The Relationship between Appearance and Personality Across the Life Span
Leslie A. Zebrowitz, Mary Ann Collins and Ranjana Dutta

_Pers Soc Psychol Bull_ 1998; 24; 736
DOI: 10.1177/0146167298247006

The online version of this article can be found at:
http://psp.sagepub.com/cgi/content/abstract/24/7/736

Published by:
SAGE Publications
http://www.sagepublications.com

On behalf of:
Society for Personality and Social Psychology, Inc.

Additional services and information for _Personality and Social Psychology Bulletin_ can be found at:

Email Alerts: http://psp.sagepub.com/cgi/alerts

Subscriptions: http://psp.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

_Citations_ (this article cites 28 articles hosted on the SAGE Journals Online and HighWire Press platforms):
http://psp.sagepub.com/cgi/content/abstract/24/7/736#BIBL
The Relationship Between Appearance and Personality Across the Life Span

Leslie A. Zebrowitz  
Brandeis University  
Mary Ann Collins  
Spring Hill College  
Ranjana Dutta  
Southwest Missouri State University

Using archival data, the relationship of appearance to personality was investigated from childhood to age 60. Lagged effects of appearance on personality and vice versa, as well as accuracy of appearance stereotypes, were assessed. For men, lagged effects of attractiveness on personality were consistent with a self-fulfilling prophecy; for adolescent boys, lagged effects of babyfacedness on personality were consistent with a self-defeating prophecy; for women, lagged effects of personality on attractiveness were consistent with a Dorian Gray effect, whereby early personality produces a congruent later appearance. There was no evidence for accuracy of the baby-face stereotype, which was significantly inaccurate for adolescent boys. Accuracy of the attractiveness stereotype required effects of a stable earlier appearance on later personality or a stable earlier personality on later appearance.

There is considerable evidence to indicate that people’s facial appearance can have a profound influence on their social environment. Most notably, there is an attractiveness halo effect, whereby people who are physically attractive are perceived more positively on a variety of dimensions, and a baby-face overgeneralization effect, whereby people with more neotenic facial structures (round faces, large eyes, high eyebrows, small noses and chins) are perceived to have more childlike traits. Evidence of such facial stereotyping is shown by and about individuals of all ages and races, and it has been demonstrated to have significant social consequences in many domains. The myriad social outcomes of appearance, discussed in more detail later in this article, have led researchers to postulate that it may create a self-fulfilling prophecy effect, whereby people develop those traits that their appearance leads others to expect (Adams, 1977; Langlois, 1986; Sorrell & Nowak, 1981). However, compelling evidence for such an effect is limited to short-term laboratory experiments (Snyder, Tanke, & Berscheid, 1977). This study sought more ecologically valid evidence using a longitudinal data archive that provided photographs of individuals and measures of their personality traits across the life span. A second possible relationship between appearance and personality was also examined—a Dorian Gray effect in which personality influences the development of facial appearance, similar to how the deeds of the protagonist in Oscar Wilde’s (1890/1974) novel influenced changes in his portrait. Finally, this research addressed the question of whether either of these developmental processes contribute to accurate facial stereotypes.

Meta-analyses have revealed that the strongest component of the attractiveness halo is the view that attractive individuals are more sociable, socially skilled, and popular than their less-attractive peers. They are also viewed as more dominant and more competent (Eagly, Ashmore, Makhijani, & Longo, 1991; Feingold, 1992). The attribution of more positive social behaviors and greater competence to more attractive individuals is seen

Authors’ Note: This research was supported by a grant from the National Institute of Mental Health (Grant No. MH42884) that was awarded to Leslie A. Zebrowitz. We express appreciation to the Institute of Human Development at the University of California, Berkeley, for providing the archival data used in this investigation. Thanks are also extended to Ron Spiro for his statistical assistance in the data analyses. Correspondence concerning this article should be addressed to Leslie A. Zebrowitz, Department of Psychology, Brandeis University, Waltham, Massachusetts 02254-9110.

© 1998 by the Society for Personality and Social Psychology, Inc.
in perceptions of infants and children as well as adults (e.g., Dion, 1972; Dion & Berscheid, 1974; also see Langlois, 1986, for a pertinent review). The components of the baby-face stereotype show some similarities and some differences to the attractiveness halo, with the baby-face stereotype being less socially desirable. Similar to attractive individuals, an individual who has a baby face is seen as warmer and less likely to show antisocial behavior. However, whereas attractive individuals are viewed as dominant, baby-faced individuals are seen as submissive. In addition, whereas a component of the competence attributed to attractive individuals is being ambitious and hardworking, baby-faced individuals tend to be seen as naive and irresponsible. The same constellation of traits is attributed to baby-faced individuals of both sexes and all ages (Zebrowitz & Montepare, 1992; see Montepare & Zebrowitz, 1998, and Zebrowitz, 1997, for pertinent reviews).

These reliable and strong effects of facial appearance on behavioral expectations may produce self-fulfilling prophecy effects via the creation of different social environments for people who vary in appearance (Adams, 1977; Langlois, 1986; Sorrell & Nowak, 1981). Although prior research on self-fulfilling prophecies has demonstrated circumscribed effects (Snyder et al., 1977), there is also considerable evidence to indicate that expectations induced by facial appearance have significant social outcomes that could produce more global influences on personality development. For instance, attractive children are treated more warmly by their parents as well as by strangers; attractive adults are preferred as dates and mates; and when in need, attractive people are more able to procure aid from bystanders, they often have greater social influence, and they are favored in the job market and in the criminal justice system (e.g., see Berscheid & Walster, 1974; Bull & Rumsey, 1988; Langlois, 1986; and Zebrowitz, 1997, for reviews of this literature). Facial maturity, with attractiveness controlled, also affects social outcomes. Thus, baby-faced people are preferred as dates by highly assertive individuals, who tend to dominate others; their misdeeds are less likely to be seen as intentional in childhood; they are favored in the criminal justice system in adulthood when accused of stereotype-inconsistent crimes; they are selected for jobs that require traits associated with the baby-face stereotype; and in childhood, they are given less-demanding chores, tasks are explained to them more slowly, and they are spoken to more affectionately (Berry & Zebrowitz-McArthur, 1988; Collins & Zebrowitz, 1995; Hadden & Brownlow, 1991; Zebrowitz, Brownlow, & Olson, 1992; Zebrowitz, Kendall-Tackett, & Fafel, 1991; Zebrowitz & McDonald, 1991; Zebrowitz, Tenenbaum, & Goldstein, 1991; also see Montepare & Zebrowitz, 1998, and Ze-

browitz, 1997, for pertinent reviews). If attractive or baby-faced people respond to these social pressures with behavior that confirms others’ expectancies, they may, over time, develop personalities that match the stereotypes associated with their appearance.

Although the evidence for differential treatment of people based on their appearance has typically yielded the prediction of self-fulfilling prophecies, such effects may in fact be elusive. There are a number of factors in real-life social interactions that can work against the behavioral confirmation of others’ expectancies and may even produce self-defeating prophecies, whereby the expectancy initiates a chain of events that cause contradictory behavior (cf. Jussim, 1991; Snyder, 1992). More specifically, self-fulfilling prophecies may fail to occur because people engage in self-verification, behaving in ways that provide corrective feedback to perceivers with an erroneous expectation rather than confirming the expectation (e.g., Swann & Ely, 1984); because the way people are treated inadvertently elicits behavior contrary to what perceivers’ expect (e.g., Zebrowitz et al., 1992); or because people show compensatory behavior, behaving in a manner that defies an expectation that is disagreeable (e.g., Miller, Rothblum, Felicio, & Brand, 1995). The multiple routes to self-defeating prophecies make it difficult to generate specific predictions in regard to when they will occur. However, a tendency to compensate for disagreeable expectations may be likely in the case of baby-faced males because the traits expected of baby-faced people are stereotypically feminine (Friedman & Zebrowitz, 1992).

The Dorian Gray effect, whereby personality influences appearance, has received almost no attention in psychological research. However, it can be readily found in folklore and literature. It is captured not only in the novel for which the effect is named but also in other novels and folktales. For example, in Dr. Jekyll and Mr. Hyde (Stevenson, 1886/1981), the ugly form of Mr. Hyde becomes more and more predominant with increases in the antisocial behavior of this split personality. Similarly, in Beauty and the Beast (Doyle, 1973), the ugly form of the beast is replaced by a handsome prince when he shows warm and loving behavior. Although we tend to think of facial structure as fixed, there is in fact some plasticity as well as some data to indicate that personality can play a causal role in the development of appearance. Thus, people with a tense, irritable temperament may tense certain facial muscles in a way that yields different jaw development from that shown in people who are more relaxed (e.g., Kreiborg, Jensen, Moller, & Bjork, 1978; Moller, 1966). Expressive habits may also influence facial appearance. Indeed, the personality dispositions of elderly people are reflected in their faces. For example,
those with a hostile disposition look angry even when posing a neutral expression (Malatesta, Fiore, & Messina, 1987). Similarly, the person who habitually raises his or her eyebrows will look more like a baby-faced individual, and one who habitually narrows his or her eyes will look less attractive as well as more like a mature-faced individual. The Dorian Gray effect may reflect a direct effect of personality on appearance, as the preceding examples illustrate, or it may be mediated through the environment. For example, individuals with a hostile temperament may be more likely to choose activities that have a deleterious effect on their appearance, such as boxing or alcohol abuse, and those who are highly sociable may choose grooming aids that have a beneficial effect on their appearance (e.g., Cash, 1990; Cash & Cash, 1982).

Both self-fulfilling prophecy and Dorian Gray effects could contribute to the accuracy of facial stereotypes over time. Existing evidence for such accuracy is mixed, with the strongest indication being a positive association between attractiveness and social skills (Feingold, 1992). There is little evidence pertaining to the accuracy of the baby-face stereotype. Consistent with the view that baby-faced adults are less dominant and threatening than mature-faced adults, baby-faced men of college age scored lower on personality tests of aggressiveness (Berry, 1991). Baby-faced men also reported less control and influence over opposite-sex interactions as well as greater disclosure and intimacy in their social interactions; this is consistent with the view that baby-faced people are submissive and warm (Berry & Landry, 1997).

This study examined self-fulfilling prophecy effects and Dorian Gray effects. Although no specific predictions were made regarding age and sex differences, these effects were explored separately for males and females across two time spans—from childhood to adolescence and from adolescence to later adulthood. The accuracy of appearance stereotypes across the life span were also examined. If appearance produces self-fulfilling prophecy effects, as hypothesized by past investigators, then early attractiveness and early babyfaceness within each time span should predict a later personality that is congruent with the attractiveness and baby-face stereotypes, respectively. If, on the other hand, appearance produces self-defeating prophecy effects, then early appearance should predict a later personality that is incongruent with the appearance stereotype. If there are Dorian Gray effects, as suggested in folklore and literature, then an early personality that matches the attractiveness stereotype should predict later attractiveness, whereas an early personality that matches the baby-face stereotype should predict later babyfaceness. Finally, to the extent that self-fulfilling prophecy effects or Dorian Gray effects have been operating, appearance stereotypes may become accurate over time.

METHOD

Participants

The participants in this study were White middle-class individuals who were drawn from the Intergenerational Studies of Development and Aging, a combination of three longitudinal studies begun between 1928 and 1933 and archived at the University of California, Berkeley, Institute of Human Development (IHD). The three studies, although different in objectives, all included photographs of the participants as well as common measures of cognitive, personality, social, and biological characteristics. Moreover, special efforts were made to derive common measures on which the three attrited samples could be pooled. (For further details about the three studies, see Eichorn, Clausen, Haan, Honzik, & Mussen, 1981; Jones, Bayley, Macfarlane, & Honzik, 1971.) Participants selected for inclusion in this investigation comprised a subset of those included in a study that investigated the stability of appearance across the life span (Zebrowitz, Olson, & Hoffman, 1993) and overlapped with those people who were included in studies that investigated the contributions of appearance to occupational outcomes (Collins & Zebrowitz, 1995) and the relationship between perceived and real honesty across the life span (Zebrowitz, Voinescu, & Collins, 1996).

To be included in this study, participants needed both appearance data and pertinent personality scores across various age levels, including childhood, puberty, adolescence, the 30s, and the 50s. Because the longitudinal panel that met this criterion from childhood to the 50s was not sufficiently large for the planned analyses employing structural equation modeling, the data were concatenated into a youth panel and an adult panel. Subjects were selected for the youth panel if they had pertinent personality scores and facial appearance data in adolescence as well as in either childhood or puberty. There were 86 male subjects and 92 female subjects who met this criterion. Subjects were selected for the adult panel if they had personality scores and facial appearance data in their 50s as well as in either adolescence or their 30s. There were 83 male subjects and 96 female subjects who met this criterion. A total of 57 males and 63 females were included in both the youth and adult panels. It should be noted that these samples were not selected for their particular appearance or personality qualities, thus minimizing possible selection bias. The appearance and personality of participants included in the youth and adult panels were compared with that of those who were excluded. There were no significant differences, indicating that our subsample was equiva-
lent on these indexes to the larger sample from which it was selected.

**Measures**

Three sets of variables were employed: ratings of facial attractiveness and babyfakeness, ratings of specific facial features, and personality measures. Q-sort data existing in the archive were used in the present research to create two global indicators of personality—a Prototypical Attractive Personality (PAP) scale and a Prototypical Babyfaced Personality (PBP) scale—that served as the primary personality measures. In addition, Q-sort scales and three California Psychological Inventory (CPI) scales that were already in the archives were used in supplementary analyses to elucidate the specific traits that contributed to significant effects on the global measures. These preexisting scales were selected because they tapped dimensions that are relevant to stereotypes of attractive and/or baby-faced individuals: dominance (stereotypically high for attractive and low for baby-faced individuals); sociability (stereotypically high for attractive and low for baby-faced individuals); goodness (stereotypically high for attractive and low for baby-faced individuals); and competence (stereotypically high for attractive and low for baby-faced individuals).

**ATTRACTIVENESS AND BABYFACENESS RATINGS**

Ratings of attractiveness and babyfakeness were taken from a previous study by Zebrowitz et al. (1993, Study 1). Reliability of these ratings was high. The average alpha coefficient across age was .87 for males’ attractiveness, .90 for females’ attractiveness, and .85 for both males’ and females’ babyfakeness. Research supporting the validity of these ratings includes significant agreement between the attractiveness ratings of adolescent girls and the prettiness of these same faces, as rated by IHD staff in the 1960s (Zebrowitz et al., 1993). Also, both attractiveness and babyfakeness have been found to predict real-life outcomes, including timing of marriage (Kalick, Zebrowitz, Langlois, & Johnson, 1998), job type, and military recognition (Collins & Zebrowitz, 1995).

**SPECIFIC FACIAL RATINGS**

Several facial qualities were rated by two male and three female judges to elucidate what specific aspects of appearance could account for any observed Dorian Gray effects on attractiveness or babyfakeness. These included ratings of how much participants were smiling, how old they looked, how flattering their hairstyles were, how much care and effort their hairstyle showed, how much makeup they wore (if any), how attractive their teeth were (if visible), how wrinkled and lined their faces were, how noticeable their laugh lines were, and how noticeable their frown lines were. All ratings were made on a 5-point scale, and age, which was freely estimated. In addition, a single judge recorded whether participants were wearing glasses. Order of age and order of participants within age were counterbalanced across judges. Judges rated all faces on one facial structure before rating another, and slides were shown for 6 seconds for each rating. The reliability of these ratings was generally high, averaging .92 for smiling, .75 for age, .73 for flattering hair, .74 for hair care, .79 for makeup, .87 for teeth, .77 for wrinkles and lines, and .84 for laugh lines. Ratings of frown lines were unreliable before age 30, but the reliability of this rating was marginally acceptable in the age ranges of the 30s and 50s, averaging .64.

**PERSONALITY ASSESSMENTS**

The PAP scale and the PBP scale described in this article were created with previously obtained longitudinal California Q-sort data in the IHD archives (Block, 1961/1978, 1971). These data, which provide a common measure of personality for the Intergenerational Studies of Development and Aging (Eichorn, 1981), consisted of expert clinicians’ sorting of 100 descriptive items, such as “seeks reassurance from others,” into a normally distributed set of nine categories according to the extent to which each item characterized the participant’s personality. At least two judges rated each participant, and no judges rated a participant at more than one age. The childhood Q-sorts were conducted at age 10, puberty Q-sorts at age 13, adolescent Q-sorts at age 16, 30s Q-sorts between the ages of 30 and 40, and 50s Q-sorts between the ages of 52 and 60, depending on the particular Intergenerational Study of Development and Aging sample. Participants in one of the three samples are missing childhood data, whereas those in the other two have Q-sort data at all age levels. Of the 100 items sorted at each age, 73 items showed reliability across all ages and constituted the longitudinal Q-sort.

The Q-sorts were based on varied sources of information, including interview, questionnaire, and observational data gathered from the participants, their families, and school records. The CPI scores, described later in this article, were withheld from judges so that the Q-sort measures would be independent. Photographs of the participants were not available to judges who were performing the Q-sorts. Individuals who interviewed participants in childhood, puberty, and adolescence did not serve as Q-sort judges. Thus, at these ages, the Q-sorts could not have been influenced by judges’ reactions to the participants’ physical appearance. Individuals who interviewed participants in their 30s and 50s also served as one of the Q-sort judges. However, there was at least one other judge who had not seen the participants. Because significant interjudge reliability was established for the Q-sorts, they could not have been greatly influenced by the
appearance information that was available to only one judge (Block, 1971; B. Burek, IHD archivist, personal communication, November 16, 1995; Eichorn, 1981).

PAP and PBP scales. Personality prototypes, representing the stereotypes of attractive and baby-faced individuals, were created to serve as personality indicators in the present investigation from the archival 73-item longitudinal Q-sort. The procedure employed followed that used in other studies to derive measures of particular constructs from personality descriptions formulated with the Q-sort (e.g., Block, 1961/1978; Reise & Wink, 1995; White et al., 1994; Whitney, Sagrestano, & Maslach, 1994). These measures assume that a personality construct can be assessed by the degree of congruence between an individual's personality profile and the personality profile of the prototypical representative of the construct (e.g., the prototypical, highly attractive individual). The prototype is created by having experts use the Q-sort to describe the personality of the prototypical representative. Descriptions of particular individuals can then be compared with the prototype. An individual whose personality profile is similar to that of the prototype is assumed to be high on the construct, whereas one with a dissimilar profile is assumed to be low. (For a more extensive discussion of the use of the Q-sort as a measure of personality, see Ozer, 1993).

Five expert judges (social psychologists conversant with the literature on facial stereotypes) sorted the 73 longitudinal Q-sort items into a normally distributed set of nine categories according to how characteristic each behavior was of the stereotype of an attractive (or baby-faced) person. Because some people have suggested that attractiveness effects may differ for males and females, five judges sorted the items according to how characteristic each was of the stereotypical attractive male (standardized alpha = .92), and three of these judges also sorted the items according to how characteristic each was of the stereotypical attractive female (standardized alpha = .90). Correlations between the three overlapping judges' Q-sorts for the stereotypical attractive male and female averaged .89, thereby justifying the use of a single prototype for both sexes, which was created by averaging ratings across judges.

Each participant's score on a PAP scale was computed by correlating his or her 73-item Q-sort profile with the 73-item PAP Q-sort, yielding an index of the degree to which the participant's personality resembled the personality of a prototypical attractive person. This correlation was called the PAP score. One Q-sort item that assessed the attractiveness of the subject was deleted before performing these correlations because this item could inflate the match to the prototype. The same procedure was used to assess the resemblance of each subject's personality to that of the prototypical baby-faced individual. Q-sort ratings for the PBP scale were averaged across judges, who showed high interjudge agreement (standardized alpha = .92). The correlation of each participant's 73-item Q-sort profile with the resultant PBP Q-sort was called the PBP score. The three items that were judged as most characteristic of the prototypical attractive individual were "has social poise and presence," "appears socially at ease," and "tends to arouse liking and acceptance"; those items that were judged as least characteristic were "feels cheated and victimized by life," "is vulnerable to real or fancied threat," and "basically submissive." The three items that were judged as most characteristic of the prototypical baby-faced individual were "arouses nurturant feelings," "basically submissive," and "has warmth"; those items that were judged as least characteristic were "behaves in an assertive fashion," "critical skeptical," "not easily impressed," and "guileful, deceitful, manipulative."

Q-sort component scales. Component analysis had been employed by previous investigators to generate six scales from the 73-item longitudinal Q-sort (Haan, Millsap, & Hartka, 1986; Millsap & Meredith, 1988). Four of these component scales were relevant to stereotypes of attractive and/or baby-faced individuals, and they were used in this study both to validate the attractive and baby-faced personality prototypes and also as secondary measures to elucidate the specific traits that contribute to significant effects on the global PAP and PBP measures. The component scales employed for these purposes were: Submissive/Assertive, Aaloof/Outgoing, Warm/Hostile, and Dependable. The Submissive/Assertive and the Aaloof/Outgoing component scores have been reversed from the original Millsap and Meredith (1988) analyses so that they are scored in the same direction as the CPI Dominance and Sociability scales, respectively. The validity of these scales has been demonstrated in previous research, including a study that found that participants' scores on each scale predicted occupational outcomes (Collins & Zebrowitz, 1995).

CPI. CPI data were available in the archives for participants in their 30s and 50s. Three of the CPI scales—Dominance, Sociability, and Responsibility—were relevant to stereotypes of attractive and/or baby-faced individuals, and they were used in this study for the same purposes as the Q-sort Component scales. (See Megargee, 1986, for a more complete discussion of these measures.)

RESULTS

Validation of the Attractive and Baby-Face Personality Prototypes

Correlations between participants' PAP scores and their scores on other personality measures provided
evidence for the construct validity of the PAP scale. The PAP scale tapped the sociability component of the attractiveness halo at all ages, as indicated by significant positive correlations of PAP scores with scores on the Q-sort Outgoing scale (.73 for females and .71 for males, averaged across age) and the CPI Sociability scale (.26 for males and .41 for females, averaged across age). Significant correlations of PAP scores with scores on the Q-sort Assertiveness scale (.23 for males and .22 for females, averaged across age) and the CPI Dominance scale (.29 for males and .34 for females, averaged across age) revealed that the PAP scale also picked up the dominance component of the attractiveness halo. The goodness aspect of the halo was seen in negative correlations of PAP scores with scores on the Q-sort Hostility scale (−.30 for males and a marginally significant −.19 for females, averaged across age). The competence component was seen in positive correlations of PAP scores with scores on the Q-sort Dependability scale (.42 for males and .34 for females, averaged across age) and with the CPI Responsibility scale (.27 for males and a marginally significant .18 for females, averaged across age).

Significant correlations between participants' PBP scores and their scores on other measures provided evidence validity for the PBP scale as well. Consistent with the perceived warmth and sociability of baby-faced individuals, PBP scores were positively correlated with scores on the Q-sort Outgoing scale (.33 for males and .28 for females, averaged across age), albeit not with the CPI Sociability scale (−.04 for males and −.08 for females, averaged across age). The PBP scale also tapped the low dominance component of the baby-face stereotype, as indicated by negative correlations of PBP scores with scores on the Q-sort Assertiveness scale (−.69 for males and −.75 for females, averaged across age) and the CPI Dominance scale (−.25 for males and females, averaged across age). Negative correlations of PBP scores with scores on the Q-sort Hostility scale (−.69 for males and −.60 for females, averaged across age) revealed that the PBP scale captured the view that baby-faced individuals are nonthreatening. Finally, the low competence aspect of the baby-face stereotype was seen in negative correlations of PBP scores with scores on the Q-sort Dependability scale (−.22 for males and −.29 for females, averaged across age), albeit not with the CPI Responsibility scale (−.03 for males and −.02 for females, averaged across age). It should be noted that the higher correlations of PAP and PBP scores with Q-sort scales than with CPI scales reflects the fact that the prototypes and the Q-sort component scales both were derived from the same set of ratings.

In sum, the individual whose own Q-sort profile correlates highly with the PAP Q-sort scale tends to be sociable, dominant, nonhostile, and dependable. The person whose Q-sort profile correlates highly with the PBP Q-sort scale tends to be sociable, submissive, nonhostile, and nondependable. Sociability is the strongest component of the PAP scale, whereas submissiveness and low hostility are most central to the PBP scale. The result is that the two scales are quite independent, as shown by the lack of a significant correlation between them (r(70) = −.05).

**Analyses of Individual Differences in Developmental Change**

PAP and PBP scores and appearance ratings were employed in structural equation modeling (SEM) to examine individual differences in developmental change that would be predicted from self-fulfilling prophecy effects and Dorian Gray effects. (See Hertzog, 1990, for a discussion of the value of SEM for this purpose.) SEM was employed rather than regression analyses because it permitted the comparison of self-fulfilling prophecy and Dorian Gray effects within a single model. The purpose of the SEM analyses was not to test the absolute goodness of fit of a particular model, but rather, the goal was to test the difference between competing models, namely, one that included self-fulfilling prophecy and/or Dorian Gray paths versus a baseline model that included only the stability paths for appearance and personality.

**MODELS TESTED**

For both the youth and the adult panels, two-step structural equation models were constructed using LISREL VII (Jöreskog & Sörbom, 1989) (see Figure 1). A latent early attractiveness construct had as indicators z scores of participants' attractiveness at Time 1 and Time 2 (childhood and puberty for the youth panel; adolescence and the 50s for the adult panel). A latent early babyfakeness construct had z scores of participants' babyfakeness at Times 1 and 2 as indicators. Latent early personality constructs had participants PAP or PBP z scores at Time 1 and Time 2 as indicators. There were only single indicators of attractiveness, babyfakeness, and personality at Time 3 (adolescence in the youth panel; 50s in the adult panel). Two indicators of the early constructs were used for both pragmatic and theoretical reasons. The pragmatic reason was that missing data resulted in inadequate sample sizes for SEM analyses using only Time 1 or Time 2 data in either the youth panel or the adult panel. The theoretical reason was that self-fulfilling prophecy effects on personality should require time to develop and should be most likely to occur when appearance has been stable over time. Including both Time 1 and 2 data in the early construct provides a more reliable index of appearance over time. Finally, it should be noted that regression analyses that averaged
Scores across the two early time periods yielded results that were identical to those obtained using two indicators of the early constructs in the SEM analyses.

The model shown in Figure 1 was tested on PAP scores and PB P scores for male participants and for female participants in the youth and adult panels. These models estimated the stability paths for each variable (represented by $s_1$, $s_2$, and $s_3$), the paths from early attractiveness and early babyfaceness to later personality (represented by $a$ and $b$), and the paths from early personality to later attractiveness and later babyfaceness (represented by $a'$ and $b'$). Thus, estimates of the lagged effects of a particular appearance variable on personality controlled for earlier personality and earlier values of the other appearance variable and estimates of the lagged effect of personality on a particular appearance variable controlled for earlier levels of that appearance factor. The paths from early babyfaceness to later attractiveness and from early attractiveness to later babyfaceness were not included in the model because they were not theoretically meaningful. The baseline LISREL model assessed only stabilities of attractiveness, babyfaceness, and the relevant personality characteristic (i.e., $a$, $b$, $a'$, $b'$) were fixed at 0). In a second model, the cross-lagged path coefficients were also freed. This nested sequence permitted an assessment of the contribution of the cross-lags to the goodness of fit of the model by a chi-square difference test. In the final model, all nonsignificant lags were trimmed. Only those models in which the cross-lagged paths provided a significant improvement over the baseline stability model are reported in Table 1.

MISSING DATA

In both panels, missing values at Time 1 and Time 2 were replaced with the value for that participant at the other age. This was considered a better estimate than the mean value for the entire sample at the missing age because appearance and personality were relatively stable from Time 1 to Time 2 (Haan et al., 1986; Zebrowitz et al., 1993). Although there was sometimes a sizable percentage of missing data at one specific age, every participant within the youth and adult panels had data for a given measure either at Time 1 or at Time 2. (Participants were selected so that all had data at Time 3.)

In most cases, missing data can be attributed to the absence of data collection for a particular sample at that age. Missing values for the appearance measures were relatively high at Time 1 in the youth panel (17% for boys, 22% for girls) because only one photograph was taken at that time for one of the three samples in the Intergenerational Studies of Development and Aging, whereas three photographs were taken for the other samples. Missing values for appearance were high at Time 2 in the adult panel (24% for men, 38% for women) because no photographs were available at this age for one entire sample. Missing values for the personality measures were elevated at Time 1 in the youth panel (52% for boys, 47% for girls) because Q-sort data were unavailable at that age for one entire sample. Missing values for personality were elevated at Time 1 in the adult panel (30% for men and women) because Q-sort data were unavailable at that age for a control group within one sample that had fewer measures taken.

When samples for whom no data were collected are excluded, the missing values become quite small, averaging about 3%, with two exceptions. The exceptions included missing personality data for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%). To examine whether attrition was systematically affected by personality or appearance in these cells, we compared appearance means for men at Time 1 in the adult panel (13%) and missing appearance data at Time 2 for women in the adult panel (26%).
TABLE 1: Standardized Path Coefficients and Change and Fit Indexes for Trimmed Cross-Lag Models

<table>
<thead>
<tr>
<th>Personality Characteristics</th>
<th>Stability Paths</th>
<th>Lagged Paths</th>
<th>Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$S_a$</td>
<td>$S_b$</td>
<td>$S_p$</td>
</tr>
<tr>
<td>Prototypical Attractive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personality (PAP) scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult panel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men ($n=83$)</td>
<td>.35**</td>
<td>.24*</td>
<td>.40****</td>
</tr>
<tr>
<td>Women ($n=96$)</td>
<td>.31**</td>
<td>.27**</td>
<td>.59****</td>
</tr>
<tr>
<td>Prototypical Baby-Faced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personality (PBP) scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Youth panel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boys ($n=86$)</td>
<td>.76****</td>
<td>.94****</td>
<td>.60****</td>
</tr>
<tr>
<td>Girls ($n=92$)</td>
<td>.60****</td>
<td>.83****</td>
<td>.47****</td>
</tr>
</tbody>
</table>

NOTE: Models for PAP scores (youth panel) and PBP scores (adult panel) are not included in this table because the cross-lagged models provided no significant improvement over baseline stability models. $S_a$ = stability of attractiveness; $S_b$ = stability of babyfaceness; $S_p$ = stability of personality; $a$ = early attractiveness to later personality; $b$ = early babyfaceness to later personality; $a^1$ = early personality to later attractiveness; $b^1$ = early personality to later babyfaceness; $df = 24$ in baseline models; CFI = Bentler's comparative fit index for final accepted model.

*p < .10. **p < .05. ***p < .01. ****p < .001.

the validity of the results, we will discuss only those significant effects that showed no substantial change when examined in a multiple regression analysis on a subsample of participants with no missing data.4

STABILITY OF MEASURES

The stability of each measure from the early to the later time periods within each panel was assessed in the baseline models. As shown by the path coefficients in Table 1, both appearance and personality showed significant stability over time. Although there was stability in personality, it was not so great as to preclude an effect of early appearance on later personality. Similarly, with the possible exception of babyfaceness in the youth panel, the stability of appearance was not so great as to preclude an effect of early personality on later appearance.5

LAGGED EFFECTS OF ATTRACTIVENESS

ON PERSONALITY (PATH a)

Youth panel. There was no evidence for self-fulfilling prophecy effects of early attractiveness on later personality for boys or girls. The models containing the cross-lagged paths did not provide a significantly better fit to the data than did the baseline stability models.

Adult panel. As shown in Table 1, the models testing PAP scores for men and women in the adult panel revealed a marginally significant self-fulfilling prophecy effect of early attractiveness on later personality for men but no effect for women. High attractiveness in adolescence and the 30s predicted a more prototypically attractive personality for men in their 50s. To clarify the specific components of the prototypical attractive personality at age 50 that were responsible for this effect, regression analyses were performed that predicted men's later PAP scores from early attractiveness and early PAP scores with four additional trait measures at age 50 entered into the regression equation. The results revealed that the effect of early attractiveness on later PAP scores was eliminated when either Hostility (Q-sort score) or Sociability (average of Q-sort Outgoing and CPI Sociability scores) at age 50 was entered ($\beta = .13$ and .11, respectively; $p > .10$). This indicated that the effect of early attractiveness on later PAP scores could be explained by its negative effect on later Hostility and its positive effect on later Sociability. On the other hand, the effect of early attractiveness on later PAP scores remained marginally significant when Competence (average of Q-sort Dependable and CPI Responsibility scores) or Dominance (average of Q-sort Assertiveness and CPI Dominance scores) were entered ($\beta = .19$ and .17, respectively; both $p = .10$), indicating that the effect of early attractiveness on later PAP scores was not due to influences on either of these traits.

LAGGED EFFECTS OF BABYFACENESS

ON PERSONALITY (PATH b)

Youth panel. As shown in Table 1, the models that tested PBP scores for boys and girls in the youth panel found no evidence for a self-fulfilling prophecy effect of early babyfaceness on later personality for either boys or girls. Rather, there was evidence for a self-defeating prophecy effect for adolescent boys, whereby higher babyfaceness in childhood and puberty predicted a less prototypically baby-faced personality in adolescence. To clarify the specific components of the prototypical baby-faced personality that were responsible for this effect, regression analyses were performed that predicted boys' adolescent PBP scores from early babyfaceness and early
PBP scores with four additional adolescent trait measures entered into the regression equation. The results revealed that the effect of early babynesseness on later PBP scores was eliminated when either Q-sort Dominance or Hostility was entered (β = .11 and .12, respectively; both ps > .10), indicating that the effect of early babynesseness on later PBP scores could be explained by its positive effects on these specific, correlated traits. On the other hand, the effect of early babynesseness on later PBP scores remained unchanged when Q-sort Sociability or Competence was entered into the regression equation (β = .26 and .23, respectively; both ps < .01), indicating that the effect of early babynesseness on later PBP scores was not due to an effect on either of these traits. Unlike the follow-up analyses for adult men, only Q-sort Scale scores were used because the CPI was not administered in adolescence.

Adult panel. The models testing PBP scores for men and women in the adult panel with the cross-lagged paths included did not provide a significantly better fit to the data than did the baseline stability model. Thus, there was no evidence for self-fulfilling prophecy effects of early babynesseness on later personality for adult men or women.

LAGGED EFFECTS OF PERSONALITY ON APPEARANCE (PATHS α^1 AND β^1)

Youth panel. As shown in Table 1, the only evidence for Dorian Gray effects in the youth panel was found in the model that tested PBP scores for girls. High PBP scores in childhood and puberty predicted the development of a marginally more baby-faced adolescent appearance as well as a more attractive appearance. However, these effects may be an artifact of missing data substitution because they were not obtained in the simple regression analysis performed on a subset of participants who had no missing data in childhood and adolescence (both β = .09, ps > .10). For this reason, these effects will not be discussed further.

Adult panel. In the adult panel, support for Dorian Gray effects was present in the model that tested PAP scores for women, as shown in Table 1. High PAP scores during adolescence and the 30s predicted the development of a greater attractiveness in the 50s as well as greater babynesseness, with early levels of each appearance variable controlled. Whereas the effect of early PAP scores on later attractiveness held true with later babynesseness controlled (β = .19, p < .05), the effect on later babynesseness was eliminated with later attractiveness controlled (β = .12, p > .20). Because the effect on later babynesseness appears to be due to the covariation between babynesseness and attractiveness for women in their 50s (r(94) = .40, p < .01), it will not be discussed further.

To clarify the specific components of an early prototypical attractive personality that were responsible for its effect on later attractiveness, regression analyses were performed that predicted women's attractiveness in their 50s from early appearance and early PAP scores with four additional trait measures entered into the regression equation. The results revealed that the effect of early PAP scores on later attractiveness was eliminated when early Sociability was entered into the regression equation (β = .14, p > .30), indicating that the greater sociability of women with a more prototypically attractive personality influenced their later attractiveness. On the other hand, the effect of early PAP scores on later attractiveness remained unchanged when early Hostility or Dominance or Competence was entered (β = .30, .27, and .26, respectively; ps < .01, .01, and .05, respectively), which indicated that neither the lower hostility nor the higher dominance and competence of women with a more prototypically attractive personality influenced their later attractiveness. The trait measures used in these analyses were a composite of the CPI and Q-sort scores in the 30s and Q-sort scales at adolescence when the CPI was not administered.

Mediators of the Dorian Gray effect. Two specific facial qualities met the two conditions necessary to be considered as potential mediators of the Dorian Gray effect reported for adult women (cf. Baron & Kenny, 1986, p. 1177). Specifically, higher PAP scores at earlier ages predicted later smiling (β = .23, p < .05) and later makeup (β = .25, p < .05). (Early attractiveness was controlled in these regressions so that they would parallel the final regression equation testing for mediation.) Smiling and amount of makeup also predicted higher levels of attractiveness in the 50s when entered as predictors together with early PAP scores and early attractiveness (for smiling: β = .33, p < .001; for makeup: β = .66, p < .0001). Finally, the Dorian Gray effect of early PAP scores on later attractiveness, controlling for early attractiveness, lost significance when later makeup was controlled (β = .10, p = .18). On the other hand, it remained significant when later smiling was controlled (β = .19, p < .05). It thus appears that the effect of early PAP scores on women's later attractiveness was due to the greater use of makeup by women with more prototypically attractive personalities.

Accuracy of Facial Stereotypes

The Dorian Gray effect for women and the self-fulfilling prophecy effect for men provide reason to think that the attractiveness stereotype may become more accurate for men and women from adolescence to their 50s. The self-defeating prophecy effect for boys suggests that the baby-face stereotype may become less accurate for boys from childhood to adolescence. To examine the accu-
racy of appearance stereotypes, correlations between perceived attractiveness and PAP scores and correlations between perceived babyfaceness and PBP scores were performed for a cross-sectional sample of male and female participants at each age. The magnitude of the significant correlations between personality traits and attractiveness or babyfaceness reported later in this article were virtually unchanged, with the other appearance variable controlled.

There was only weak evidence for the accuracy of appearance stereotypes, with only 3 of 20 possible correlations showing even marginal significance: more baby-faced adolescent boys had a less prototypically baby-faced personality, \( r(84) = -0.26, p < 0.05 \), and more attractive women had more prototypically attractive personalities in their 30s, \( r(79) = 0.19, p < 0.10 \), and 50s, \( r(113) = 0.18, p < 0.10 \). An inspection of the mean PBP z-scores for adolescent boys in the top, bottom, and middle two quartiles of babyfaceness confirmed that it was the highly baby-faced boys who accounted for the significant negative correlation between appearance and personality at adolescence. Boys who scored in the top quartile on babyfaceness had significantly lower PBP scores (\( M = -0.59 \)) than those in the middle two quartiles (\( M = 0.16 \)), \( t(63) = 3.08, p < 0.01 \), and the latter did not differ significantly from those in the bottom quartile (\( M = 0.29 \)), \( t < 1 \). An inspection of the mean PAP z-scores for women in the top, bottom, and middle two quartiles of attractiveness revealed that it was highly attractive women who accounted for the marginally significant correlation between appearance and personality in the 50s. Women in the top quartile on attractiveness in their 50s had marginally higher PAP scores (\( M = 0.35 \)) than those in the middle two quartiles (\( M = -0.03 \)), \( t(83) = 1.71, p < 0.10 \), and the latter did not differ significantly from those in the bottom quartile (\( M = -0.18 \)), \( t < 1 \). On the other hand, both highly attractive and highly unattractive women contributed to the overall relationship between attractiveness and PAP scores in the 50s, with neither group differing significantly in personality from those in the middle of the distribution (both \( t < 1 \)).

**Effects of Appearance Stability**

**And Personality Stability on the Accuracy of Facial Stereotypes**

Two sets of exploratory analyses considered the possibilities that (a) self-fulfilling and self-defeating prophecy effects would yield significant correlations between appearance and personality for participants whose appearance was stable across time but not for those who were perceived as highly attractive or highly baby-faced at some ages and less so at others, and (b) Dorian Gray effects would yield significant correlations between appearance and personality for participants whose personality was stable across time but not for those whose personality was unstable. Although the results provided some support for these hypotheses, they should be interpreted with caution because the sample sizes are very small when participants are broken down by both gender and stability.

Participants within each panel were designated as stable in attractiveness if they remained in the same third of the perceived attractiveness distribution from their early years (z score averaged across Time 1 and Time 2 within each panel) to their later years (z score at Time 3 for each panel). Those within each panel who moved from one third of the distribution to another from the early to the later years were designated as unstable in attractiveness. The same procedure was used to designate participants in each panel who were stable and unstable in babyfaceness, PAP scores, and PBP scores.

Men in their 50s who had been stable in attractiveness since adolescence showed a significantly greater positive correlation between attractiveness and PAP scores, \( r(39) = 0.25 \), than those who had been unstable in attractiveness, \( r(44) = -0.20, z = 2.01, p < 0.05 \). There also was a significantly greater correlation between attractiveness and PAP scores for women in their 50s whose PAP scores had been stable across time, \( r(44) = 0.42 \), than for those whose PAP scores had been unstable, \( r(52) = 0.00, z = 2.11, p < 0.05 \). Thus, stability of attractiveness contributed to a more accurate attractiveness stereotype for adult men, consistent with a self-fulfilling prophecy effect, and stability of a prototypically attractive personality contributed to an accurate attractiveness stereotype for adult women, consistent with a Dorian Gray effect. None of the stability comparisons were significant for the youth panel or for babyfaceness and PBP scores.

**Discussion**

The present findings reveal that the relationship between appearance and personality over the life span is complex, varying as a function of age, gender, and the appearance quality under investigation. For adult men, there was a marginally significant lagged effect of attractiveness on personality, consistent with a self-fulfilling prophecy effect. More specifically, high attractiveness in adolescence and the 30s predicted a personality in the 50s that was more like the prototypical attractive person, an effect that was due to greater sociability and less hostility. An influence of early appearance on men’s later personality was also recently reported by Zebrowitz et al. (1996). In that study, it was found that men with an honest-looking appearance early in life became more honest later on. Although attractiveness was one of several factors that contributed to an honest-looking appearance, it did not mediate the effect of looking honest on subsequent honesty. It thus appears that an honest...
appearance and an attractive appearance can each produce self-fulfilling prophecy effects for men, with the former increasing honesty and the latter increasing sociability and diminishing hostility.

For boys, unlike men, early attractiveness did not influence personality development. Having a baby face, however, did influence personality development. Interestingly, the lagged effect of early babyfaceness on adolescent personality was opposite to that predicted by a self-fulfilling prophecy effect. High babyfaceness in childhood and puberty predicted an adolescent personality that was less like the prototypical baby-faced person, an effect that was due to more assertiveness and more hostility. These results suggest a compensation effect: Baby-faced boys may manifest personality traits that counteract the expectation that they will be babyish, thereby yielding a self-defeating prophecy (e.g., Jussim, 1991). Such compensation effects are consistent with evidence that people try to disconfirm expectations that are inconsistent with their self-concepts (e.g., Swann, 1984) and that obese individuals can behave in ways that compensate for others’ negative reactions to their appearance (Miller et al., 1995). A compensatory reaction to the baby-face stereotype was not shown by adult men. Babyfaceness at adolescence and the 30s had no effect on personality in the 50s, which suggests that men are less motivated than boys to compensate for infantlike expectations elicited by appearance.

Whereas the results for male participants revealed lagged effects of appearance on personality that were consistent with self-defeating prophecies in childhood and self-fulfilling prophecies in adulthood, there were no significant lagged effects of appearance for female participants. The absence of a self-defeating prophecy effect of babyfaceness in childhood is not so surprising. Baby-faced girls may be less motivated than boys to compensate for the expectancy that they will be childlike because childlike traits are more compatible with a feminine identity than a masculine one (cf. Friedman & Zebrowitz, 1992). The divergent effects of appearance stereotypes on the personality development of boys and girls recalls earlier evidence that the stability of personality across age depends on the sex-role congruence of the traits in question (e.g., Kagan & Moss, 1962).

The absence of lagged effects of attractiveness consistent with a self-fulfilling prophecy for adult women is surprising, particularly given the widespread assumption that attractiveness has more social consequences for women than for men. It is possible that this failure to confirm predictions actually derives from those greater social consequences. In particular, it may reflect a countervailing compensation effect for unattractive women who attempt to defeat the expectation that they will be unsociable, yielding no net effect of attractiveness on women’s personality development. Consistent with this suggestion, other investigators have reported evidence of stereotype-defying compensation effects among physically unattractive women and girls (Dion & Stein, 1978; Reis, Spiegel, Kernis, Wezlek, & Perri, 1982).

What the lagged effects for female subjects did reveal was a Dorian Gray effect, whereby facial appearance shows an influence of earlier personality. Women who had a more prototypical attractive personality in adolescence and the 30s, with greater sociability being the key factor, developed higher attractiveness in their 50s. An examination of various facial qualities that might account for the increased attractiveness of these women, including smiling, facial lines, hairstyles, and youthfulness, revealed that the Dorian Gray effect was explained only by their tendency to use more makeup in their later years. This mediation can account for the fact that the Dorian Gray effect was limited to female participants. It also indicates that rather than reflecting a direct effect of personality on structural facial qualities, such as laugh lines or wrinkles, it reflected an indirect influence of personality on appearance via the selection of grooming aids. An influence of early personality on later appearance for women but not for men has also been revealed in a recently reported artefact effect: More dishonesty in early years predicted a more honest and more attractive appearance in later years only for women (Zebrowitz et al., 1996). These sex differences may reflect greater ability of women to manipulate their appearance or greater motivation to do so because they have less power to achieve their social goals through other means.

The Dorian Gray effect observed for early sociability in this study and the artefact effect previously observed for early dishonesty may appear to be inconsistent because the socially desirable trait of sociability and the undesirable trait of dishonesty both foster later attractiveness. However, it should be noted that the early sociability and honesty of participants in these two studies were not positively correlated. Indeed, there was a slight negative correlation for females, which is consistent with other recent evidence on the correlates of dishonesty. In particular, Kashy and DePaulo (1996) found that more sociable people reported lying more in social interactions. Those who lied more also scored higher on a measure of public self-consciousness, which assesses the extent to which people are concerned about their appearance and the impressions they convey, and this effect was most pronounced for women. It thus appears that the traits of sociability and dishonesty are both conducive to the development of a more attractive appearance in women, perhaps because women with either of these traits are more concerned with impression management than with their less sociable or more honest counterparts. Whereas sociability consequently
fosters the development of an appearance that is congruent with social expectations—a Dorian Gray effect—dishonesty fosters the development of an appearance that is incongruent with expectations—an artifice effect.

The present findings have some interesting implications for the accuracy of appearance stereotypes, although they should be interpreted with caution due to small sample sizes in analyses performed on subgroups of participants. What the results suggest is that a relationship between appearance and personality may require lagged effects of earlier appearance on later personality or vice versa and that the direction of the relationship may depend on the social desirability of the appearance stereotype. Correlations between participants’ attractiveness and their scores on the PAP scale revealed that the attractiveness stereotype was significantly more accurate for men in their 50s whose attractiveness had been stable since adolescence than for those whose attractiveness had been unstable. The attractiveness stereotype was also more accurate for women in their 50s whose personality had been stable since adolescence than for those whose personality had been unstable. These results are consistent with a self-fulfilling prophecy effect of attractiveness for men and a Dorian Gray effect of a stereotypically attractive personality for women. The fact that these effects were not observed in the youth panel suggests that they take more time to develop than the youth panel encompassed or that they occur only in adulthood.

A very different pattern of results was obtained for the baby-face stereotype. Correlations between participants’ babyfakeness and scores on the PBP scale revealed that the baby-face stereotype was significantly inaccurate for adolescent boys, the group for whom there was a lagged effect of early babyfakeness on later personality. Although this effect did not differ significantly for boys whose appearance had been stable versus unstable, this may reflect the high stability of babyfakeness for all boys at this age (see Table 1). These results suggest that boys who elicit the baby-face stereotype come to reject that stereotype by showing compensatory, disconfirming behaviors. Evidence to support such a process was provided by the finding that it was the most highly baby-faced boys who differed in personality from the others. Whereas the personalities of baby-faced boys contradicted the baby-face stereotype, there were no significant relationships between babyfakeness and personality in adulthood.

It has long been argued that facial attractiveness can produce self-fulfilling prophecy effects on personality development. This study provides the first empirical evidence to support this presumption. At the same time, it highlights the limitations of a focus on self-fulfilling prophecies when considering the relationship between appearance and personality over time. Although the results should be replicated with other samples, they suggest that this relationship may be bidirectional rather than unidirectional and that it may depend not only on the expectations induced by appearance but also on normative expectations for age and gender-appropriate behavior. Age differences in the effects of attractiveness and babyfakeness on the personality development of male participants suggest that the salience and social consequences of attractiveness are stronger in adulthood than in childhood, whereas the reverse is true for babyfakeness. The contrasting direction of these effects further indicates that expectancies for socially desirable attributes are more likely to be fulfilled than expectancies for undesirable ones, which may be defeated. Finally, the present findings indicate that accurate appearance stereotypes sometimes may be explained not only by self-fulfilling prophecy effects but also by Dorian Gray effects, which heretofore have been affirmed only in works of fiction. For example, the greater social skills of attractive women (e.g., Feingold, 1992) may reflect a tendency for women who have such skills to develop their attractiveness as well as a tendency for attractive women to develop those skills.

NOTES

1. The authors would like to thank Avshalom Caspi for suggesting this procedure.

2. z scores were employed in these analyses because averaging raw scores from two age levels could distort subjects’ standing relative to their peers. The regression coefficients of Time 1 and Time 2 indicators on the constructs (i.e., lambda values) were constrained to be equal and freely estimated. At Time 3, in which there were single indicators, the path from the latent variable to the indicator variable was fixed at 1.0, and the error fixed at (1 - reliability) x variance for these indicators (Hayduk, 1987).

3. Separate analyses for male and female participants, rather than multisample analyses with nested comparisons, were performed because we began with no a priori model of sex differences to guide decisions regarding what paths to constrain to be similar or different in multisample analyses.

4. Tests for whether data were missing at random were carried out according to the procedure recommended by Cohen and Cohen (1983). Regressions were constructed so that appearance or personality at Time 3 was predicted first by the early level of that variable (as in the stability model of the structural equation modeling, SEM, analyses) and, on a second step, by a dummy variable representing the presence or absence of early personality or appearance data, respectively (as in the cross-decoding model of the SEM analyses). The change in R^2 produced by adding the dummy variable revealed whether subjects who had data at both early time periods differed on the criterion from those for whom data were missing at either time. Of the 16 regressions performed (4 Criterion Variables x 2 Panels x 2 Genders), only two yielded a significant change in R^2 due to the presence/absence of a predictor variable. Early personality data was not missing at random for girls in the youth panel with respect to the criterion of adolescent attractiveness (Δ R^2 = .03, p < .05). Early appearance data also was not missing at random for girls in the youth panel with respect to the criterion of adolescent Prototypical Baby-Faced Personality (PBP) scores (Δ R^2 = .10, p < .001). It should be noted that in both of these instances, almost all of the missing values were due to data not being collected from a subsample of participants. In all other cases, including all for which significant results are discussed in the following sections, there was no evidence that missing data was
not random. Although advanced techniques for handling missing data in SEM analyses have been developed (e.g., Allison, 1987), these typically use LISREL to fit a limited number of missing data patterns created by attrition over time in a longitudinal data set. In our sample, the missing data problem is not primarily due to attrition, and the resulting patterns are too numerous to be suitably addressed by these techniques.

5. There was also evidence for the stability of measures within the early time period. Attractiveness was stable from Time 1 to Time 2 for males and females in the youth and adult panels (mean r = .57); babyfacedness was stable from Time 1 to Time 2 for boys and girls in the youth panels and for men in the adult panel (mean r = .57). Prototypical Attractive Personality (PAP) scores were stable from Time 1 to Time 2 for males and females in the youth and adult panels (mean r = .46); and PBP scores were stable from Time 1 to Time 2 for boys and girls in the youth panel and for women in the adult panels (mean r = .51). As one would expect, the stability of these measures was somewhat higher when missing data were replaced with the participants' scores at the other early time. (For attractiveness, mean r = .70; for babyfacedness, mean r = .65; for PAP scores, mean r = .65; for PBP scores, mean r = .70). The only measures that did not show stability from Time 1 to Time 2 when no missing data were substituted were babyfacedness for women in the adult panel, r(54) = .09, and PBP scores for men in the adult panel, r(55) = .20. Because the SEM analyses yielded no significant effects for these two measures, the inflated stability that resulted from replacing missing data with the participants' scores at the other early time produced spurious significant effects.

6. Because the personality prototypes were generated with adult targets in mind, it is possible that they failed to match the personalities of children because some of the manifestations of an attractive or baby-faced personality are different at younger ages. (See Asendorpf, 1992, for a discussion of this problem in developmental research.) To explore this possibility, we examined correlations between appearance and additional behavior ratings available for some subjects at child-

REFERENCES


Langlois, J. H. (1986). From the eye of the beholder to behavioral reality: The development of social behaviors and social relations as a function of physical attractiveness. In C. P. Herman, M. P. Zanna, &