3]We re-estimate all the main models from this section using classical split-plot ANOVA in Appendix B.

[4]Rather than one intercept per participant, each participant gets a family of intercepts — one per target gender level — and the by-participant-by-target-gender term captures how much those shift relative to the participant's overall intercept.

two methods are cases of the general linear model

$$Y = XB + U \tag{1}$$

[5] and, with appropriate coding of the categorical explanatory variables, produce the same underlying model fit and equivalent test statistics. The approaches differ in what they emphasize as output. A regression specification emphasizes prediction. Its primary output is a set of coefficients that describe the direction and magnitude of each predictor's relationship with the outcome, representing the estimated difference between each non-reference level and the reference category, holding other predictors constant. The ANOVA framing, by contrast, emphasizes hypothesis testing. It partitions the total variance in the outcome into components attributable to each factor (target race, target gender, their interaction, and residual error) and uses an F-statistic to determine whether each component is larger than would be expected by chance. This produces a single test per factor, asking whether that factor accounts for meaningful variance in age evaluations across all its levels simultaneously, rather than reporting differences level by level. The individual group means, and the distances between them, are background quantities in this framework; they become the object of analysis only in subsequent post-hoc procedures such as Tukey's HSD or Scheffé's method, which conduct pairwise comparisons of the factor-level means. The ANOVA framing is, in our view, more readily interpretable given the categorical structure of the design.

4.1 Morph Stop Time

The first set of analyses examines whether the point in the morph sequence at which observers stopped the face — hereafter *stop time* — differed as a function of target race and gender. Models were estimated separately for participants who initiated the task from the young-face

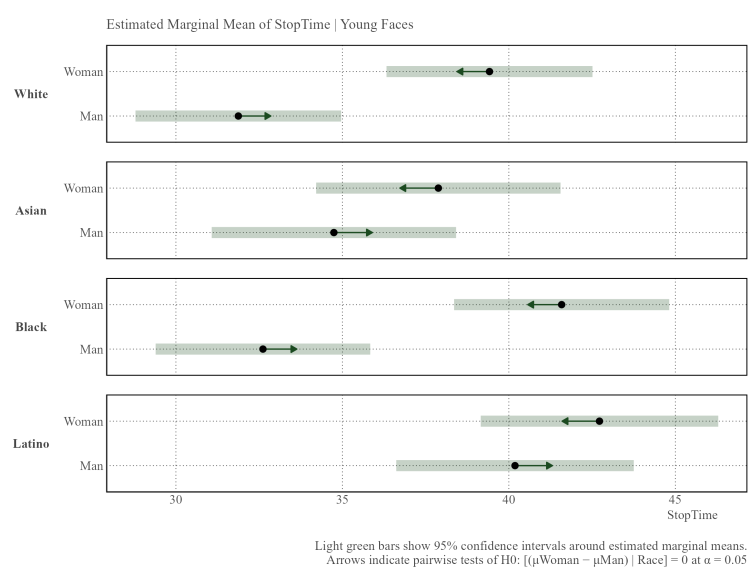
[5] Where $\mathbf{Y}$ is a matrix of multivariate measurements (each column representing a dependent variable), $\mathbf{X}$ is a design matrix of observations on independent variables (each column representing an independent variable), $\mathbf{B}$ is a matrix of parameters to be estimated, and $\mathbf{U}$ is a matrix of errors assumed to be uncorrelated across measurements and following a multivariate normal distribution, $\mathbf{U} \sim \mathcal{MN}(0, \Sigma)$.

endpoint and those who initiated from the old-face endpoint. Each sequence was 50 images.

We observe a large main effect of target gender ($F(1,\ 77.24) = 106.59$, $p < 0.001$, $\eta_p^2 = 0.58$), indicating that stop time differed substantially between male and female faces. This gender effect is heterogeneous by race, with a significant Race × Gender interaction ($F(3,\ 77.28) = 8.81$, $p < 0.001$, $\eta_p^2 = 0.25$), indicating that the magnitude of the gender difference in stop time varied across racial groups.

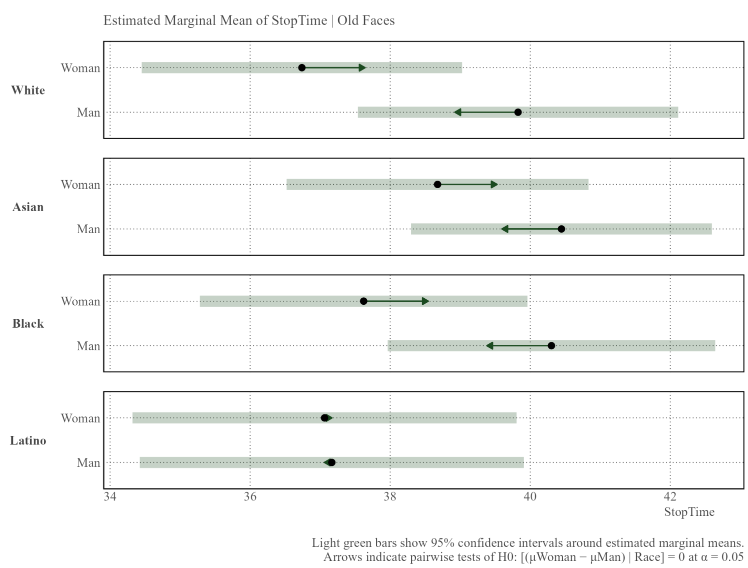
Figure 2 contains the point estimate of the marginal mean (dot), a 95% confidence interval (shaded region), and pairwise comparisons of the mean stop time of male versus female targets within a given target race using Tukey's HSD adjustment (arrows). The arrows represent a visual representation of these pairwise comparisons and can be interpreted such that any overlay implies we fail to reject the null hypothesis at the 5% level. The pairwise comparisons reveal that female faces were stopped later in the morph sequence (measured in terms of images, 1-50) than male faces across all racial groups and that this represents a statistically significant difference at the $\alpha = 0.05$ level. The gender gap was most pronounced for White faces (men: mean = 31.9, 95% CI [28.8, 35.0]; women: mean = 39.4, 95% CI [36.3, 42.5]; $\Delta = 7.5$) and Black faces (men: mean = 32.6, 95% CI [29.4, 35.8]; women: mean = 41.6, 95% CI [38.4, 44.8]; $\Delta = 9.0$).

Figure 2: Estimated Marginal Means of Stop Time By Condition



Among participants who initiated the task from the old-face endpoint, we again estimate a significant main effect of gender, though it is attenuated relative to the young-face condition ($F(1,\ 82.99) = 16.23$, $p < 0.001$, $\eta_p^2 = 0.16$). Neither the main effect of race ($F(3,\ 83.00) = 0.83$, $p = 0.481$, $\eta_p^2 = 0.03$) nor the Race × Gender interaction ($F(3,\ 83.00) = 1.69$, $p = 0.175$, $\eta_p^2 = 0.06$) are statistically significant at the 5% level. Figure 3 shows the estimated marginal means when participants see an old-face first. We see that male faces were stopped at a later point in the morph sequence than female faces across most racial groups (e.g., White men: $M = 39.8$, 95% CI [37.5, 42.1]; White women: $M = 36.7$, 95% CI [34.5, 39.0]). Since participants are now stopping the videos when they view the target as looking young, this implies the same result as the previous group: that women experience the transition between old and young age at a later point in the morph sequence than men.

Figure 3: Estimated Marginal Means of Stop Time By Condition



These findings are inconsistent with predictions derived from the double standard of aging (Sontag 1972; Kite, Deaux, and Miele 1991). Under that framework, female faces would be expected to cross the threshold for "old" appearance at an earlier point in the young-old morph sequence than male faces, and a later point in the old-young morph sequence, reflecting a harsher standard applied to visible aging in women. The present data demonstrate the opposite pattern: across all racial groups, observers permitted female (male) faces to advance further along the aging (de-aging) continuum before stopping the morph.

4.2 Stop Time Versus Perceived Age

While stop time provides a direct observation of the point in facial feature space at which observers judged a face to have crossed some threshold of "old", it is unclear exactly what it means for age. Stop time is a perfect proxy for target age only under the assumption all participants agree on how they perceive the age of the start and end faces within the morphing sequence.[6]

Figure 4: Distribution of Ratings of Anchor Faces for Morphs

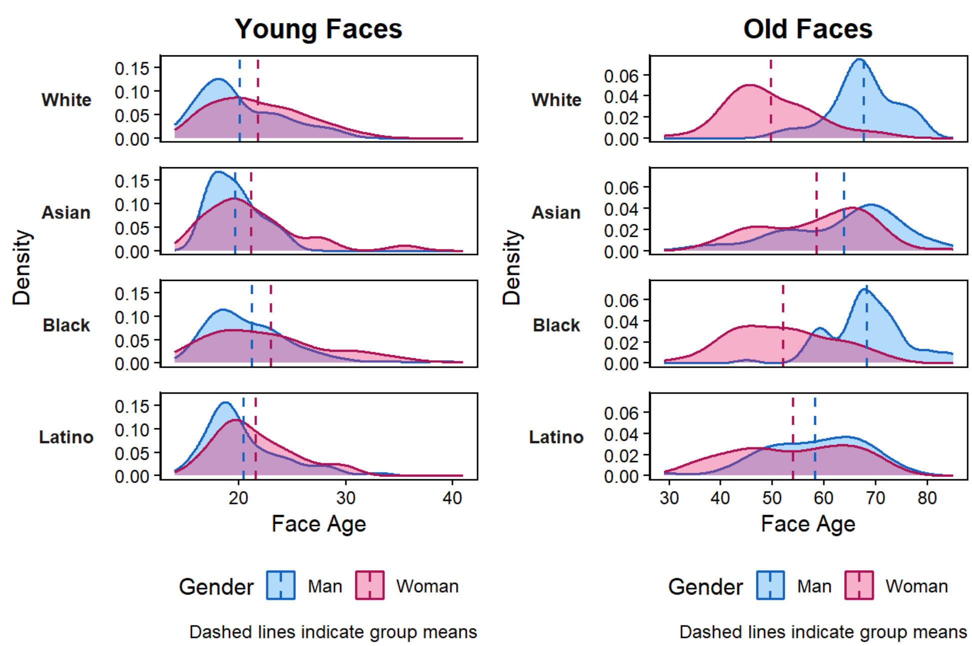


Figure 4 illustrates that this assumption does not hold in the present data. The left panel displays the distribution of age ratings assigned to the *young* anchor of each morph sequence—that is, how old participants judged the youngest face in the sequence to be—while the right panel displays the equivalent ratings for the *old* anchor. Critically, this variability is

[6]While this is a necessary condition, it is not sufficient. Sufficiency requires that all participants agree on the time-path of face ages throughout the morph. We explore this issues further in Appendix A.

not random noise: the distributions differ systematically as a function of the conditions under investigation, namely the gender and racial group of the target face. For the young anchor (left panel), female faces received consistently higher age ratings than male faces across all racial groups, indicating that participants perceived the youngest point of the female morph as older than the equivalent point of the male morph. For the old anchor (right panel), the distributions are far wider, spanning approximately 30 to 85 years, and the between-gender gap is considerably more pronounced and variable across racial groups—White faces in particular show marked separation, with male old anchors rated as substantially older than female old anchors.

This is a key observation for the interpretation of *stop time* as an outcome measure. If the anchor ages themselves vary as a function of target gender and race, then group differences in stop time may reflect, at least in part, differences in how the morph sequence was perceptually scaled rather than differences in the age at which observers judged a face to appear old. Disentangling these two sources of variance requires an outcome measure that accounts for each participant's own perception of the endpoints, which is the motivation for the participant-specific age estimation approach adopted in the following section.

4.3 Participant-Specific Estimated Target Age

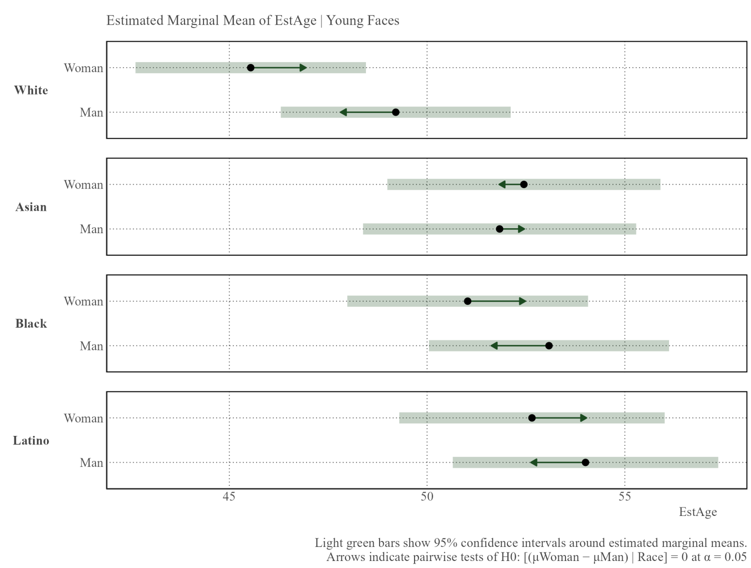
To obtain a measure of perceived age, we create a participant-specific estimate for each morph by interpolation between their ratings of the age of the faces which constitute the youngest and oldest within a morph. For a participant who assigned age y to the young endpoint and age o to the old endpoint of a given face, and who stopped the morph at image t, the estimated age at the point of stopping is computed as

$$\widehat{\text{Age}} = y + t\,(\frac{o-y}{50}) \tag{2}$$

Under this formulation, perceived age is modeled as advancing in equal increments across the full stimulus range.

Applying the same mixed-effects model structure to the estimated age outcome, significant main effects of both race ($F(3,\ 76.72) = 3.83$, $p = 0.013$, $\eta_p^2 = 0.13$) and gender ($F(1,\ 77.27) = 4.40$, $p = 0.039$, $\eta_p^2 = 0.05$) remain in the young-face-starting condition. However, Figure 5 shows that among the estimated marginal means, White women are estimated as youngest at the point of stopping (mean = 45.5, 95% CI [42.6, 48.5]) and Latino men as oldest (mean = 54.0, 95% CI [50.7, 57.4]). When age is estimated as described above, the only target gender difference that remains statistically significant is for white targets.

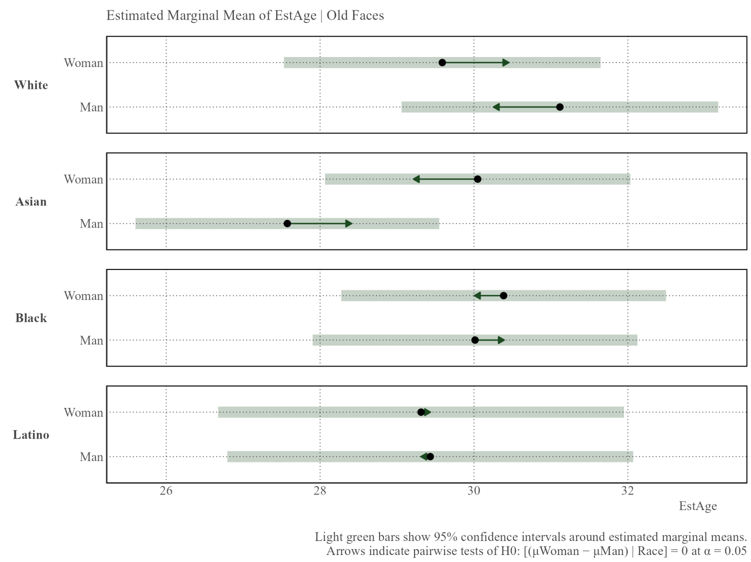
Figure 5: Estimated Marginal Means of Stop Time By Condition



In the old-face-starting condition, neither race ($F(3,\ 79.98) = 0.60$, $p = 0.620$, $\eta_p^2 = 0.02$) nor gender ($F(1,\ 412.28) = 0.40$, $p = 0.528$, $\eta_p^2 < 0.01$) featured significant main effects. However there is a significant Race × Gender interaction ($F(3,\ 412.30) = 3.66$, $p = 0.013$, $\eta_p^2 = 0.03$), reflecting divergence in estimated age across racial groups as a function of target gender. The Race × Gender interaction, as seen in Figure 6, is driven by Asian Man targets being rated as much younger when they reach the threshold between adulthood and elderhood than other races.[7]

[7]The old-face estimated age model converged with a singular fit, indicating that the variance of at

Figure 6: Estimated Marginal Means of Stop Time By Condition



4.4 Covariate-Adjusted Models

To assess whether the primary effects of target race and gender on *stop time* and $\widehat{\text{Age}}$ are robust to individual differences in task engagement and stimulus perception, we re-estimate each model with three participant-level covariates: rated attractiveness of the target face when old, self-reported tiredness at the time of participation, and self-reported focus during the task .

In the young-face-starting condition, the inclusion of covariates meaningfully alters the pattern of results relative to the unadjusted model for *stop time*. The main effect of target race is more precisely estimated with covariates ($F(3,\ 66.25) = 3.39$, $p = 0.023$, $\eta_p^2 = 0.13$). The main effect of gender remains large and statistically significant ($F(1,\ 78.80) = 98.70$, $p < 0.001$, $\eta_p^2 = 0.56$), as does the Race $\times$ Gender interaction ($F(3,\ 78.04) = 9.40$, $p < 0.001$, $\eta_p^2 = 0.27$). The adjusted estimated marginal means indicate that the gender gap in *stop time* persisted across all racial groups, with the largest disparities observed for Black faces (men: mean = 34.9, 95% CI [30.4, 39.3]; women: mean = 44.1, 95% CI [39.7, 48.6]; $\Delta = 9.2$) and Latino faces (men: mean = 42.8, 95% CI [38.0, 47.6]; women: mean = 45.2, 95% CI

least one random effect was estimated at the boundary of the parameter space. Estimates are reported for completeness but should be interpreted with caution.

[40.4, 50.0]; $\Delta = 2.4$). Notably, the covariate-adjusted means for Latino men shift upward relative to the unadjusted estimates, accounting for the emergence of a significant race main effect. In the old-face-starting condition, the effects are unchanged from the model without covariates.

When considering how the inclusion of covariates impacted group differences in $\widehat{\text{Age}}$, the effects are largely preserved. The main effect of race remains statistically significant ($F(3,\ 64.99) = 4.56$, $p = 0.006$, $\eta_p^2 = 0.17$) and is somewhat larger than in the unadjusted model ($\eta_p^2 = 0.13$), and the main effect of gender remains statistically significant as well ($F(1,\ 78.53) = 3.98$, $p = 0.049$, $\eta_p^2 = 0.05$). The adjusted estimated marginal means reveal that White women were assigned the youngest $\widehat{\text{Age}}$ at the point of stopping (mean = 47.4, 95% CI [43.5, 51.4]) while Black men and Latino men were assigned the oldest (mean = 56.4, 95% CI [52.2, 60.6] and mean = 56.4, 95% CI [51.9, 60.9], respectively). Compared with the unadjusted model, the covariate-adjusted means shifted upward across all cells, for White men in particular, as a results of adjustment for attractiveness and tiredness. As with *stop time*, the old-face-starting condition results for $\widehat{\text{Age}}$ were unchanged by the inclusion of covariates.

Across all four models, controlling for face attractiveness did not attenuate the gender effect on either stop time or estimated age. Effect sizes for the gender main effect are virtually unchanged before and after adjustment, with no reduction in precision. These findings are inconsistent with the theoretical prediction that gender differences in perceived aging are driven by the attractiveness penalty associated with aging in women. Instead, the results suggest that observers' age-related judgments of female faces reflect something distinct from attractiveness evaluations, a pattern that complicates straightforward attractiveness-mediation accounts of the double standard of aging.

5 Conclusion

This study examines the perceived threshold between young and old across gender and race using a face-morphing paradigm with two outcome measures: stop time, the morph position at which participants judged the face to have become old, and estimated age, a participant-specific measure derived by linear interpolation between each participant's ratings of the morph anchors. The two measures produced effects in opposing directions. In the young-face-starting condition, stop time was later for female than male faces across all four racial groups, whereas estimated age was lower for female than male faces, with White women judged old at the youngest estimated age and Latino and Black men at the oldest. The estimated-age pattern is consistent with the double standard of aging (Sontag 1972; Kite, Deaux, and Miele 1991; Seccombe and Ishii-Kuntz 1991) and with prior evidence from verbal age judgments (Seccombe and Ishii-Kuntz 1991; AARP 2014) and evaluative ratings of photographs (Berman, O'Nan, and Floyd 1981; Deutsch, Zalenski, and Clark 1986; Henss 1991).

The divergence between the two measures is attributable to systematic variation in participant ratings of the morph anchors. Young anchors of female morphs were rated as older than young anchors of male morphs, and old-anchor ratings varied by gender and race. Consequently, the morph sequence did not span an equivalent age range across conditions: a later stop time on a female morph could correspond to a lower interpolated age than an earlier stop time on a male morph. The source of this variation in anchor ratings cannot be identified from the present data and may reflect properties of the source photographs, the aging transformation applied by the stimulus-generation software, the perception of identical images under different target attributions, or some combination of these.

Adjustment for rated attractiveness of the aged face did not attenuate the gender effect on estimated age. This result is inconsistent with a strong form of the hypothesis that gendered ageism in age categorization is mediated by an attractiveness penalty (Chrisler, Rossini, and Newton 2015; Harris 1994). Two interpretations are compatible with the data:

perceptual age categorization and attractiveness evaluation may be partially dissociable components of gendered ageism, or the single post-task attractiveness item employed here may be insufficiently sensitive to detect a mediating relationship. Designs incorporating more extensive attractiveness measurement, or experimental manipulation of attractiveness, would be required to distinguish these accounts.

Several limitations constrain the generalizability of these findings. The sample was drawn from a single Midwestern university and was predominantly composed of young women; prior work indicates that the magnitude and direction of double-standard effects can vary with observer age and gender (Berman, O'Nan, and Floyd 1981; Foos and Clark 2011; AARP 2014). The stimulus set comprised three male and three female faces per racial group, which does not rule out stimulus-specific contributions to the observed group-level effects. The aged endpoints were generated by a commercial application whose underlying aging model is not publicly documented and may encode gender-specific aging trajectories; the anchor-rating data are consistent with this possibility but do not establish it. Finally, the old-face-starting condition produced smaller effects, which may not be robust to broader assumptions about the functional form of the age-morph mapping, as documented in Appendix A. Future work could consider whether reverse aging stimuli could serve to inoculate individuals against biases in perceived aging.

References

AARP (2014). "You're old, I'm not: How Americans really feel about aging". In: *AARP: The Magazine* 40, pp. 40–43.

Bates, Douglas et al. (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bazzini, D. G. et al. (1997). "The aging woman in popular film: Underrepresented, unattractive, unfriendly, and unintelligent". In: *Sex Roles* 36, pp. 531–543. DOI: 10.1007/BF02766689.

Berman, P. W., B. A. O'Nan, and W. Floyd (1981). "The double standard of aging and the social situation: Judgments of attractiveness of the middle-aged woman". In: *Sex Roles* 7, pp. 87–96. DOI: 10.1007/BF00287795.

Bugental, Daphne B. and Jessica A. Hehman (2007). "Ageism: A Review of Research and Policy Implications". In: *Social Issues and Policy Review* 1.1, pp. 173–216. DOI: 10.1111/j.1751-2409.2007.00007.x.

Burt, D. Michael and David I. Perrett (1995). "Perception of Age in Adult Caucasian Male Faces: Computer Graphic Manipulation of Shape and Colour Information". In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 259.1355, pp. 137–143. DOI: 10.1098/rspb.1995.0021.

Calasanti, T. M. and K. F. Slevin (2003). *Gender, Social Inequalities, and Aging*. AltaMira Press.

Canetto, S. S., P. L. Kaminski, and D. M. Felicio (1995). "Typical and optimal aging in women and men: Is there a double standard?" In: *International Journal of Aging and Human Development* 40, pp. 187–207. DOI: 10.2190/RX0U-T56B-1G0F-266U.

Cecil, Vanessa et al. (Mar. 2022). "Gendered ageism and gray hair: must older women choose between feeling authentic and looking competent?" en. In: *Journal of Women & Aging* 34.2, pp. 210–225. ISSN: 0895-2841, 1540-7322. DOI: 10.1080/08952841.2021.1899744. URL: https://www.tandfonline.com/doi/full/10.1080/08952841.2021.1899744 (visited on 05/15/2025).

Chan, W. et al. (2012). "Stereotypes of age differences in personality traits: Universal and accurate?" In: *Journal of Personality and Social Psychology* 103, pp. 1050–1066. DOI: 10.1037/a0029712.

Chrisler, J. C., A. Barney, and B. Palatino (2016). "Ageism and women's health". In: *Journal of Social Issues* 72, pp. 86–104.

Chrisler, J. C., M. Rossini, and J. R. Newton (2015). "Older women, power, and the body". In: *Women and Aging: An International, Intersectional Power Perspective*. Ed. by V. Muhlbauer, J. C. Chrisler, and F. L. Denmark. Cham, Switzerland: Springer International Publishing, pp. 9–30.

Christensen, K. et al. (2009). "Perceived age as biomarker". In: *BMJ* 339, b5262.

Clark, L. H. and M. Griffin (2008). "Visible and invisible ageing: Beauty work as a response to ageism". In: *Ageing & Society* 28, pp. 653–674. DOI: 10.1017/S0144686X07007003.

Clarke, L. H. and A. Korotchenko (2010). "Shades of grey". In: *Ageing and Society* 30, pp. 1011–1026.

Cuddy, A. J. C. and S. T. Fiske (2002). "Doddering but dear: Process, content, and function in stereotyping of older persons". In: *Ageism: Stereotyping and Prejudice against Older Persons*. Ed. by T. D. Nelson. Cambridge, MA: MIT Press, pp. 3–26.

Cuddy, A. J. C., M. I. Norton, and S. T. Fiske (2005). "This old stereotype: The pervasiveness and persistence of the elderly stereotype". In: *Journal of Social Issues* 61, pp. 267–285. DOI: 10.1111/j.1540-4560.2005.00405.x.

Dehon, Hedwige and Serge Brédart (2001). "An 'Other-Race' Effect in Age Estimation from Faces". In: *Perception* 30.9, pp. 1107–1113. DOI: 10.1068/p3122.

Deutsch, F. M., C. M. Zalenski, and M. E. Clark (1986). "Is there a double standard of aging?" In: *Journal of Applied Social Psychology* 16, pp. 771–785. DOI: 10.1111/j.1559-1816.1986.tb01167.x.

— (1991). "Is there a double standard of aging?" In: *Journal of Applied Social Psychology* 21.13, pp. 1096–1108.

Eagly, A. H. et al. (1991). "What is beautiful is good, but ...: A meta-analytic review of research on the physical attractiveness stereotype". In: *Psychological Bulletin* 110, pp. 109–128. DOI: 10.1037/0033-2909.110.1.109.

England, P. and E. McClintock (2009). "The gendered double standard of aging in U.S. marriage markets". In: *Population and Development Review* 35, pp. 797–816. DOI: 10.1111/j.1728-4457.2009.00309.x.

Etcoff, N. L. et al. (2011). "Cosmetics as a feature of the extended human phenotype: Modulation of the perception of biologically important facial signals". In: *PLoS ONE* 6, pp. 1–9. DOI: 10.1371/journal.pone.0025656.

Fileborn, Bianca et al. (Jan. 2015). "Sex, desire and pleasure: considering the experiences of older Australian women". en. In: *Sexual and Relationship Therapy* 30.1, pp. 117–130. ISSN: 1468-1994, 1468-1749. DOI: 10.1080/14681994.2014.936722. URL: http://www.tandfonline.com/doi/abs/10.1080/14681994.2014.936722 (visited on 05/15/2025).

Foos, P. W. and M. C. Clark (2011). "Adult age and gender differences in perceptions of facial attractiveness: Beauty is in the eye of the older beholder". In: *Journal of Genetic Psychology* 172, pp. 162–175. DOI: 10.1080/00221325.2010.526154.

Grabe, S., L. M. Ward, and J. S. Hyde (2008). "The role of the media in body image concerns". In: *Psychological Bulletin* 134, pp. 460–476.

Hamermesh, D. S. (2011). *Beauty Pays: Why Attractive People Are More Successful*. Princeton, NJ: Princeton University Press.

Harris, M. B. (1994). "Growing old gracefully: Age concealment and gender". In: *Journal of Gerontology: Psychological Sciences* 49, P149–P158. DOI: 10.1093/geronj/49.4.P149.

Henss, R. (1991). "Perceiving age and attractiveness in facial photographs". In: *Journal of Applied Social Psychology* 21.11, pp. 933–946. DOI: 10.1111/j.1559-1816.1991.tb00451.x.

Hummert, M. L., T. A. Garstka, and J. L. Shaner (1997). "Stereotyping of older adults: The role of target facial cues and perceiver characteristics". In: *Psychology and Aging* 12, pp. 107–114. DOI: 10.1037/0882-7974.12.1.107.

Johnson, D. F. and J. B. Pittenger (1984). "Attribution, the attractiveness stereotype, and the elderly". In: *Developmental Psychology* 20, pp. 1168–1172. DOI: 10.1037/0012-1649.20.6.1168.

Kay, A. and A. Furnham (2013). "Age and sex stereotypes in British television advertisements". In: *Psychology of Popular Media Culture* 2.3, pp. 171–186. DOI: 10.1037/a0033083.

Kissler, J. and K.-H. Bäuml (2000). "Effects of beholder's age on the perception of facial attractiveness". In: *Acta Psychologica* 104, pp. 145–166. DOI: 10.1016/S0001-6918(00)00018-4.

Kite, M. E., K. Deaux, and M. Miele (1991). "Stereotypes of young and old: Does age outweigh gender?" In: *Psychology and Aging* 6, pp. 19–27. DOI: 10.1037/0882-7974.6.1.19.

Kogan, N. (1979). "A study of age categorization". In: *Journal of Gerontology* 34, pp. 358–367. DOI: 10.1093/geronj/34.3.358.

Lauzen, M. and D. Dozier (2005). “Maintaining the double standard: Portrayals of age and gender in popular films”. In: *Sex Roles* 52.7-8, pp. 437–446. DOI: 10.1007/s11199-005-3710-1.

Lee, M. M., B. Carpenter, and L. S. Meyers (2007). “Representations of older adults in television advertisements”. In: *Journal of Aging Studies* 21, pp. 23–30. DOI: 10.1016/j.jaging.2006.04.001.

Lemish, D. and V. Muhlbauer (2012). ““Can’t have it all”: Representations of older women in popular culture”. In: *Women & Therapy* 35.3-4, pp. 165–180. DOI: 10.1080/02703149.2012.684541.

Luke, Steven G. (2017). “Evaluating Significance in Linear Mixed-Effects Models in R”. In: *Behaviour Research Methods* 49.4, pp. 1494–1502. DOI: 10.3758/s13428-016-0809-y.

Ma, Debbie S., Joshua Correll, and Bernd Wittenbrink (2015). “The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data”. In: *Behavior Research Methods* 47.4, pp. 1122–1135. DOI: 10.3758/s13428-014-0532-5.

Mobius, M. M. and T. S. Rosenblat (2006). “Why beauty matters”. In: *American Economic Review* 96, pp. 222–235. DOI: 10.1257/000282806776157515.

Muise, A. and S. Desmarais (2010). “Women’s perceptions and use of “anti-aging” products”. In: *Sex Roles* 63, pp. 126–137. DOI: 10.1007/s11199-010-9791-5.

Narayan, C. (2008). “Is there a double standard of aging? Older men and women and ageism”. In: *Educational Gerontology* 34.9, pp. 782–787. DOI: 10.1080/03601270802042123.

Neumark, D., I. Burn, and P. Button (2017). “Age discrimination and hiring of older workers”. In: *FRBSF Economic Letter*.

Norja, R. et al. (2021). “Accuracy of assessing age from faces”. In: *Nordic Psychology* 74, pp. 70–85.

Palmore, E. B. (1999). *Ageism: Negative and Positive*. New York: Springer.

Prieler, M. et al. (2011). “Gender representation of older people in Japanese television advertisments”. In: *Sex Roles* 64, pp. 405–415. DOI: 10.1007/s11199-010-9923-y.

Raman, P. et al. (2008). “Portrayals of older adults in U.S. and Indian magazine advertisements: A cross-cultural comparison”. In: *Howard Journal of Communications* 19.3, pp. 221–240. DOI: 10.1080/10646170802218214.

Rhodes, G., L. W. Simmons, and M. Peters (2005). “Attractiveness and sexual behavior: Does attractiveness enhance mating success?” In: *Evolution and Human Behavior* 26, pp. 186–201. DOI: 10.1016/j.evolhumbehav.2004.08.014.

Rupp, Deborah E., Stephen J. Vodanovich, and Marcus Credé (2005). “The Multidimensional Nature of Ageism: Construct Validity and Group Differences”. In: *Journal of Social Psychology* 145.3, pp. 335–362. DOI: 10.3200/SOCP.145.3.335-362.

Saucier, M. G. (2004). “Midlife and beyond: Issues for aging women”. In: *Journal of Counseling & Development* 82, pp. 420–425.

Scholz, J. K. and K. Sicinski (2014). “Facial attractiveness and lifetime earnings: Evidence from a cohort study”. In: *Review of Economics and Statistics* 97, pp. 14–28. DOI: 10.1162/REST_a_00435.

Seccombe, K. and M. Ishii-Kuntz (1991). “Perceptions of problems associated with aging: Comparisons among four older age cohorts”. In: *The Gerontologist* 31, pp. 527–533. DOI: 10.1093/geront/31.4.527.

Sherman, S. R. (1986). “Sex role stereotypes of middle and old age”. In: *Journal of Social Service Research* 8, pp. 23–35. DOI: 10.1300/J079v08n04_02.

Sontag, S. (1972). *The double standard of aging.* Saturday Review of the Society.

Ward, R. and C. Holland (2011). “Hair and later-life image dilemmas”. In: *Ageing and Society* 31, pp. 288–307.

Wernick, M. and G. J. Manaster (1984). “Age and the perception of age and attractiveness”. In: *The Gerontologist* 24, pp. 408–414. DOI: 10.1093/geront/24.4.408.

Wettstein, Markus et al. (2024). “Postponing Old Age: Evidence for Historical Change Toward a Later Perceived Onset of Old Age”. In: *Psychology and Aging* 39.6, pp. 625–636. DOI: 10.1037/pag0000812.

Wilcox, S. (1997). “Age and gender in relation to body attitudes”. In: *Psychology of Women Quarterly* 21, pp. 549–565.

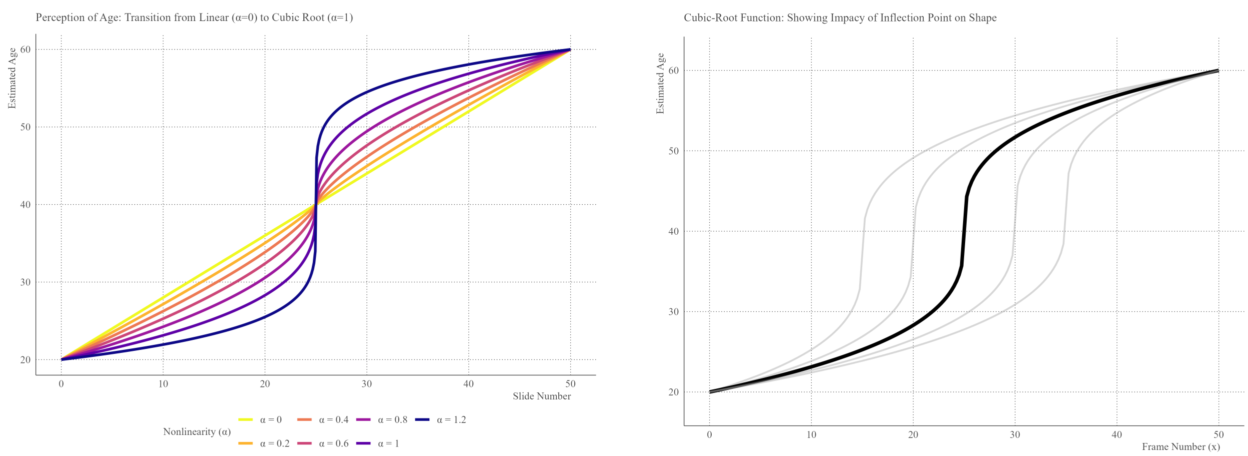
Willner, Paul and Glyn Rowe (2001). “Alcohol Servers’ Estimates of Young People’s Ages”. In: *Drugs: Education, Prevention and Policy* 8.4, pp. 375–383. DOI: 10.1080/09687630110063960.

Zepelin, H., R. A. Sills, and M. W. Heath (1986). “Is age becoming irrelevant? An exploratory study of perceived age norms”. In: *International Journal of Aging and Human Development* 24, pp. 241–256. DOI: 10.2190/1RAF-8YEW-QKU8-RTF8.

Appendix

A Sensitivity to the Assumed Form of Age Progression

Equation 2 assumes that perceived age increases as a linear function of morph progression. This assumption may not capture the dynamics of facial age perception. For example, perceptual sensitivity to age-related facial change may be relatively low during early adulthood and increase suddenly in middle age. A power function with exponent $1/3$ (i.e., a cubic root transformation) represents the limiting case of such concavity within the parameter range examined here.



(a): Effect of the non-linearity parameter (b): Effect of the inflection point parameter

Figure A.1: Parameterization of the power function

To assess the robustness of the reported findings to this modeling assumption, we recompute estimated age across a grid of functional forms parameterized by an exponent $\alpha \in [1/3,\ 1]$, where $\alpha = 1$ recovers the linear model and $\alpha = 1/3$ corresponds to the cubic root, and an inflection point parameter governing the locus of maximum curvature within the morph sequence. The full mixed-effects model was refit at each grid point and the resulting p-values for the Race $\times$ Gender interaction, target gender main effect, and target race main effect are shown as contours in Figures A.2 and A.3 for young-face and old-face starting conditions respectively, with lighter hues encoding smaller p-values and white contours demarcating the $p = 0.05$ (solid) and $p = 0.01$ (dashed) boundaries.

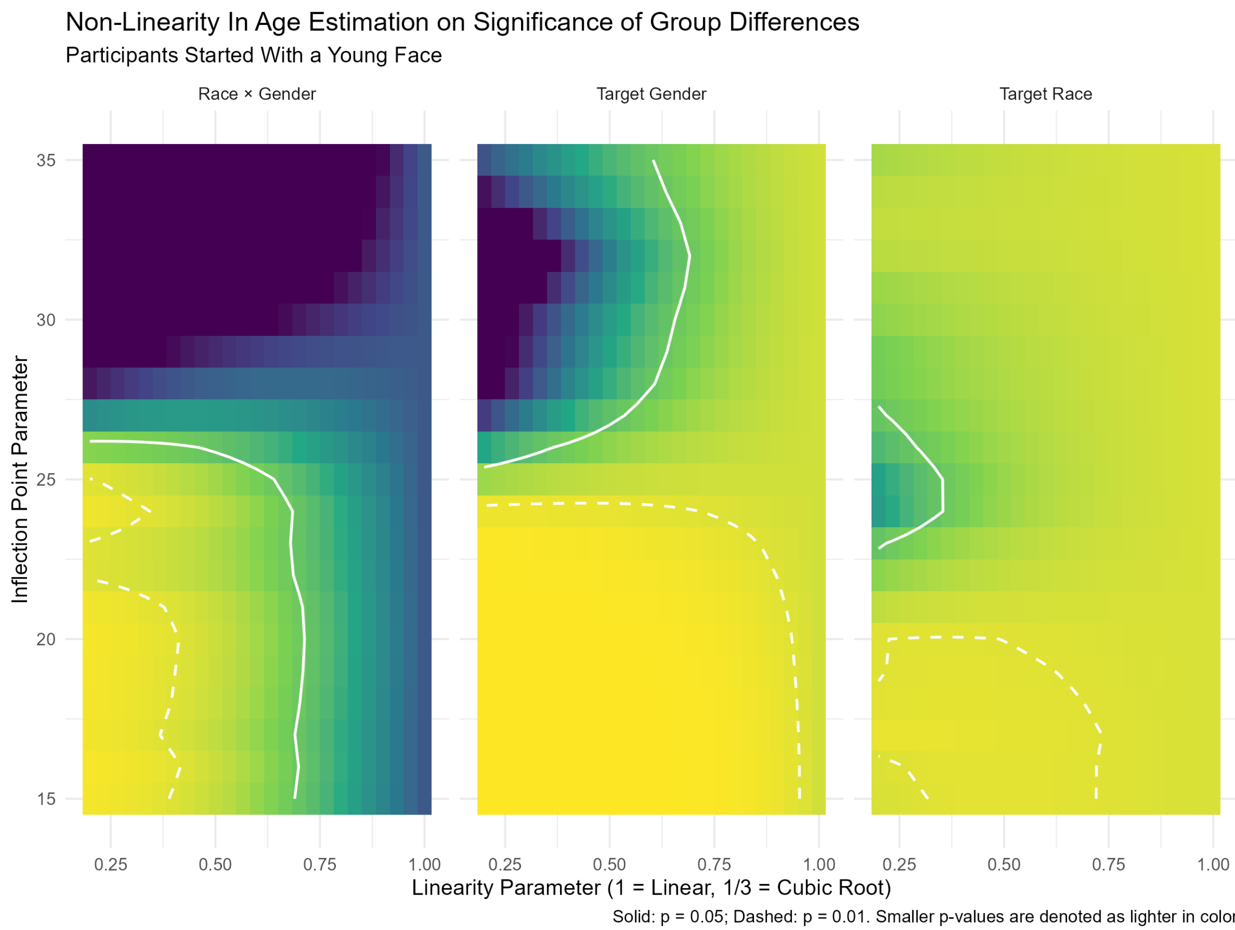


Figure A.2: Significance landscape for the three omnibus tests (Race × Gender interaction, target gender main effect, target race main effect) as a function of the linearity exponent α (x-axis; 1 = linear, $1/3$ = cubic root) and inflection point parameter (y-axis), for participants who initiated the task from the young-face endpoint. Cell color encodes the p-value from the corresponding mixed-effects model (lighter = smaller p). Solid and dashed white contours mark $p = 0.05$ and $p = 0.01$ respectively.

As shown in Figure A.2, the target gender main effect in the young-face condition is robust across the majority of the parameter space, retaining statistical significance at $p < 0.05$ across most combinations of α and inflection point in the Target Gender panel. The Race × Gender interaction attains statistical significance only within a circumscribed region corresponding to lower exponents and moderate inflection points, indicating that this effect is contingent on the assumed functional form of the age-morph mapping. The target race main effect was near uniformly statistically significant across parameterizations. In the old-face condition (Figure A.3), the contours are different: the Race × Gender interaction is nearly always statistically significant, with neither gender nor race on their own being statistically significant across most of the space examined (an exception is for highly non-linear, late-inflection models for gender). The old-face interaction reported under the linear model should accordingly be treated as hypothesis-generating rather than confirmatory.

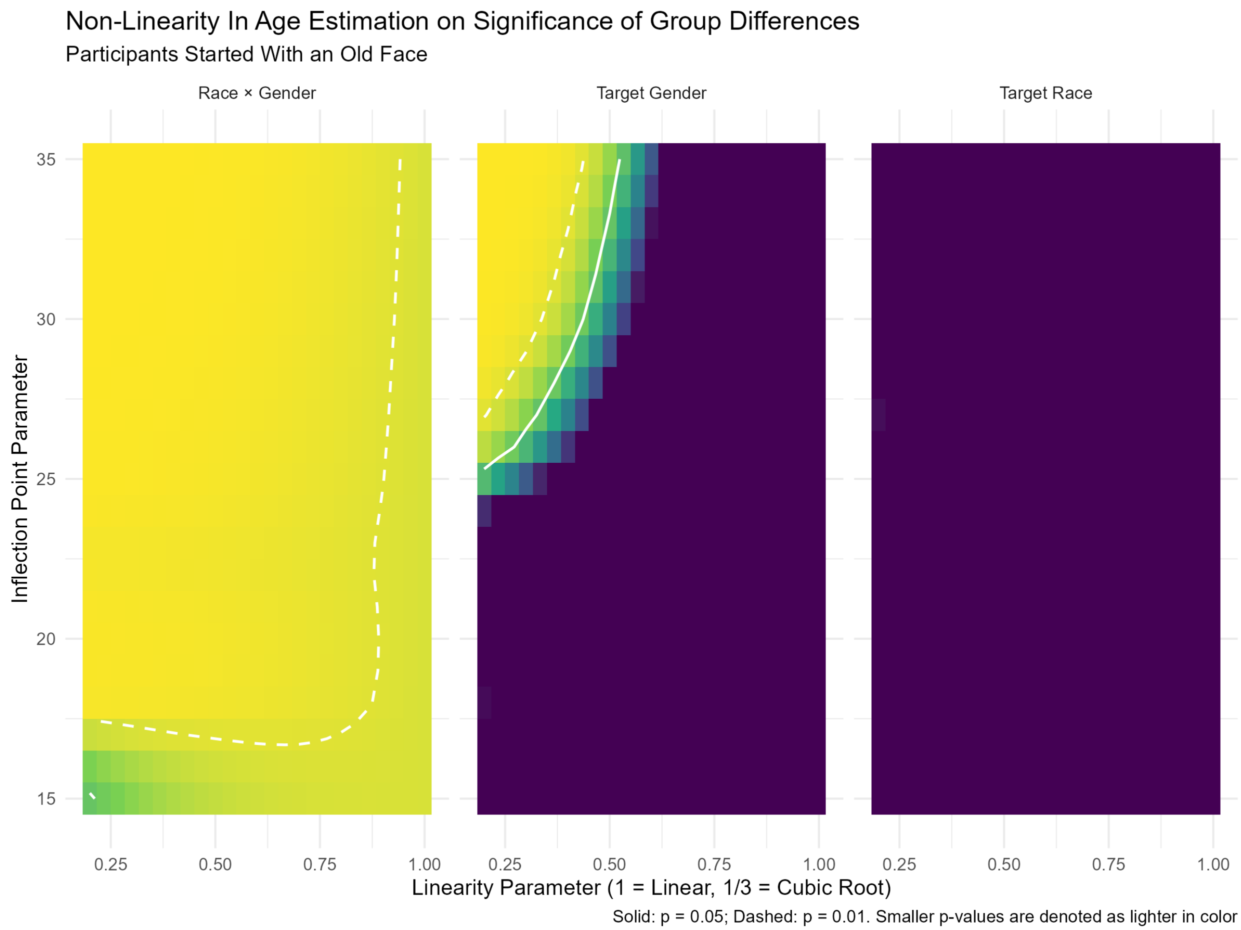


Figure A.3: Significance landscape as in Figure A.2, for participants who initiated the task from the old-face endpoint.

B Robustness Check: Classical Split-Plot ANOVA

The primary analyses are estimated as linear mixed-effects models via REML, a specification that offers several advantages — accommodation of unbalanced cell sizes, partial missingness handled through model-based estimation rather than listwise deletion, and Satterthwaite-approximated degrees of freedom that account for the random-effect structure. As a robustness check, all four models were re-estimated using a classical split-plot ANOVA, which differs from the primary specification in three consequential ways.

First, the classical approach computes F-ratios using a fixed error term structure derived from the balanced-design assumption: between-subjects effects are tested against a between-subjects error mean square, and within-subjects effects are tested against a pooled within-subjects error mean square. The mixed-effects approach, by contrast, constructs separate variance components for each random effect and derives denominator degrees of freedom from the estimated covariance structure via the Satterthwaite approximation. In practice, this means the two approaches can yield meaningfully different F-ratios and p-values when random-effect variance is poorly identified or when the data are unbalanced.

Second, the classical split-plot ANOVA requires complete data across all within-subjects cells: any participant missing a response for even one condition is excluded entirely from the analysis. The mixed-effects specification retains partial observations, estimating cell means from whatever data are available under the assumption of missing at random. This distinction is consequential in the present data, where several participants are missing responses.

Third, the classical specification does not include random slopes — it partitions variance into a participant random intercept and a single pooled within-subjects error term. Where that random slope accounts for meaningful individual variation in the gender effect, the classical approach will absorb this variance into the error term, potentially altering power and Type I error rates.

For participants who initiated the task from the young-face endpoint, the classical split-plot results for stop time are consistent with the primary analysis. The main effect of gender is large and highly significant ($F(1,\ 76) = 104.42$, $p < 0.001$, $\eta_p^2 = 0.58$), closely matching the primary estimate ($\eta_p^2 = 0.58$). We replictae the Race $\times$ Gender interaction ($F(3,\ 76) = 8.64$, $p < 0.001$, $\eta_p^2 = 0.25$), while the main effect of race did not reach the 5%

threshold of statistical significance ($F(3,\ 76) = 2.55$, $p = 0.062$, $\eta_p^2 = 0.09$). One participant was excluded due to a missing data; under the mixed-effects specification this participant was retained via partial-data estimation.

For participants who initiated from the old-face endpoint, the classical results again closely reproduce the primary analysis. The gender main effect is significant ($F(1,\ 83) = 16.23$, $p < 0.001$, $\eta_p^2 = 0.16$), with an effect size identical to the primary estimate. Neither the race main effect ($F(3,\ 83) = 0.84$, $p = 0.474$, $\eta_p^2 = 0.03$) nor the interaction ($F(3,\ 83) = 1.69$, $p = 0.175$, $\eta_p^2 = 0.06$) was statistically significant.

When looking at $\widehat{\text{Age}}$, the young-face-starting condition yields significant main effects of both race ($F(3,\ 75) = 4.23$, $p = 0.008$, $\eta_p^2 = 0.15$) and gender ($F(1,\ 75) = 4.33$, $p = 0.041$, $\eta_p^2 = 0.06$), consistent with the primary results ($\eta_p^2 = 0.13$ and $\eta_p^2 = 0.05$, respectively). The interaction is not statistically significant ($F(3,\ 75) = 1.35$, $p = 0.266$, $\eta_p^2 = 0.05$). Two participants were excluded due to missing data; the modest upward shift in the race effect size ($\eta_p^2 = 0.15$ vs. 0.13) most plausibly reflects the composition of the analytic sample under the non-missing restriction rather than a substantive difference between specifications.

For the old-face-starting condition, the two approaches diverge. As in the primary model, neither race ($F(3,\ 76) = 0.41$, $p = 0.748$, $\eta_p^2 = 0.02$) nor gender ($F(1,\ 76) = 0.67$, $p = 0.415$, $\eta_p^2 = 0.01$) are significant. The Race $\times$ Gender interaction, however, is significant under the classical specification ($F(3,\ 76) = 4.41$, $p = 0.006$, $\eta_p^2 = 0.15$), with a considerably larger effect size than in the primary model ($\eta_p^2 = 0.03$). This divergence is most plausibly attributable to the combined influence of exclusion of seven participants with missing data and the singular fit of the primary `lme4` model, which indicated that the participant-by-gender random slope is estimated at the boundary of the parameter space. When the random slope contributes negligible variance — as the singular fit implies — the classical approach's pooled error term may be more efficient, but the non-missing restriction simultaneously alters the sample. Given the uncertainty introduced by both the singular fit and the differential exclusions, neither specification should be treated as definitive for this model, and the interaction effect in the old-face estimated age condition warrants cautious interpretation regardless of which approach is preferred.