School, Studying, and Smarts:

Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009*

Andrei Boutyline
University of Michigan

Alina Arseniev-Koehler
University of California, Los Angeles

Devin J. Cornell
Duke University

Revised draft –
January 10, 2022

Running Head: School, Studying, and Smarts

* We thank Elizabeth Armstrong, Elizabeth Bruch, Erin Cech, Fabiana Silva, Karin Martin, and Mark Mizruchi for reading the manuscript and providing insightful commentary. We also thank attendees at the University of Michigan Measuring and Modeling Culture workshop and the University of Chicago Computational Social Science workshop for their feedback. We are grateful to Yinan Wang and Hannah Simon for their help as undergraduate research assistants. We are also grateful to Jacob Foster and Erica Cartmill at the Diverse Intelligences Summer Institute for bringing us together and providing the opportunity to begin this collaboration. Please direct all correspondence to Andrei Boutyline at Department of Sociology, University of Michigan, Room 3115 LSA Building, 500 S State St, Ann Arbor, MI 48109. Email: aboutyl@umich.edu

ABSTRACT

In this article, we apply computational word embeddings to a 200-million-word corpus of American print media (1930-2009) to examine how education-relevant gender stereotypes changed as women's educational attainment caught up with and eventually surpassed men's. This case presents a rare opportunity to observe how cultural components of the gender system transform alongside the reversal of an important pattern of stratification. We track six stereotypes that prior work linked to academic outcomes. Our results suggest that stereotypes most closely tied to the core stereotypical distinction between women as communal and men as agentic remained unchanged. The other stereotypes we tracked, however, became increasingly gender polarized: as school and studying gained feminine associations, intelligence and unintelligence gained masculine ones. Unexpectedly, we observe that trends in the gender associations of intelligence and studying are near-perfect mirror opposites, suggesting an interrelationship. We use these observations to further elaborate contemporary theoretical accounts of the gender system that argue it persists partly because stereotypes shift to reinterpret social change in terms of a durable hierarchical distinction between women and men.

School, Studying, and Smarts: Gender Stereotypes and Education Across 80 Years of American Print Media, 1930-2009

A large body of recent work illustrates that the ways that girls and boys enact and enforce gender stereotypes powerfully shapes their academic outcomes. For example, ethnographic and survey evidence documents how displaying an affinity for school often violates boys' norms of masculinity while fitting girls' norms of femininity (Legewie and DiPrete 2012; Morris 2008; Musto 2019). Meanwhile, girls are dissuaded from disciplines which are perceived to require brilliance, such as physical sciences, engineering, and mathematics (Leslie et al. 2015). Many social and behavioral skills required for overall academic success (such as attentiveness, politeness and diligence) also appear to fit hegemonic feminine stereotypes, while problem behaviors like disobedience and aggression fit hegemonic masculine stereotypes (Jackson and Dempster 2009; Marrs 2016; Ridgeway 2011; Ridgeway and Correll 2004).

This article leverages eight decades of U.S. print media (1930-2009) to track these gender stereotypes at scale. Despite the wealth of recent work illustrating the importance of gender stereotypes for academic outcomes, we know little about whether and how these stereotypes changed as gender differences in educational attainment drastically transformed since the 1940s.[1] While men reached higher levels of education than women for much of the 20th century, in recent decades this gender gap reversed, with women's educational attainment increasingly surpassing men's (Goldin, Katz, and Kuziemko 2006). Because gender stereotypes are partly the product of observed material differences,[2] they respond to social change (Koenig and Eagly 2014; Ridgeway et al. 2009; Seguino 2007). Therefore, the dramatic changes in women's

[1] Our corpus covers the years 1930 to 2009, and we partition it into 20-year-wide sliding windows (see details below). Our earliest window is thus centered on 1940, and latest on 2000.
[2] E.g., in women's versus men's educational achievement.

educational achievement throughout the 20th century may have been accompanied by changes in the gender associations of school, studiousness, intelligence, and classroom behaviors, among other concepts.

These changes in the educational landscape present a rare opportunity to examine how stereotypes change alongside the reversal of a core pattern of stratification. Our longitudinal examination enables us to observe which stereotypes transform along with corresponding material changes, and which remain despite them. We use these observations to contribute to theoretical discussions about the persistence of the gender system (Ridgeway 2011; Ridgeway and Correll 2004). Additionally, given the powerful role of education in stratification and the salience of gender in the classroom, our observations of these stereotypes are important contributions in their own right. These observations may enable scholars of gender in contemporary classrooms to contextualize their observations within broader historical trends. They may also inform future scholarship in the sociology of education on the gender gap reversal, which has implicitly assumed that these gender stereotypes are an important but unchanging contributor to the gender gap.

As Ridgeway (2011) notes, the lack of systematic longitudinal data on gender stereotypes has hampered previous efforts to understand their historical trajectories. To overcome this problem, we examine a 200-million-word corpus of American print media (1930-2009) with Word2Vec (Mikolov et al. 2013). Word2vec is a computational text analysis method that has found wide application across many disciplines, but has only recently begun to see sociological uses (e.g., Kozlowski, Taddy, and Evans 2019). Our approach lets us estimate how often each of over 10,000 words in our corpus vocabulary occurs in feminine versus masculine contexts at

different time points in our study period. We use this to track changes in stereotypes that prior work has linked to gender differences in contemporary academic outcomes.

The rest of our manuscript proceeds as follows. We first survey existing work to identify relevant gender stereotypes and derive predictions about their longitudinal changes. We next detail our methodological approach, which uses bootstrapped Word2Vec embeddings. We then introduce our corpus and operationalize the gender stereotypes via keyword scales. We then detail our findings. We conclude by discussing the implications of our results for broader theoretical understandings of stereotypes within the gender system.

GENDER STEREOTYPES AND EDUCATION

Gender stereotypes are widespread beliefs about what men and women are (or should be) typically like (Fiske et al. 2002; Ridgeway et al. 2009). They provide blueprints for how men and women should act in social situations, such as how to behave in the classroom. Stereotypes are socially reinforced: when individuals deviate from gender stereotypes, they may be penalized by others—especially when their behaviors are prescriptive for the *opposite* gender (Prentice and Carranza 2002; Ridgeway 2011:59; Ruble, Martin, and Berenbaum 2007). Stereotypes also provide default assumptions that "fill in the gaps" in what is ambiguously perceived or imperfectly remembered with culturally established expectations. Thus, while stereotypes are partly based in objective material differences between social groups, they also amplify and perpetuate those differences, thereby contributing to the persistence of the gender system (Ridgeway 2011:159).

In the next section, we review scholarship on gender stereotypes in education to identify six gender stereotypes strongly linked to academic outcomes and develop predictions about their longitudinal changes. This includes both stereotypes specific to education, and those tied to the

core stereotypical distinctions between men and women. This mixture of specific and general stereotypes makes education an especially informative domain for understanding dynamics of stereotypes within the gender system.

Socio-behavioral skills and problem behaviors

We begin with stereotypes for two sets of classroom behaviors. First, academic success requires a range of stereotypically feminine (i) "socio-behavioral skills"—a label that education researchers apply to a range of positive behaviors, traits, and competencies including attentiveness, responsibility, communication, cooperation, helpfulness, and respect for teachers (e.g., Downey and Vogt Yuan 2005). Socio-behavioral skills affect student outcomes both because they are immediately critical to students' learning in the classroom environment, and because they are rewarded by teachers and educational institutions. Girls are socialized into these skills from a young age (DiPrete and Jennings 2012). At the same time, these skills may be stigmatized for boys because they are considered effeminate (e.g., Cohen 1998; Epstein 1998). Conversely, (ii) "problem behaviors" such as interrupting, fighting, disobeying teachers, and behaving aggressively are stereotypical of boys, and hinder their academic outcomes (Jackson and Dempster 2009; Morris 2008; Musto 2019).

Previous work has not directly investigated longitudinal changes in stereotypes around classroom behaviors. However, we note that stereotypes of socio-behavioral skills and problem behaviors appear to largely overlap with the stereotypes of *communality* and *agency*, which existing work has argued are the core dimensions by which attributes are distinguished as feminine versus masculine (Eagly 1987; Eagly and Wood 1999; Ridgeway 2011).[3] Scholars link

[3] Competence is also sometimes included as a component of agency; however, like Ridgeway (2011:168–69) and others, we view competence as a conceptually distinct dimension of masculinity.

these core gender stereotypes to the distribution of power in society, arguing that, as women occupy lower-status and caregiving roles, femininity comes to be associated with the supportive, agreeable, and conflict-avoidant behavior more characteristic of those roles; and, similarly, as men occupy higher-status, higher-power positions, masculinity comes to be associated with the agentic, forceful, dominant behavior more characteristic of those roles (Eagly and Wood 1999; Ridgeway 2011; Wood and Eagly 2002). Empirical work confirms that the stereotype of women as communal (Spence and Buckner 2000) and men as forcefully agentic are indeed highly persistent across time (Cejka and Eagly 1999; Diekman and Eagly 2000; Haines, Deaux, and Lofaro 2016; Koenig and Eagly 2005).[4]

Examination of the survey scales existing work uses to track these gender stereotypes suggests that communality and agency have substantial intersections with socio-behavioral skills and problem behaviors, respectively. For example, Eagly and Steffen's widely-used communality scale taps perceptions of people as "kind, helpful, understanding, warm, aware of others' feelings, and […] able to devote self to others" (Eagly and Steffen 1984:738)—traits that, in a classroom context, would describe a well-behaved student. In much the same way, problem behaviors map closely onto forcefully agentic attributes which characterize stereotypical masculinity. For example, Eagly and Steffen's agentic scale includes items for "active, not easily influenced, aggressive, independent, dominant, self-confident, [and] competitive" (Eagly and Steffen 1984:738)—traits which, within the academic setting, would describe a difficult student. The stereotypes of socio-behavioral skills and problem behaviors may thus be closely related to the core stereotypical distinction between women as communal and men as agentic.

[4] But see Leupetow et al. (2001).

Identifying this relationship between classroom gender stereotypes and core stereotypical distinctions lets us use scholarship on the latter to make predictions about the former. As Ridgeway (2011) and others argue, the core stereotypes of women as communal and men as agentic may be particularly durable because they are reinforced by the persistent distribution of power in society. Indeed, a large body of older cross-sectional studies confirms that gender differences in socio-behavioral skills and problem behaviors were present in schooling throughout the 20th century (Blatz and Bott 1927; Hartley 1959; Hayes 1943; Meyer and Thompson 1956; Peltier 1968; Tuddenham 1952). Thus, we expect that the gender stereotypes of socio-behavioral skills and problem behaviors have likely remained broadly unchanged throughout our time period.

Schooling and School Effort

Contemporary work shows that adolescents view a wholehearted embrace of school as stereotypically feminine (Heyder and Kessels 2013; Jackson 2003; Kessels et al. 2014; Morris 2008; Warrington, Younger, and Williams 2000). Since masculinity is partly constructed in opposition to femininity, an affinity for school is often stigmatizing for boys (Adler, Kleiss, and Adler 1992; Epstein 1998; Pascoe 2007). An even stronger stigma centers around effort put into schoolwork. While studiousness may be rewarded for girls (Adler et al. 1992; Entwisle, Alexander, and Olson 2007), boys who are seen as putting in substantial effort into school can be ostracized or teased by other boys as effeminate or queer (Epstein 1998; Morris 2008; Pascoe 2007). We will thus track the gender associations of (iii) schooling and (iv) academic effort.

As scholarship on social role theory and status construction theory demonstrates, gender beliefs arise partly from observations of material differences between the social positions of men and women (Eagly 1987; Ridgeway et al. 2009; Ridgeway and Correll 2004). We thus conjecture

that radical changes in the educational attainment of American women and men over the 20th century likely led to shifts in the cultural understanding of gendered academic potential. While men made up the majority of bachelor's degree recipients for much of the 20th century, by the 1980s, women's educational attainment caught up to (and eventually surpassed) men's across all racial groups (DiPrete and Buchmann 2013:27, 39). Given the salience, scope, and durability of this transformation, we expect that the stereotypical image of a student—and especially a high-achieving student—may have also feminized across this period.

Scholarship on the history of education, and older cross-sectional studies, offer support for this prediction. Well into the 20th century, academic attainment was perceived as a "man's pursuit" (Cohen 1998; Solomon 1985). For example, Harley's (1959) interview study of elementary schoolchildren found that boys believed academic success was more important for boys than girls. This was also reflected in college aspirations, with various older studies finding that boys were more likely than girls to aspire to attend college and to list college as a motivation to get good grades (Bordua 1960; Hartley 1959; Hicks and Hayes 1938).[5] In contrast with these older results, contemporary work shows that more girls than boys now aspire to attend college—a change that chronologically coincided with the reversal in the educational attainment gap (Fortin, Oreopoulos, and Phipps 2015). This evidence is consistent with a changing gender landscape where schooling and academic achievement came to be gradually seen as more characteristic of women than men.

Since the 1940s, various social changes enabled and encouraged increases in women's academic engagement. The continuous growth of women's labor force participation between the

[5] Clark (1967) instead suggested that, in elementary school, girls had higher academic aspirations than boys, but this difference reversed in later grades.

1940s and 1990s increased the rewards women received from having a college degree (Juhn and Potter 2006). At the same time, a variety of sweeping structural changes opened educational opportunities for women—such as the increased availability of birth control in the 1960s, the rise of second-wave feminism in the 1960s and 1970s, and the passage of Title IX in 1972—and women took advantage of these newfound opportunities (Goldin et al. 2006). For example, a study of 9th graders found that between 1935 and 1953 girls caught up to and surpassed boys in the amount they studied in high school—a change that continued growing throughout the 1950s (Jones 1960). Various longitudinal surveys also show that, beginning in the 1950s, girls' enrollment in high school math and science classes caught up to and then surpassed boys'. More broadly, girls began to take more advanced courses and more college preparatory courses than boys (DiPrete and Buchmann 2013; Goldin et al. 2006). Thus, we predict that not only did schooling become seen as more feminine, but, alongside this, academic effort also gained feminine associations.

Intelligence and Unintelligence

Recent work also points to the importance of two additional stereotypes—(v) intelligence and (vi) unintelligence. While studiousness is the stereotypic feminine route to academic success, stereotypes for masculine educational success instead suggest that boys should *effortlessly* do well in school, succeeding from sheer intelligence (Cohen 1998:26; Jackson and Dempster 2009; Morris 2008; Musto 2019). Effortlessness, however, is rarely a viable strategy for academic success, thus potentially leading boys to struggle in school. Meanwhile, scholars note that a noticeable lack of intellectual ability—i.e., *unintelligence*—is also especially undesirable for boys. Thus, some boys may excessively signal their own lack of effort to shift attention away from their intellectual ability as a cause for their academic failure (Adler et al. 1992; Jackson

2002). Similarly, some boys may also "overdo" other displays of masculinity—namely, problem behaviors. Thus, as Musto (2019) argues, boys may have two routes for enacting masculinity in school: being "brilliant" or "bad".

Gender stereotypes of intelligence may also have substantial negative effects for girls (Bian, Leslie, and Cimpian 2018; Leslie et al. 2015; Storage et al. 2016). Importantly, girls are less likely to pursue school subjects (and ultimately careers) that are believed to require exceptional intelligence—namely, science, technology, engineering or mathematics, i.e., STEM (Leslie et al. 2015; Storage et al. 2016). Further, as Cohen (1998) notes, the stereotype of girls as hard-working also inadvertently undermines the perception of girls' intellect: the harder they work, the less "evidence" there is for their innate intellectual potential (see also Dweck 2006). Particularly in STEM subjects, girls' success is often attributed to effort rather than intelligence (e.g., Espinoza, Arêas, and Arms-Chavez 2014).

We can gain insights about changes in gender stereotypes of intelligence from a combination of historical work on discourses concerning gender and intelligence and longitudinal studies of popular stereotypes. Because this literature does not study gender stereotypes of *unintelligence* separately from intelligence, we do not make separate predictions for unintelligence here. Early 20th-century discourses viewed males as intellectually superior to females, but did not view exceptional intelligence as *masculine*: rather, the dominant view of a gifted child imagined a weak, eccentric, effeminate, and possibly gay male (Elfenbein 1999; Fox 1968; Hegarty 2007). Much research on giftedness and intellect throughout the early- to mid-20th century was designed in part to combat these negative stereotypes, instead arguing that intellect is a sign of health, success, and, in the case of boys, masculinity and virility (Hegarty

2007; Jolly 2008; Warne 2019). This intellectual history suggests that the association of intelligence and masculinity may have grown stronger during the 20th century.

Empirical work on gender stereotypes presents a more mixed picture. One meta-analysis synthesizing results of sixteen cross-sectional surveys suggested that, between 1946 and 2018, the likelihood of respondents considering men as more intelligent than women fell across time. Instead, respondents became more likely to state that women are more intelligent than men, or that women and men are equally intelligent (Eagly et al. 2020). This finding, however, conflicts with an abundance of contemporary literature that finds intelligence is more closely associated with stereotypes of men rather than women (Bian, Leslie, and Cimpian 2017; Bian et al. 2018; Leslie et al. 2015; Prentice and Carranza 2002; Storage et al. 2016).[6] Because different strands of scholarship point to conflicting conclusions, we make no predictions regarding changes in the gender stereotypes of intelligence.

Predictions

In the previous sections, we identified six gender stereotypes relevant to academic outcomes: socio-behavioral skills, problem behaviors, schooling, school effort, high intelligence, and low intelligence. We summarize our predictions for their changing gender associations in Table 1.

[Table 1 about here]

DATA

To investigate these stereotypes across time, we use the Corpus of Historical American English (COHA) for the years 1930 to 2009 (Davies 2012). COHA is commonly used in historical

[6] This discrepancy may be due to measurement. Items analyzed by Eagly et al. (2020) ask the respondent to *explicitly* compare men's and women's intelligence—a format known to produce substantial social desirability bias (Klonis, Plant, and Devine 2005). In contrast, most of the second body of work uses more subtle measures (e.g., vignettes), which may reduce socially desirable responding (Hughes and Huby 2012). The trend recorded by Eagly et al. may thus reflect the increasing social unacceptability of sexist attitudes about intelligence.

linguistics to study American English language change. It consists of a purposive sample of full-text books, magazine articles, and newspaper articles that is balanced across the decades by medium and Library of Congress category. It is designed to capture a consistently wide variety of American print media: roughly 50% of the corpus is fiction, with the remainder sourced from popular magazines (e.g., *TIME*), newspapers (e.g., *New York Time*s), and a small number of nonfiction books.

While COHA was designed and validated for studying language change (Davies 2012) and has been used in other studies using Word2Vec (e.g., Garg et al. 2018), its validity as a record of popular culture is less well established. To verify that it provides a good window into the kinds of works that were widely read by Americans within each time period, we validated its contents against historical Publisher's Weekly lists of top-selling fiction. We randomly selected 30% of the titles for each decade and examined whether the title and the author were in the corpus. We found that, on average, COHA contains books by 75% (s.d. = 9.5%) of the best-selling authors within each decade, and 30% (s.d. = 10%) of the specific best-selling novels. COHA thus indeed contains a collection of some of the works most widely read by Americans this century. Many Americans have been exposed to the gendered portrayals in this corpus.

To study longitudinal change, we segmented the corpus into overlapping 20-year windows, which we "slide" by 5 years at a time. Each window is centered on January 1 of a focal year, starting with 1940. This yields the following series of 13 windows (written as *focal year: [range]*):

1940 : [1930,1949], 1945 : [1935,1954], . . . , 2000: [1990,2009]

In the remainder of this manuscript, we often use the focal year as shorthand for the full window.

METHODS

To track our six stereotypes across time, we estimated a series of bootstrapped Word2Vec models from our thirteen sliding COHA windows (1940, 1945, … 2000). We detail this process here, beginning with a brief overview of Word2Vec for readers who are unfamiliar with this method (for a detailed sociological introduction, see Kozlowski et al. 2019).

Word2Vec is a machine-learning algorithm that models the meaning of words by locating them relative to one another in a high-dimensional vector space (i.e., by representing each word as an array of n coordinates; in our case, $n = 200$.) To map words to vectors in a 200-dimensional space, the Word2Vec algorithm estimates parameters of a model such that, given a target word, it can predict which words appear near it in the corpus—i.e., its "context windows." In an estimated model, semantically or syntactically similar words are located closer to each other than dissimilar words (as measured by cosine similarity). Different areas of the embedding space thus represent different contexts. The words with the highest proximity are those that (i) frequently co-occur with one another, or (ii) or can be used interchangeably, even though they may rarely be used together (e.g., synonyms) (Mrksic et al. 2016). Together, the collection of these word vectors forms a *word embedding*.

Measuring Gender Stereotypes with Word2Vec

Existing work demonstrates that vector arithmetic can be used to measure cultural associations of terms in embedding models (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018; Grand et al. 2018; Jones et al. 2020; Joseph and Morgan 2020; Kozlowski et al. 2019). Following these approaches, we (i) "*extract*" a gender axis from each embedding using sets of keywords, and then (ii) estimate the gender associations of each keyword of interest by "*projecting*" it onto this axis.

In embedding space, word vectors are more proximal if they correspond to words that tend to be used in similar contexts—i.e., they tend to co-occur with the same sets of words. To differentiate words that occur in feminine versus masculine contexts, we *extract* a gender axis by identifying pairs of *anchor terms* that differ primarily in gender (e.g., 'girl', 'boy'). Within each pair, we subtract the vector corresponding to the masculine word from the vector corresponding to the feminine word (e.g., 'she' – 'he'). We average the resulting difference vectors to estimate a single vector corresponding to a gender axis pointing from masculinity (-) to femininity (+). We used the following anchor term pairs:

('she', 'he'), ('her', 'him'), ('her', 'his'), ('girl', 'boy'), ('girls', 'boys'), ('daughter', 'son'), ('daughters', 'sons'), ('sister', 'brother'), ('sisters', 'brothers'), ('woman', 'man'), ('women', 'men'), ('mother', 'father'), ('mothers', 'fathers'), and ('female', 'male').

To evaluate the relative extent to which a word appears in feminine- or masculine-gendered contexts, we *project* it onto the gender axis by computing the cosine similarity between the gender axis vector and the word vector. This measures the extent to which it is more likely to be used around feminine versus masculine words.[7]

An abundance of prior work validates this approach to measuring the meanings of terms found in print media (see below). However, an important caveat is in order: this approach measures word usage in texts rather than conceptual associations in authors' minds. An author could conceivably use a characteristic to describe primarily men or women without herself holding a stereotype of this characteristic as feminine or masculine. For example, prior scholarship showed that the gendering of occupations in popular print media correlates with

[7] More precisely, this reflects both *first-order* and *second-order contexts* of the anchors. Words we identify as feminine consist of (i) "first-order" feminine words that co-occur with feminine anchor terms; and (ii) "second-order" feminine words that co-occur with first-order feminine words (and conversely for masculine).

those occupations' gender composition in historical Census data (Garg et al. 2018). The authors may have thus conceivably been reflecting their observed social reality rather than following internalized stereotypes of femininity or masculinity.

However, even if this were the case, the depictions of gender in popular media would still have important consequences for popular stereotypes. An abundance of scholarship suggests that mass media play a key role in shaping widespread perceptions of the social world (Bandura 2001; Gamson et al. 1992; see also Cultivation theory, e.g., Gerbner et al. 2002; Potter 2014). Prior work has also empirically shown that exposure to mass media depictions of social groups can influence the audience's stereotypes and attitudes across many domains (e.g., Dixon 2006; Grabe, Ward, and Hyde 2008; Kahn 1994). Since the corpus we use here contains some of the most popular books published during each of the time periods (see Data section above), large segments of the population have been exposed to the portrayals of men and women found in this corpus. The gender stereotypes evoked by these texts are thus sociologically important.

Indeed, a substantial body of validation studies shows that word embedding-based estimates of the gendered usage of terms in large corpora (i.e., gender projections) track hegemonic stereotypes in the population (e.g., Bolukbasi et al. 2016; Caliskan et al. 2017; Grand et al. 2018; Joseph and Morgan 2020; Kozlowski et al. 2019). For example, Caliskan et al.'s (2017) well-known validation study shows that these gender projections corresponds to human participants' implicit gender associations of these same terms as measured by the Implicit Association Test. Previous work has also repeatedly shown that gender projections estimated by this method closely match the gender stereotypes reported by survey respondents (Grand et al.

2018; Joseph and Morgan 2020; Kozlowski et al. 2019).[8] Word embeddings trained on popular media thus offer proxies for widely held gender stereotypes. Following this literature, we occasionally refer to these projections as gender "associations."

To account for potential differences in the overall gendering of language across the decades (for example, due to shifts away from using "man" to mean "human"), we project each of over 10,000 unique words that comprise each window's lexicon onto the gender axis, and then z-score each window's set of resulting cosine similarities to calculate our final projection scores. A projection score of 1 thus indicates that a word is 1 standard deviation more feminine than the average word in the vocabulary in that time window, whereas -1 indicates that it is one standard deviation more masculine.[9]

Constructing the Keyword Scales

We used multi-word scales to measure the gender associations of each of the six education-relevant constructs we identified above: (i) socio-behavioral skills; (ii) problem behaviors; (iii) schooling; (iv) studying (school effort); (v) high intelligence; and (vi) low intelligence. Our use of scales serves two purposes: it reduces measurement error stemming from eccentric changes to any one term's usage, and it allows us to capture a variety of different dimensions associated with each concept.

[8] This work has similarly validated embedding-based estimates of terms' associations with intelligence and unintelligence (Grand et al. 2018).

[9] Some words in the corpus are present in some decades but absent in others. Because each window's baseline is calculated using all the words in that window's vocabulary, these differences could conceivably produce an artifact where estimated changes in gendering are actually due to changes in the vocabulary. To ascertain that this is not happening, we calculated an alternate baseline by first subsetting each window's vocabulary to only those words that are present in every window, and then repeating the above z-scoring process with this constant vocabulary. This yielded substantively identical results to our main analyses.

To select the keywords, we identified candidate terms from academic and primary literature, and then expanded that list using modern-day thesauruses and the *Historical Thesaurus of English* (Kay et al. 2020). We then followed two filtering steps to ensure term quality: we (1) removed words that occurred less than 100 times in each corpus time window, and (ii) read corpus excerpts to manually verify that the candidate words are used in relevant ways. To estimate the gendering of each scale within a given corpus window, we projected the scale keywords that met these filter criteria onto that window's gender axis and averaged the resulting scores. See Appendix A for further details of this process and our specific reasoning behind each set of keywords.

Bootstrapping

We performed all of our statistical tests using bootstrapping (Efron 1979). Bootstrapping is a well-established nonparametric inferential statistical technique that works by resampling the observed data[10] to examine how estimates vary in response to sampling error. We drew 200 bootstrapped resamples of the documents that make up our corpus to create bootstrapped versions of each of our embedding models (13 corpus windows X 200 resamples = 2600 bootstrapped embeddings). We then recalculated all our estimates for each bootstrapped resample. The *p*-values we report in this manuscript reflect the reliability of our estimates across these resamples (with two-sided tests.) So, for example, if we report that "studying" had significant feminine associations in our last corpus window at $p < 0.01$—or, equivalently, "was significantly feminine at $p < 0.01$"—this means that we estimated the gender associations of

[10] I.e., sampling again with replacement from our original sample.

"studying" across all 200 bootstrapped embeddings for that window and found that it was more feminine than the average word in at least 198 of them.[11]

In addition to variation due to sampling, Word2Vec embeddings also vary due to estimation error. Word2Vec is a stochastic algorithm, and the word embedding models it trains differ across estimation runs.[12] To increase the reliability of our results, we adopt a machine learning technique called "bootstrap aggregating" or "bagging" (Breiman 1996): to arrive at our primary estimate of each term's or scale's gendering, we average its projection score across all 200 bootstrapped embeddings. This greatly reduces the impact of estimation error and lessens the effects of outlier observations. The correlations between scales we report here are also between these bagged estimates.

RESULTS

To aid exposition, we analyze our scales in a different order than we introduced them: (i) schooling, (ii) socio-behavioral skills, (iii) behavioral problems, (iv) school effort, (v) intelligence, and finally (vi) unintelligence.

Schooling

Our schooling scale combines terms for students, educational institutions, classrooms, and academic achievement. As the results in Table 2 indicate, in the most recent time window (centered on 2000), this scale was roughly one standard deviation more feminine (mean = 0.98, $p < 0.01$) than the average word in the vocabulary. This is consistent with our expectations based on recent work.

[11] Similarly, if we report that trends for "school effort" and "intelligence" were significantly negatively correlated at $p < 0.01$, this means that these trends were negatively correlated in at least 198 of the 200 resamples.
[12] Our bootstrapping accounts for both sources of variability.

[Figure 1 about here]

[Table 2 about here]

As the trend line in Figure 1 indicates, however, these gender associations of schooling may be relatively recent. In 1940, the schooling scale was not significantly more feminine or masculine than the average word (mean = .16, n.s.). However, it steadily gained feminine associations in every subsequent window, and by 1955 became significantly feminine (mean = 0.44, $p<.01$). As the gender gap in educational attainment closed over the 20th century, schooling continued to move monotonically towards femininity before plateauing in 1985, right as college degree attainment reached gender parity (DiPrete and Buchmann 2013:27). Since 1985, the schooling scale has remained nearly a full standard deviation more feminine than the average word (in 1985, mean = 0.97, $p<0.01$; in 2000, mean = 0.98, $p<0.01$). This change between 1940 and 2000 is consistent with our theoretical predictions (see table 1).

To explore this change in more detail, we examine some of the individual keywords that make up the schooling scale. Among these keywords, "classroom", "graduation", "education", and "student" underwent the biggest changes. For example, in 1940, "student," was 0.35 s.d. more *masculine* than the average word, indicating that it had moderate masculine associations. By the last window, centered on 2000, this flipped to 1.12 s.d. more *feminine* than the average word, indicating a strong feminine association. In fact, among all 16,336 words in the vocabulary for both 1940 and 2000, "student" experienced 213th-largest move towards femininity (top 1.5% of words). The terms "education", "graduation", and "classroom" were similarly in this top 1.5%.

For context, we can contrast these keywords with the word "empowered." Google Books Ngrams first register the use of the phrase "women's empowerment" in print in 1968. The phrase

remained relatively obscure until 1980, when its frequency began to rapidly increase. It reached a symbolic pinnacle during the 1995 United Nation's Conference in Beijing, which "marked the apex of 20 years of sustained endeavor to secure women's empowerment as a central element in international development discourse" (Eyben and Napier-Moore 2009:258). From 1940 to 2000, the term "empowered" shifted 1.34 standard deviations towards femininity (top 2%)—a large change which nonetheless is noticeably smaller than the 1.47 standard deviation shift undergone by "student." This contextualizes the magnitude of the cultural shift which resulted in the association of schooling and femininity that we observe today.

Classroom behaviors

We next turn to two scales related to classroom behaviors: socio-behavioral skills (like attentiveness and politeness) and problem behaviors (like aggression and fighting). In contrast with the schooling scale, figure 2 and table 2 indicate that socio-behavioral skills are significantly feminine even in 1940 (mean = 0.84, $p < 0.01$). They remain consistently feminine through the whole time series and are still significantly feminine in 2000 (mean = 0.57, $p < 0.01$), if at a slightly lower level (difference between 1940 and 2000 = -0.27, n.s). Thus, as we expected, the feminine associations of socio-behavioral skills appear to pre-date this historical period.

[Figure 2 about here]

Figure 3 depicts results for problem behaviors. We see that problem behaviors are significantly masculine in our first corpus time window (mean for 1940 = -0.42, $p < 0.01$), and remain consistently as masculine throughout the rest of the time series (mean for 2000 = -0.55, $p < 0.01$). In any given decade, these terms for problem behaviors are thus approximately a half standard deviation more masculine than the average word. If these terms underwent any change

between 1940 and 2000, it was a small shift towards even greater masculinity—albeit one that is not statistically significant (difference = -0.13, n.s.). Thus, as we expected, the masculine associations of being a poorly behaved student also appear to pre-date the period in our study.

[Figure 3 about here]

School effort

We now turn to our scale for school effort, which tracks the associations of being a hard-working student. In the 1940s, school effort was not significantly gendered (mean = -0.13, n.s.; see figure 4). The associations of school effort then steadily feminized, first becoming significantly feminine in 1975 (mean = 0.43, $p < 0.05$). They reached their current level in 1985 (mean = 0.66, $p < 0.01$), where they have since remained. This is again consistent with our theoretical expectations.

[Figure 4 about here]

Intelligence and Unintelligence

We now turn to intelligence and unintelligence, for which existing work offered no consistent predictions. The gender associations of intelligence, too, appear dynamic (see Figure 5). As table 2 indicates, between 1940 and 1955, intelligence remained neutral to slightly feminine (mean = 0.26, n.s.), but this association was not significant except in 1950 (mean = 0.26, $p < 0.05$). Starting 1955, however, the association moved persistently towards masculinity, until plateauing with a significant masculine association starting 1990 (mean = -0.55, $p < 0.01$). Thus, intelligence gained significant masculine associations.

[Figure 5 about here]

One term in the intelligence scale that deserves special attention is “genius.” This is by far the most masculine term in the scale across all the corpus windows. Indeed, out of the 23,467

words in our 2000 window, "genius" is the 146th-most masculine (top 0.6%) with a standardized projection of -2.39, as compared to -1.26 in 1940—an increase of more than a standard deviation. For context, this means "genius" is slightly more masculine in 2000 than "brother" (-2.26) or "football" (-2.26) and roughly as masculine as "marines" (-2.33), "bullet" (-2.39), and "buddy" (-2.37). Looking at terms that are the same distance from zero but in the direction of femininity, we find that "genius" is roughly as masculine as "estrogen" (2.33), "fragrance" (2.38), "quilts" (2.40), "sari" (2.40), and "motherhood" (2.43) are feminine. In our most recent decade, "genius" thus has exceptionally strong masculine associations, which echoes recent sociological work on the "genius effect" (e.g., Bian et al. 2017; Musto 2019).

In Figure 6, we observe that not only is intelligence becoming increasingly masculine, but so is unintelligence. Unintelligence begins in the 1940 window with a significant feminine association (mean = 0.49, $p < 0.01$). It steadily sheds these associations until it eventually stops being significantly associated with either gender in 1970. It then continues moving towards masculinity, becoming significantly masculine in the 1990 window. By our final window (2000), unintelligence has an average projection of -0.55 ($p < 0.01$), which makes it approximately as masculine as it was feminine in our first window.

[Figure 6 about here]

Similarities between trend lines

As Figure 7 illustrates, the intelligence and unintelligence scales follow nearly identical trajectories to one another (cor = 0.98, $p < 0.01$). We did not predict this similarity, as none of the literature we reviewed suggested that the gender associations of intelligence and unintelligence would undergo the same (rather than opposite) changes. We interpret this unexpected simultaneous movement in detail in the discussion section below.

[Figure 7 about here]

Finally, we note that the gender associations of intelligence and school effort also move with surprising synchrony, albeit in opposite directions. Overall, the correlation between them is $r = -0.93$ ($p < 0.01$), which indicates that they have more than 86% of their variance in common. Figure 8 highlights the surprising extent to which these two trends mirror one another. As Table 3 details, in 1940, the intelligence and school effort scales are slightly feminine (mean = 0.26, n.s.) and masculine (-0.13, n.s.), respectively (difference = -0.39, n.s.). Starting in 1950, both scales move in the direction of the opposite gender. Except for a shared pause in 1960-1965, these changes continue until the scales reach their present associations in 1990 for intelligence (masculine; mean = -0.55, $p < 0.01$) and 1985 for school effort (feminine; mean = 0.66, $p < 0.01$), after which both scales plateau. We offer one possible explanation for this synchrony in the Discussion.[13]

[Figure 8 about here]

[Table 3 about here]

DISCUSSION

Gender stereotypes have powerful consequences for adolescents' academic outcomes. Because of a lack of systematic repeated measures of gender stereotypes across the years (Ridgeway 2011:167), it has remained largely unknown whether and how these stereotypes transformed amid radical shifts in men and women's educational attainment across the 20th century. In this article, we leveraged bootstrapped word embeddings and a large corpus of American print media (1930-2009) to construct a novel historical barometer for these changes. We examined six sets of

[13] In Appendix C, we demonstrate that our results cannot be attributed to differences in the part-of-speech composition of the scales.

gender associations that prior work linked to educational outcomes: (i) schooling, (ii) socio-behavioral skills, (iii) behavioral problems, (iv) school effort, (v) intelligence, and (vi) unintelligence. We found that while the stereotypes of socio-behavioral skills and problem behaviors remained unchanged since the middle of the 20th century, the other four stereotypes have undergone dramatic changes, gaining substantial associations with either femininity or masculinity. Contra Jones et al. (2020), none of the six stereotypes we tracked experienced a substantial decrease in gender associations. Below, we revisit these empirical results to develop their theoretical implications. We begin with trends for which we made theoretical predictions (see Table 1). We then examine the remaining trends and their interrelationships, drawing on these unexpected results to contribute to contemporary theoretical understandings of the gender system.

Expected Trends

First, we noted that (i) socio-behavioral skills and (ii) problem behaviors have substantial overlap with the core stereotypes of women as communal and men as agentic—stereotypes that have been shown to be especially resistant to change (Cejka and Eagly 1999; Diekman and Eagly 2000; Haines et al. 2016; Koenig and Eagly 2005; Spence and Buckner 2000).[14] This durability may be because these core stereotypes are particularly closely linked to the basic structure of the gender system, where men are more likely to occupy positions of power and prestige, and women to occupy lower-power positions that involve deference and caregiving (Ridgeway 2011). We thus conjectured that socio-behavioral skills and problem behaviors would also remain relatively fixed in their gender associations.

[14] There is evidence that female survey respondents came to see themselves as more agentic. However, women's individual *self-perceptions* need not follow the same trends as stereotypical beliefs about *typical* women (Ridgeway 2011:168).

And indeed, our results in figures 2 and 3 indicate that both scales already had strong gender associations at the beginning of our time series. These associations then remained largely unchanged across the 20th century, which is consistent with the image of the core stereotypical distinction between women as communal and men as agentic acting as a cultural bedrock for the gender system. Gender differences in these stereotypes of classroom behaviors may have thus also potentially provided a longstanding classroom advantage for girls and disadvantage for boys, as suggested by previous scholars (DiPrete and Buchmann 2013; Goldin et al. 2006).

Second, to arrive at predictions regarding schooling and school effort, we began by noting that, since the mid-twentieth century, women's academic attainment rose drastically, and empirical evidence suggests that girls' academic effort increased as well. Various theoretical accounts hold that cultural beliefs about gender are tightly intertwined with material arrangements, such as the proportion of men and women in higher education (Eagly 1987; Ridgeway 2011). This pointed us to the prediction that schooling and school effort should become more closely associated with femininity.

Our results for scales that track (iii) schooling (e.g., school, students) and (iv) school effort support this supposition. Whereas neither scale had a statistically significant association with either gender in our earliest time window, we find that both steadily gained significant feminine associations across the 20th century. Additionally, the relatively recent emergence of these gender associations suggests that the stereotype of schooling as "for girls" observed in contemporary empirical work (Jackson and Dempster 2009; Morris 2008) may be a relatively recent cultural phenomenon—and thus possibly one that is less durable and more open to classroom intervention than stereotypes around socio-behavioral skills and problem behaviors.

Other Observations

We now turn to results for which we did not make predictions, beginning with (v) intelligence and (vi) unintelligence. We found that intelligence did not have a significant gender association in 1940, whereas unintelligence had an association with femininity. Across the decades, both stereotypes shifted substantially towards masculinity, ending in 2000 with significant masculine associations. This invites two questions: first, why might both antonyms become associated with the same gender (rather than with different genders); and second, why might intelligence follow the opposite trend from school effort? We examine these in order.

To interpret the counterintuitive simultaneous movement of both intelligence and unintelligence towards masculinity, recall that Word2Vec positions words relative one another in vector space so that different regions of this space represent different contexts. There are some contexts that words for intelligence and unintelligence do not share: for example, terms for intelligence are more likely to occur in the context of flattery or praise, while terms for unintelligence are more likely to double as insults. Conversely, the context that is perfectly shared by terms for intelligence and unintelligence—i.e., when terms for either are equally likely to be observed[15]—occurs specifically when the intelligence of a person is judged, compared, or questioned: e.g., X is "smart" or "dumb." Because trends for intelligence and unintelligence move in nearly perfect unison, we suspect that this is exactly the context that has gained strong masculine associations. This means that the *judgement of intelligence* (whether high or low) became an increasingly salient attribute of masculinity across time. This result recalls findings from recent ethnographic work, which has observed that boys are more likely than girls to be seen as *either* "super smart" or unintelligent (Morris 2008; Musto 2019).

[15] Net of how common those terms are overall.

Perhaps our most striking finding concerns the mirror-opposite trends followed by intelligence and school effort. This result recalls scholarship about students' educational mindsets (Dweck 2006), where the perception that intelligence and school effort are mutually contradictory is a core aspect of the "fixed mindset"—the popular belief that intelligence is fixed from birth rather than acquired through effort. From the point of view of the fixed mindset, "either you have ability or you expend effort"; thus, "effort is only for people with deficiencies" (Dweck 2006:40–42). Indeed, research suggests that the appearance of effortlessness is held in high esteem across many social domains, including evaluations of the quality of inventor's ideas (Elmore and Luna-Lucero 2017), entrepreneurs' business proposals (Tsay 2016), and musician's talents (Tsay and Banaji 2011). Thus, if intelligence and school effort are seen as competing routes to academic achievement, intelligence is clearly the higher-status and more desirable route.

A growing body of scholarship demonstrates that this opposition between intelligence and school effort is strongly gendered, with students frequently associating effortless intelligence with men and school effort with women (Elmore and Luna-Lucero 2017; Heyder and Kessels 2017; Jackson and Nyström 2015). Our results reinforce the conclusions of this work, but also indicate that these gender associations may be a relatively recent phenomenon. Indeed, the exceptionally strong negative correlation we observed between the gender associations of intelligence and school effort seemingly suggests that these two sets of cultural associations are interlinked and may have arisen as part of one broader change to the gender system.

While our methods do not allow us to observe the mechanism behind this social change, we can draw on contemporary accounts of the gender system to offer a tentative explanation. First, Ridgeway (2011) argues that the core of the gender system can persist because its surface

components are flexible: even amid material and cultural changes, women and men continue to be differentiated, with new social arrangements reframed so that men continue to be understood as higher status than women. From this perspective, the timing of our observed simultaneous change is conspicuous, with men's intelligence and women's school effort gaining in emphasis just as women came to challenge men in education. It is thus possible that, as educational attainment—an important social characteristic—stopped being a marker that advantaged men, women's achievements in school became reframed as a product of effort—a less lauded and lower status avenue to success than effortless brilliance (Heyder and Kessels 2017; Jackson and Nyström 2015). Meanwhile, men's academic performance became reframed as a question of effortless intelligence, thus possibly allowing men to reap higher social standing and academic self-confidence from equal (or even slightly lower) academic achievement as women (e.g., Tormala, Jia, and Norton 2012).

Second, our findings on effort and intelligence might reflect changing gender ideologies. Historically, men's and women's traits were perceived as innate and fixed based on their biological sex (Bohan 1993). During our study period, ideologies about femininity – but not masculinity – may have changed: drawing on interviews with youth sports coaches, Messner (2011) argues that girls' traits are now perceived as malleable and their lives as full of *choice*, while boys' traits are still perceived as fixed and biologically determined (see also Williams 2016). Gender stereotypes around school effort and intelligence may reflect this new opposition: girls may put in effort and thus succeed (if they choose); but boys' success comes from an innate quality — they are innately brilliant or innately dull.

Taken together, our results appear consistent with a dynamic landscape of gender stereotypes across the 20th century anchored by a persistent, underlying gender system whereby

men and women are differentiated and men are afforded higher status than women. Prior work has argued that core aspects of gender stereotypes remain largely unchanged partly because the distribution of men and women to high-prestige, powerful positions and lower-prestige, caretaking positions remains highly unequal. Thus, women continue to be seen as more communal and men as more agentic. Accordingly, we found that the schooling-relevant gender stereotypes that directly relate to these characteristics—girls' socio-behavioral skills and boys' problem behaviors—also did not shift. Conversely, throughout the 20th century, women caught up to and surpassed men in educational attainment. We observed that two gender stereotypes arose in parallel with this distinction: the association of school and studying with femininity, and of intelligence with masculinity. This pattern reflects *both* change and persistence in the gender system. Further, since effortless intelligence is a higher-status route to achievement than effortful studying, this gendering of school effort may partly enable the hierarchical distinction at the core of the gender system to persist, even amidst sweeping material changes in women's educational achievement.

Limitations and Future Directions

Our findings leave several avenues for future research. First, we did not break down schooling-relevant stereotypes by academic discipline. We thus could not engage with the unevenness of the rise in women's attainment across different fields. Indeed, women continue to be underrepresented in disciplines perceived to require innate exceptional intelligence, such as STEM fields (Leslie et al. 2015). How does this segregation of men and women into different disciplines relate to the changing landscape of gender stereotypes we examined here? Future work could address this question by contrasting the changing gender associations of different disciplines and examining the specific stereotypes relevant to success in those fields. Second,

while our study focused on gender, educationally-relevant skills and behaviors also have stereotypical associations with social class, race/ethnicity, and sexual orientation (Hegarty 2007; Hsin 2018; Morris 2008; Williams 2016). Future work should use our approach to investigate how chronological changes to education-relevant stereotypes map onto a broader array of social categories and their intersections.

Our use of the Corpus of Historical American English (COHA) also presented several data limitations. COHA is a 400-million-word full-text rigorously documented academic corpus that was specifically designed to study historical change (Davies 2012). This makes it a "small and tidy" corpus (Mair 2006), whereas Word2Vec was designed for "large and messy" corpora like the 130-billion-word Google Books English Ngrams. While existing work had applied Word2Vec to COHA, our pre-testing showed that the estimates this produces may not be robust. To overcome this limitation, we pioneered the use of bootstrap aggregating (or "bagging", Breiman 1996) to stabilize the embedding estimates. We repeated our analyses across 200 bootstrapped resamples of our corpus and "bagged" these estimates to yield robust final results. We urge other scholars to incorporate bagging into their Word2Vec analyses of small corpora to ensure the robustness of their findings.

Our use of a "small and tidy" corpus also limited the range of keywords available to our analyses. To further increase the robustness of our results, we examined only keywords that occurred at least 100 times in each corpus window. If future scholars approached our questions with a larger corpus, they may have access to a wider set of keywords that enable finer distinctions than we could make here—for example, contrasting between pairs of closely related stereotypes, or studying stereotypical associations with less populous social categories. Indeed,

embedding-based analyses of associations with intersectional social groups may necessitate this kind of larger corpus.

CONCLUSION

As Ridgeway (2011:167) notes, "measures of gender stereotypes have not been systematically administered over the years," which has limited our understanding of how gender stereotypes have changed across time. In this manuscript, we demonstrated how recently developed word embedding methods can be used to construct new historical measures of these stereotypes—even when no traditional data are available. Moreover, embedding methods make it possible to investigate larger sets of related stereotypes than would generally be found in a longitudinal survey—and to do so across vast time scales. A similar approach can be used to observe many further "missing" longitudinal variables in other sociological subfields. We thus hope our work helps further popularize embedding-based investigations in sociology—work that may be able to uncover important cultural shifts that have hitherto seemed empirically intractable.

REFERENCES

Adler, Patricia, Steven Kleiss, and Peter Adler. 1992. "Socialization to Gender Roles: Popularity among Elementary School Boys and Girls." *Sociology of Education* 65(3):169–87.

Bandura, Albert. 2001. "Social Cognitive Theory of Mass Communication." *Media Psychology* 3(3):265–99.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2017. "Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children's Interests." *Science* 355(6323):389–91. doi: 10.1126/science.aah6524.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2018. “Evidence of Bias against Girls and Women in Contexts That Emphasize Intellectual Ability.” *American Psychologist* 73(9):1139.

Blatz, W. E., and E. A. Bott. 1927. “Studies in Mental Hygiene of Children I. Behavior of Public School Children—A Description of Method.” 34.

Bohan, Janis S. 1993. “Regarding Gender: Essentialism, Constructionism, and Feminist Psychology.” *Psychology of Women Quarterly* 17(1):5–21.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” *Adv. Neural Inf. Process. Syst.*

Bordua, David J. 1960. “Educational Aspirations and Parental Stress on College.” *Social Forces* 38(3):262–69.

Breiman, Leo. 1996. “Bagging Predictors.” *Machine Learning* 24(2):123–40. doi: 10.1007/BF00058655.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Science* 356(6334):183–86.

Cejka, Mary Ann, and Alice H. Eagly. 1999. “Gender-Stereotypic Images of Occupations Correspond to the Sex Segregation of Employment.” *Personality and Social Psychology Bulletin* 25(4):413–23. doi: 10.1177/0146167299025004002.

Clark, Edward. 1967. “Sex Differences in the Perception of Academic Achievement among Elementary School Children.” *The Journal of Psychology* 67(2):249–56.

Cohen, Michele. 1998. "'A Habit of Healthy Idleness': Boys' Underachievement in Historical Perspective." in *Failing Boys? Issues in gender and achievement*. Buckingham: Open University Press.

Davies, Mark. 2012. "Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English." *Corpora* 7(2):121–57. doi: 10.3366/cor.2012.0024.

Diekman, Amanda, and Alice Eagly. 2000. "Stereotypes as Dynamic Constructs: Women and Men of the Past, Present, and Future." *Personality and Social Psychology Bulletin* 26(10):1171–88.

DiPrete, Thomas A., and Claudia Buchmann. 2013. *The Rise of Women: The Growing Gender Gap in Education and What It Means for American Schools*. Russell Sage Foundation.

DiPrete, Thomas A., and Jennifer L. Jennings. 2012. "Social and Behavioral Skills and the Gender Gap in Early Educational Achievement." *Social Science Research* 41(1):1–15. doi: 10.1016/j.ssresearch.2011.09.001.

Dixon, Travis L. 2006. "Psychological Reactions to Crime News Portrayals of Black Criminals: Understanding the Moderating Roles of Prior News Viewing and Stereotype Endorsement." *Communication Monographs* 73(2):162–87. doi: 10.1080/03637750600690643.

Downey, Douglas B., and Anastasia S. Vogt Yuan. 2005. "Sex Differences in School Performance During High School: Puzzling Patterns and Possible Explanations." 46(2):229–321.

Dweck, Carol S. 2006. *Mindset : The New Psychology of Success*. New York: Random House.

Eagly, Alice. 1987. *Sex Differences in Social Behaivor: A Social-Role Analysis*. Hillsdale, NJ: Erlbaum.

Eagly, Alice H., Christa Nater, David I. Miller, Michèle Kaufmann, and Sabine Sczesny. 2020. "Gender Stereotypes Have Changed: A Cross-Temporal Meta-Analysis of U.S. Public Opinion Polls from 1946 to 2018." *American Psychologist* 75(3):301–15. doi: 10.1037/amp0000494.

Eagly, Alice H., and Valerie J. Steffen. 1984. "Gender Stereotypes Stem from the Distribution of Women and Men into Social Roles." *Journal of Personality and Social Psychology* 46(4):735–54. doi: 10.1037/0022-3514.46.4.735.

Eagly, Alice H., and Wendy Wood. 1999. "The Origins of Sex Differences in Human Behavior." *American Psychologist* 54(6):408–23.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7(1):1–26.

Elfenbein, Andrew. 1999. *Romantic Genius: The Prehistory of a Homosexual Role*. Columbia University Press.

Elmore, Kristen C., and Myra Luna-Lucero. 2017. "Light Bulbs or Seeds? How Metaphors for Ideas Influence Judgments About Genius." *Social Psychological and Personality Science* 8(2):200–208. doi: 10.1177/1948550616667611.

Entwisle, Doris R., Karl L. Alexander, and Linda S. Olson. 2007. "Early Schooling: The Handicap of Being Poor and Male." *Sociology of Education* 80(2):114–38. doi: 10.1177/003804070708000202.

Epstein, Debbie. 1998. “Real Boys Don’t Work:“underachievement’, Masculinity, and the Harassment of ‘Sissies.’” in *Failing Boys? Issues in gender and achievement*. Philadelpha, PA: Open University Press.

Espinoza, Penelope, da Luz Fontes Arêas, and Clarissa Arms-Chavez. 2014. “Attributional Gender Bias: Teachers’ Ability and Effort Explanations for Students’ Math Performance.” *Social Psychology of Education* 17:105–26.

Eyben, Rosalind, and Rebecca Napier-Moore. 2009. “Choosing Words with Care? Shifting Meanings of Women’s Empowerment in International Development.” *Third World Quarterly* 30(2):285–300. doi: 10.1080/01436590802681066.

Fiske, Susan T., Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. “A Model of (Often Mixed) Stereotype Content: Competence and Warmth Respectively Follow from Perceived Status and Competition.” *Journal of Personality and Social Psychology* 82(6):878–902. doi: 10.1037//0022-3514.82.6.878.

Fortin, Nicole M., Philip Oreopoulos, and Shelley Phipps. 2015. “Leaving Boys Behind: Gender Disparities in High Academic Achievement.” *Journal of Human Resources* 50(3):549–79. doi: 10.3368/jhr.50.3.549.

Fox, Gudelia. 1968. “The Gifted: How Are They Viewed?” *Gifted Child Quarterly* 12(1):23–33.

Gamson, William A., David Croteau, William Hoynes, and Theodore Sasson. 1992. “Media Images and the Social Construction of Reality.” *Annual Review of Sociology* 18(1):373–93.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. “Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes.” *Proceedings of the National Academy of Sciences* 115(16):E3635–44. doi: 10.1073/pnas.1720347115.

Gerbner, George, Larry Gross, Michael Morgan, Nancy Signorielli, and James Shanahan. 2002. "Growing up with Television: Cultivation Processes." Pp. 53–78 in *Media effects*. Routledge.

Goldin, Claudia, Lawrence F. Katz, and Ilyana Kuziemko. 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap." *Journal of Economic Perspectives* 20(4):133–56. doi: 10.1257/jep.20.4.133.

Grabe, Shelly, L. Monique Ward, and Janet Shibley Hyde. 2008. "The Role of the Media in Body Image Concerns among Women: A Meta-Analysis of Experimental and Correlational Studies." *Psychological Bulletin* 134(3):460.

Grand, Gabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. "Semantic Projection: Recovering Human Knowledge of Multiple, Distinct Object Features from Word Embeddings." *ArXiv Preprint* 1802(01241).

Haines, Elizabeth, Kay Deaux, and Nicole Lofaro. 2016. "The Times They Are A-Changing… or Are They Not? A Comparison of Gender Stereotypes." *Psychology of Women Quarterly* 40(3):353–63.

Hartley, Ruth E. 1959. "Sex-Role Pressures and the Socialization of the Male Child." *Psychological Reports* 5:457–68.

Hayes, Margaret Louise. 1943. *A Study of the Classroom Disturbances of Eighth Grade Boys and Girls*. New York: Teachers College, Columbia University.

Hegarty, Peter. 2007. "From Genius Inverts to Gendered Intelligence: Lewis Terman and the Power of the Norm." *History of Psychology* 10(2):132.

Heyder, Anke, and Ursula Kessels. 2013. “Is School Feminine? Implicit Gender Stereotyping of School as a Predictor of Academic Achievement.” *Sex Roles* 69(11–12):605–17. doi: 10.1007/s11199-013-0309-9.

Heyder, Anke, and Ursula Kessels. 2017. “Boys Don’t Work? On the Psychological Benefits of Showing Low Effort in High School.” *Sex Roles* 77(1):72–85. doi: 10.1007/s11199-016-0683-1.

Hicks, Allan J., and Margaret Hayes. 1938. “Study of the Characteristics of 250 Junior High School Children.” *Child Development* 9(2):19–242.

Hsin, Amy. 2018. “Hegemonic Gender Norms and the Gender Gap in Achievement: The Case of Asian Americans.” *Sociological Science* 5:752–74. doi: 10.15195/v5.a32.

Hughes, Rhidian, and Meg Huby. 2012. “The Construction and Interpretation of Vignettes in Social Research.” *Social Work and Social Sciences Review* 11(1):36–51. doi: 10.1921/swssr.v11i1.428.

Jackson, Carolyn. 2002. “‘Laddishness’ as a Self-Worth Protection Strategy.” *Gender and Education* 14(1):37–50.

Jackson, Carolyn. 2003. “Motives for ‘Laddishness’ at School: Fear of Failure and Fear of the ‘Feminine.’” *British Educational Research Journal* 29(4):583–498.

Jackson, Carolyn, and Steven Dempster. 2009. “‘I Sat Back on My Computer … with a Bottle of Whisky next to Me’: Constructing ‘Cool’ Masculinity through ‘Effortless’ Achievement in Secondary and Higher Education.” *Journal of Gender Studies* 18(4):341–56. doi: 10.1080/09589230903260019.

Jackson, Carolyn, and Anne-Sofie Nyström. 2015. “‘Smart Students Get Perfect Scores in Tests without Studying Much’: Why Is an Effortless Achiever Identity Attractive, and for

Whom Is It Possible?" *Research Papers in Education* 30(4):393–410. doi: 10.1080/02671522.2014.970226.

Jolly, Jennifer. 2008. "Historical Perspectives: Lewis Terman: Genetic Study of Genius—Elementary School Students." *Gifted Child Today* 31(1):27–33.

Jones, Jason, Mohammad Amin, Jessica Kim, and Steven Skiena. 2020. "Stereotypical Gender Associations in Language Have Decreased Over Time." *Sociological Science* 7:1–35. doi: 10.15195/v7.a1.

Jones, Mary Cover. 1960. "A Comparison of the Attitudes and Interests of Ninth-Grade Students over Two Decades." *Journal of Educational Psychology* 51(4):175–86.

Joseph, Kenneth, and Jonathan H. Morgan. 2020. "When Do Word Embeddings Accurately Reflect Surveys on Our Beliefs About People?" *ArXiv:2004.12043 [Cs]*.

Juhn, Chinhui, and Simon Potter. 2006. "Changes in Labor Force Participation in the United States." *Journal of Economic Perspectives* 20(3):27–46. doi: 10.1257/jep.20.3.27.

Kahn, Kim Fridkin. 1994. "Does Gender Make a Difference? An Experimental Examination of Sex Stereotypes and Press Patterns in Statewide Campaigns." *American Journal of Political Science* 162–95.

Kay, Christian, Marc Alexander, Fraser Dallachy, Jane Roberts, Michael Samuels, and Irené Wotherspoon, eds. 2020. *The Historical Thesaurus of English*. 4.21. Glasgow: University of Glasgow.

Kessels, Ursula, Anke Heyder, Martin Latsch, and Bettina Hannover. 2014. "How Gender Differences in Academic Engagement Relate to Students' Gender Identity." *Educational Research* 56(2):220–29.

Klonis, Suzanne C., E. Ashby Plant, and Patricia G. Devine. 2005. “Internal and External Motivation to Respond Without Sexism.” *Personality and Social Psychology Bulletin* 31(9):1237–49. doi: 10.1177/0146167205275304.

Koenig, Anne, and Alice Eagly. 2014. “Evidence for the Social Role Theory of Stereotype Content: Observations of Groups’ Roles Shape Stereotypes.” *Journal of Personality and Social Psychology* 107(3):371.

Koenig, Anne M., and Alice H. Eagly. 2005. “Stereotype Threat in Men on a Test of Social Sensitivity.” *Sex Roles* 52(7–8):489–96. doi: 10.1007/s11199-005-3714-x.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.” *American Sociological Review* 84(5). doi: 10.1177/0003122419877135.

Legewie, Joscha, and Thomas A. DiPrete. 2012. “School Context and the Gender Gap in Educational Achievement.” *American Sociological Review* 77(3):463–85. doi: 10.1177/0003122412440802.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. 2015. “Expectations of Brilliance Underlie Gender Distributions across Academic Disciplines.” *Science* 347(6219):262–65.

Lueptow, Lloyd B., Lori Garovich-Szabo, and Margaret B. Lueptow. 2001. “Social Change and The Persistence of Sex Typing: 1974–1997.” *Social Forces* 80(1):1–36. doi: 10.1353/sof.2001.0077.

Mair, Christian. 2006. “Tracking Ongoing Grammatical Change and Recent Diversification in Present-Day Standard English: The Complementary Role of Small and Large Corpora.” Pp. 355–76 in *The changing face of corpus linguistics*. Brill Rodopi.

Marrs, Heath. 2016. “Conformity to Masculine Norms and Academic Engagement in College Men.” *Psychology of Men & Masculinity* 17(2):197–205. doi: 10.1037/men0000015.

Messner, Michael. 2011. “Gender Ideologies, Youth Sports, and the Production of Soft Essentialism.” *Sociology of Sport Journal* 28(2):151–70. doi: 10.1123/ssj.28.2.151.

Meyer, William J., and George G. Thompson. 1956. *Sex Differences in the Distribution of Teacher Approval and Disapproval among Sixth-Grade Children*. Vol. 47.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” *Adv. Neural Inf. Process. Syst.* 26:3111–19.

Morris, Edward W. 2008. “‘Rednecks,’ ‘Rutters,’ and `Rithmetic: Social Class, Masculinity, and Schooling in a Rural Context.” *Gender & Society* 22(6):728–51. doi: 10.1177/0891243208325163.

Mrksic, Nikola, Diarmuid Ó. Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. “Counter-Fitting Word Vectors to Linguistic Constraints.” Pp. 142–48 in *Proceedings of NAACL-HLT*.

Musto, Michela. 2019. “Brilliant or Bad: The Gendered Social Construction of Exceptionalism in Early Adolescence.” *American Sociological Review* 84(3):369–93. doi: 10.1177/0003122419837567.

Pascoe, CJ. 2007. *Dude, You’re a Fag: Masculinity and Sexuality in High School*. Berkeley: University of California Press.

Peltier, Gary. 1968. “Sex Differences in the School: Problem and Proposed Solutions.” *The Phi Delta Kappan* 50(3):182–85.

Potter, W. James. 2014. “A Critical Analysis of Cultivation Theory: Cultivation.” *Journal of Communication* 64(6):1015–36. doi: 10.1111/jcom.12128.

Prentice, Deborah, and Erica Carranza. 2002. “What Women and Men Should Be, Shouldn’t Be, Are Allowed to Be, and Don’t Have to Be: The Contents of Prescriptive Gender Stereotypes.” *Psychology of Women Quarterly* 26(4):269–81.

Ridgeway, Cecilia L. 2011. *Framed by Gender*. New York: Oxford University Press.

Ridgeway, Cecilia L., Kristen Backor, Yan E. Li, Justine E. Tinkler, and Kristan G. Erickson. 2009. “How Easily Does a Social Difference Become a Status Distinction? Gender Matters.” *American Sociological Review* 74(1):44–62. doi: 10.1177/000312240907400103.

Ridgeway, Cecilia L., and Shelley J. Correll. 2004. “Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations.” *Gender & Society* 18(4):510–31. doi: 10.1177/0891243204265269.

Ruble, Diane, Carol Lynn Martin, and Sheri Berenbaum. 2007. “Gender Development.” in *Handbook of Child Psychology*. Vol. 3. John Wiley & Sons, Inc.

Seguino, Stephanie. 2007. “PlusÇa Change? Evidence on Global Trends in Gender Norms and Stereotypes.” *Feminist Economics* 13(2):1–28. doi: 10.1080/13545700601184880.

Solomon, Barbara Miller. 1985. *In the Company of Educated Women: A History of Women and Higher Education in America*. New Haven: Yale University Press.

Spence, Janet, and Camille Buckner. 2000. “Instrumental and Expressive Traits, Trait Stereotypes, and Sexist Attitudes: What Do They Signify?” *Psychology of Women Quarterly* 24(1):44–53.

Storage, Daniel, Zachary Horne, Andrei Cimpian, and Sarah-Jane Leslie. 2016. "The Frequency of 'Brilliant' and 'Genius' in Teaching Evaluations Predicts the Representation of Women and African Americans across Fields." *PloS One* 11(3).

Tormala, Zakary L., Jayson S. Jia, and Michael I. Norton. 2012. "The Preference for Potential." *Journal of Personality and Social Psychology* 103(4):567–83. doi: 10.1037/a0029227.

Tsay, Chia-Jung. 2016. "Privileging Naturals Over Strivers: The Costs of the Naturalness Bias." *Personality and Social Psychology Bulletin* 42(1):40–53. doi: 10.1177/0146167215611638.

Tsay, Chia-Jung, and Mahzarin R. Banaji. 2011. "Naturals and Strivers: Preferences and Beliefs about Sources of Achievement." *Journal of Experimental Social Psychology* 47(2):460–65. doi: 10.1016/j.jesp.2010.12.010.

Tuddenham, Read D. 1952. *Studies in Reputation: I. Sex and Grade Differences in School Children's Evaluations of Their Peers. II. The Diagnosis of Social Adjustment*. Vol. 66.

Warne, Russell. 2019. "An Evaluation (and Vindication?) Of Lewis Terman: What the Father of Gifted Education Can Teach the 21st Century." *Gifted Child Quarterly* 63(1):3–21.

Warrington, Molly, Mike Younger, and Jacquetta Williams. 2000. "Student Attitudes, Image and the Gender Gap." *British Educational Research Journal* 26(3):393–407.

Williams, Juliet A. 2016. *The Separation Solution?: Single-Sex Education and the New Politics of Gender Equality*. Berkeley, CA: University of California Press.

Wood, Wendy, and Alice H. Eagly. 2002. "A Cross-Cultural Analysis of the Behavior of Women and Men:Implications for the Origins of Sex Differences." *Psychological Bulletin* 128:699–727.

TABLES

Table 1. Predictions based on the existing literature.

	Present associations	Change,1940 - 2000
Schooling	Feminine	Towards femininity
Socio-behavioral skills	Feminine	No change
Problem behaviors	Masculine	No change
School effort	Feminine	Towards femininity
High intelligence	Masculine	X
Low intelligence	X	X

Note: An 'X' indicates situations when the existing literature does not identify an unambiguous prediction.

Table 2. Average projection of six keyword scales onto the Feminine (positive) to Masculine (negative) axis across 13 sliding 20-year windows.

	Schooling		Social and Behavioral Skills		Problem Behaviors		School Effort		Intelligence		Unintelligence	
Window	Mean	Change	Mean	Change	Mean	Change	Mean	Change	Mean	Change	Mean	Change
1940	.16		.84**		-.42 **		-.13		.26		.49 **	
1945	.20	.03	.84**	.00	-.31 **	.11	-.12	.01	.23	-.04	.46 **	-.03
1950	.30	.14	.85**	.02	-.33 **	.09	-.09	.04	.26 *	.00	.57 **	.09
1955	.44**	.28	.73**	-.11	-.45 **	-.03	-.02	.11	.19	-.07	.42 *	-.07
1960	.44**	.28	.66**	-.18	-.45 **	-.03	.13	.26	.06	-.20	.46 *	-.03
1965	.51**	.35	.68**	-.16	-.42 **	.00	.12	.25	.07	-.19	.41 *	-.08
1970	.67**	.51	.72**	-.11	-.42 **	.00	.28	.41	.02	-.24	.27	-.22
1975	.80**	.64*	.70**	-.13	-.45 **	-.03	.43 *	.56	-.04	-.31	.13	-.36
1980	.86**	.70**	.75**	-.08	-.44 **	-.02	.43 *	.56	-.09	-.35	.01	-.48*
1985	.97**	.80**	.73**	-.10	-.56 **	-.14	.66 **	.79**	-.23	-.49*	-.19	-.68**
1990	.95**	.79**	.61**	-.23	-.62 **	-.20	.60 **	.73*	-.55 **	-.81***	-.41 *	-0.90***
1995	1.01**	.85**	.63**	-.21	-.59 **	-.17	.72 **	.85**	-.55 **	-.82***	-.46 *	-0.95***
2000	.98**	.82***	.57**	-.27	-.55 **	-.13	.63 **	.76*	-.54 **	-.80***	-.55 **	-1.03***

Note: The midpoint of each time window *t* is the start of the year indicated in the "Window" column, i.e., "1940" corresponds to the time window from January 1st, 1930 to December 31st, 1949. Each result $mean_t$ in each Mean column is the average projection of the corresponding keyword scale across embeddings estimated from bootstrap resamples of the corresponding COHA window (200 embeddings per window, 2600 embeddings total). Change is from the 1940 mean. Significance tests for mean against $H_0: mean = 0$ are based on CIs from 200 bootstrapped embeddings. Tests for change against $H_0: mean_t = mean_{1940}$ compare each of 200 bootstrapped embeddings for window *t* to each embedding for 1940, since embeddings for different windows are independent (40,000 comparisons per significance test).

* $p < .05$; ** $p < .01$; *** $p < .001$ (two-tailed tests).

Table 3. Differences in the average projections of School Effort and Intelligence onto the Feminine (positive) to Masculine (negative) axis across 13 sliding 20-year windows.

Window	Intelligence	School Effort	Difference (effort – intelligence)
1940	.26	-.13	-.40
1945	.23	-.12	-.35
1950	.26 *	-.09	-.36
1955	.19	-.02	-.20
1960	.06	.13	.07
1965	.07	.12	.05
1970	.02	.28	.26
1975	-.04	.43 *	.48
1980	-.09	.43 *	.53 *
1985	-.23	.66 **	.89 **
1990	-.55 **	.60 **	1.15 **
1995	-.55 **	.72 **	1.27 **
2000	-.54 **	.63 **	1.17 **

Note: The midpoint of each time window is January 1st of the year indicated in the "Window" column. Each result in the Studying and Intelligence column is the mean of that scale across 200 bootstrapped embeddings estimated from resamples of the corresponding part of COHA (total of 2600 embeddings). Difference(year) = Studying(year) – Intelligence(year).

* $p < .05$; ** $p < .01$ (two-tailed tests); significance is estimated by comparing the mean projection of the Studying scale to that of the Intelligence scale within each bootstrapped embedding for a given window (row) (200 comparisons per significance test).

FIGURES

Figure 1. Gendered associations (y) of schooling across time (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

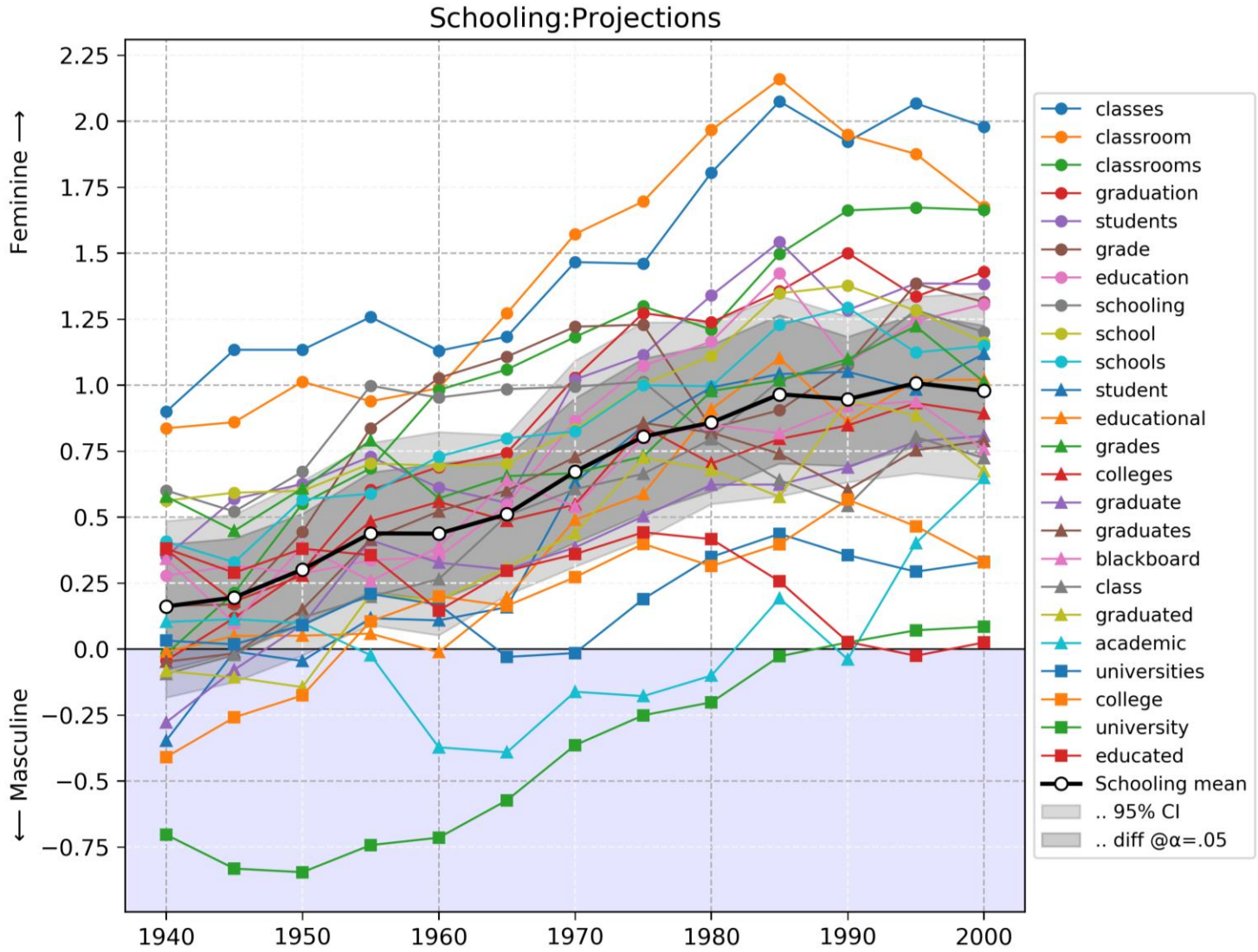


Figure 2. Gendered associations (y) of social-behavioral skills across time (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

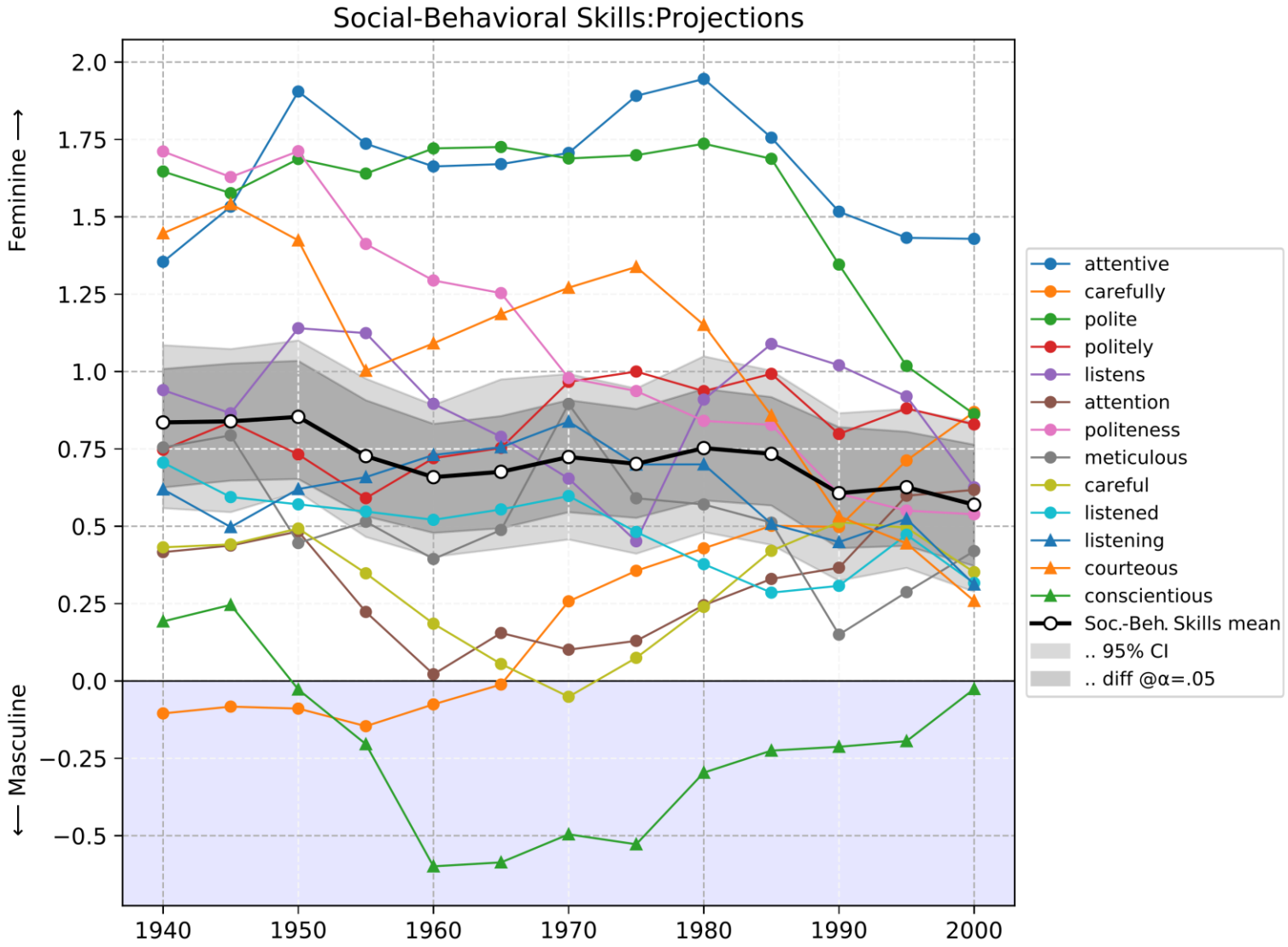


Figure 3. Gendered associations (y) of problem behaviors across time (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

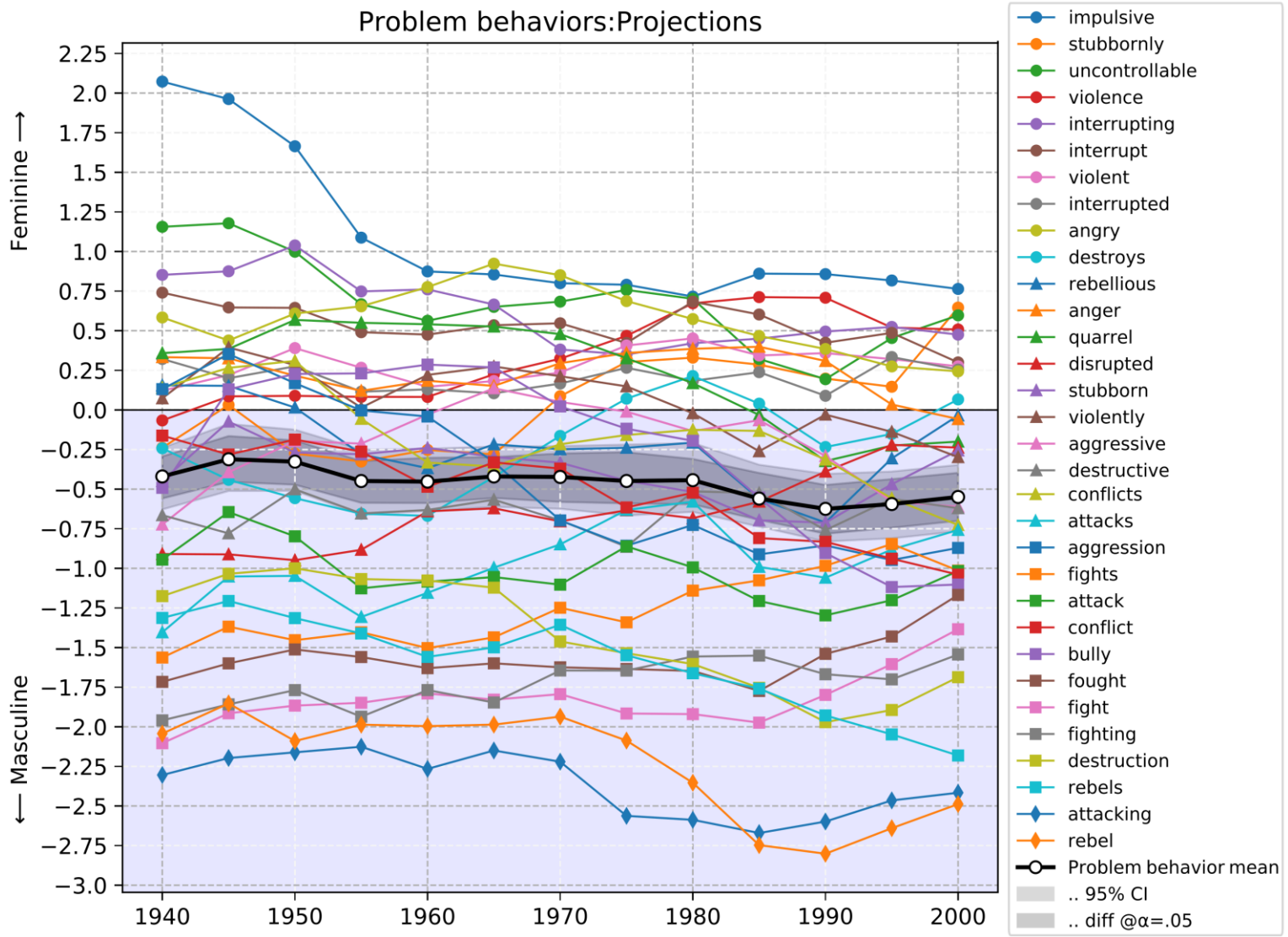


Figure 4. Gendered associations (y) of studying across time (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

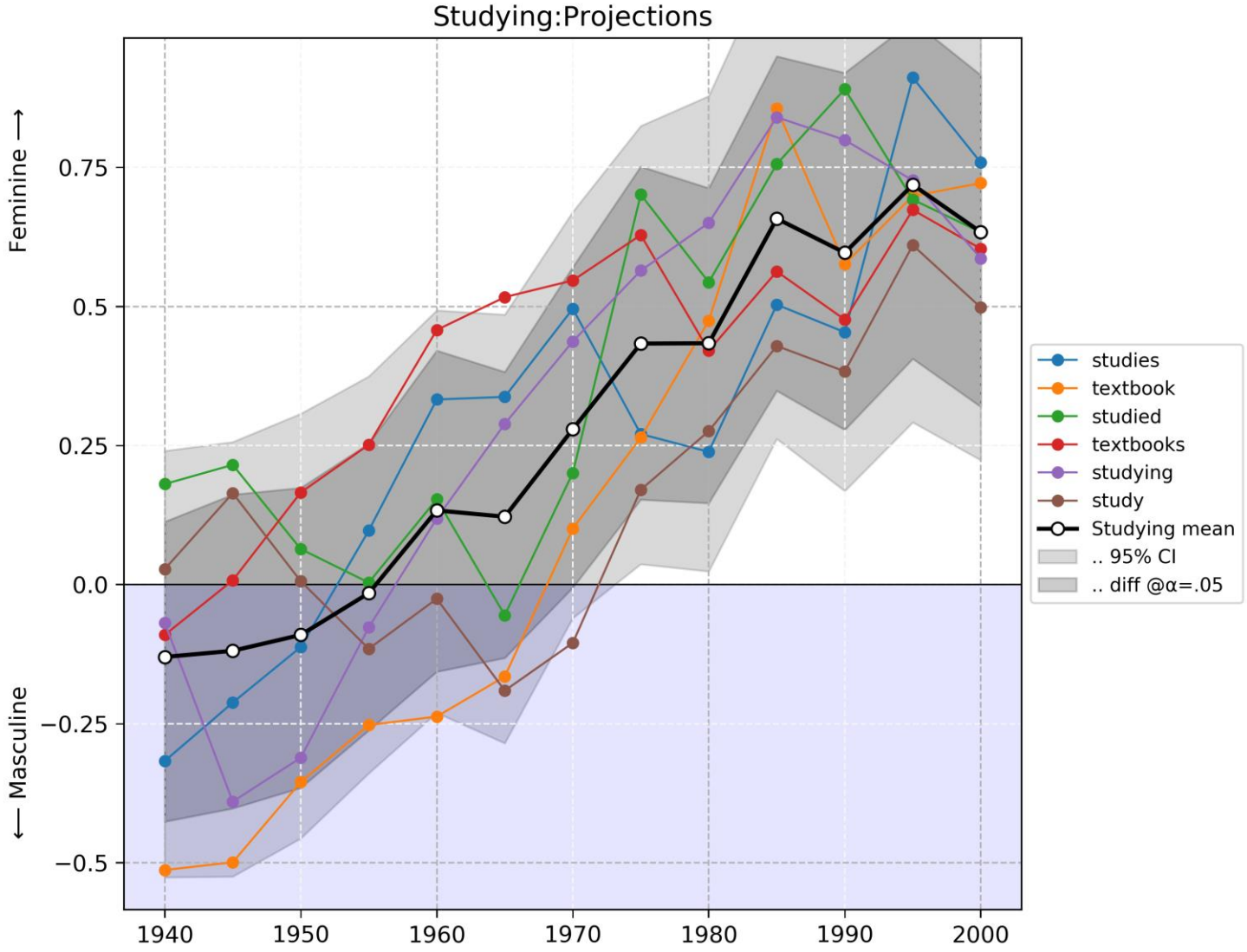


Figure 5. Gendered associations (y) of intelligence (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

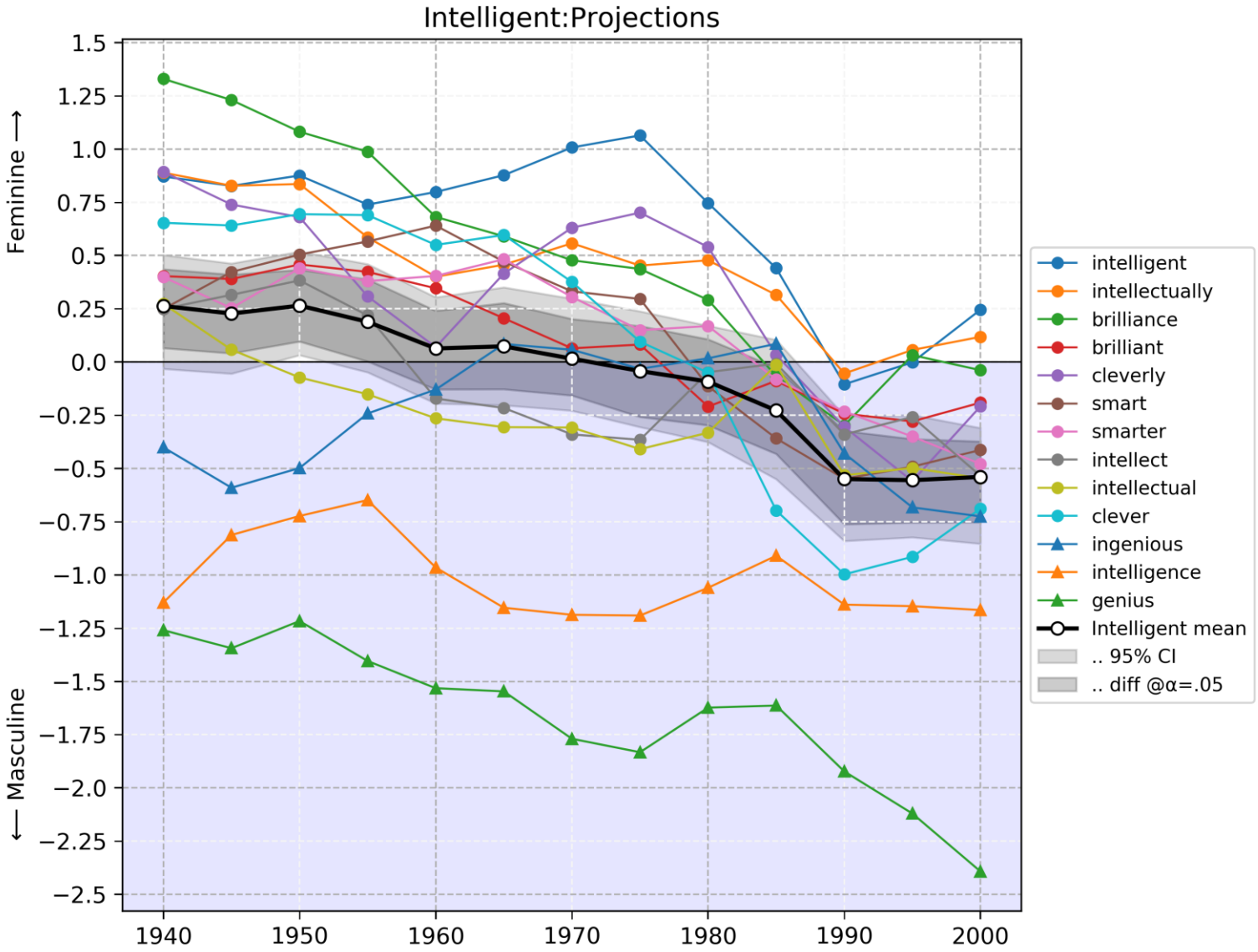


Figure 6. Gendered associations (y) of unintelligence (x). Embedding-based estimates using COHA (200 bootstraps). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point or trend line.

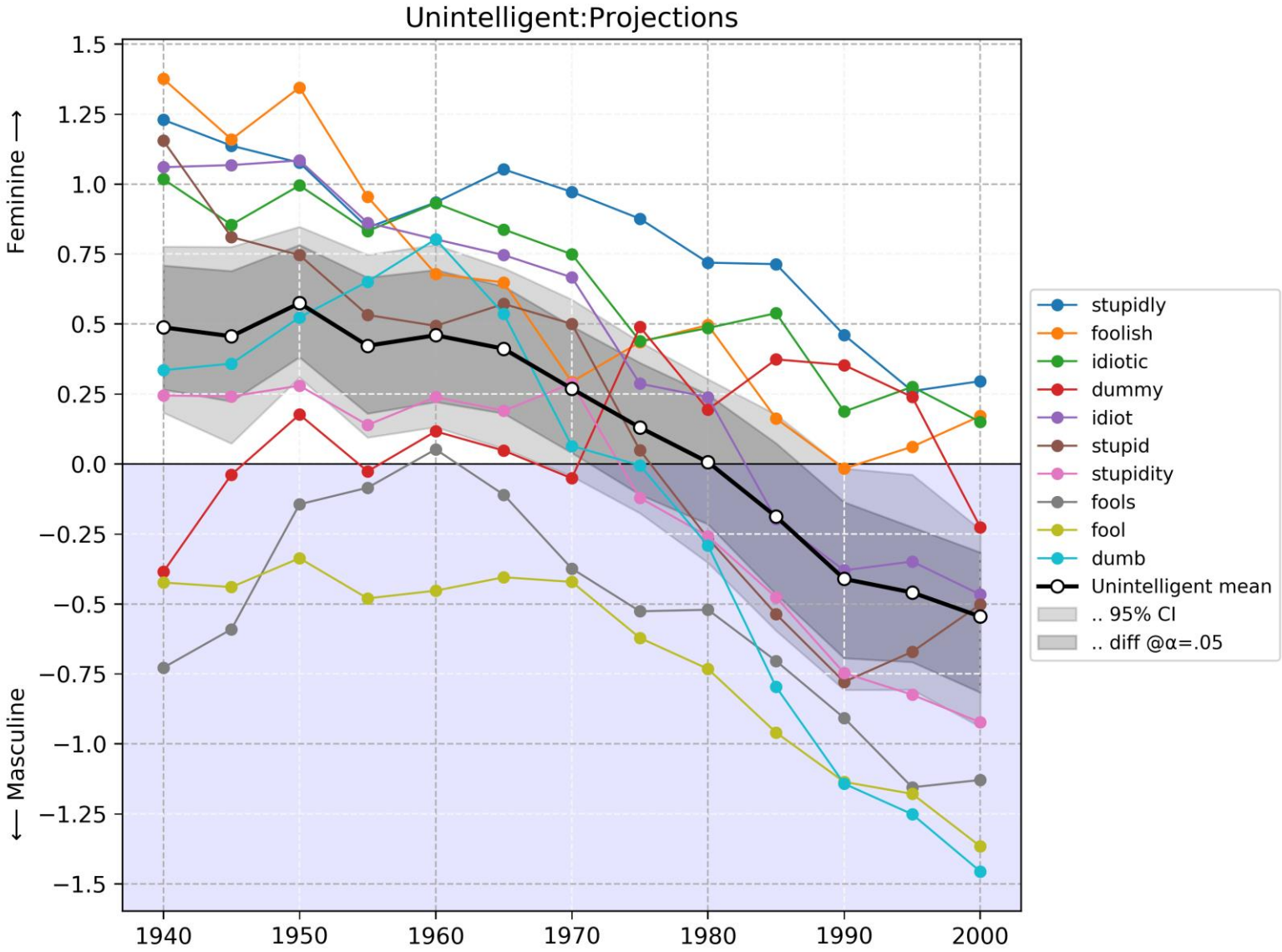


Figure 7. Gendered associations (y) of intelligence and unintelligence across time (x). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point, or across the two scales.

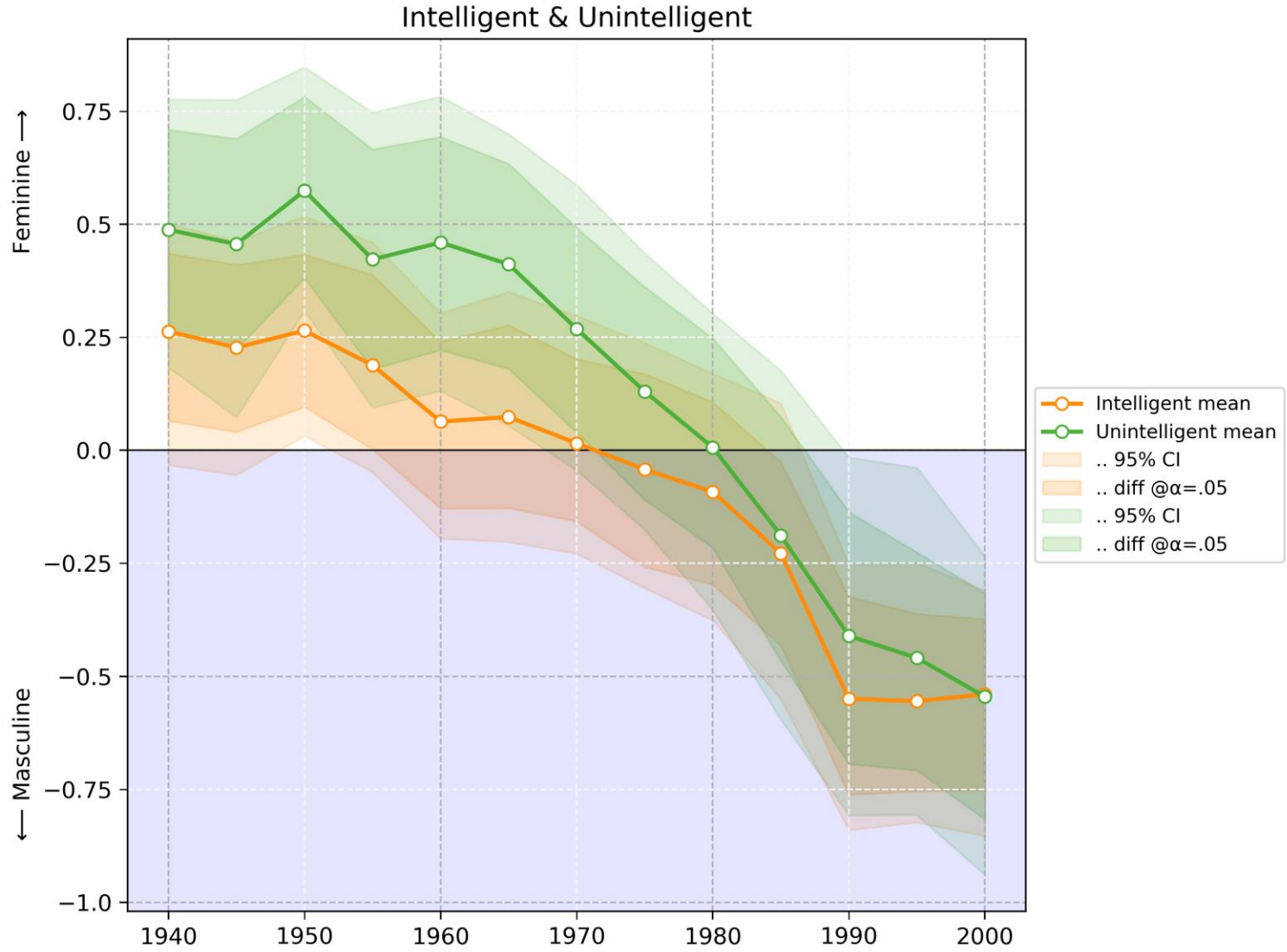
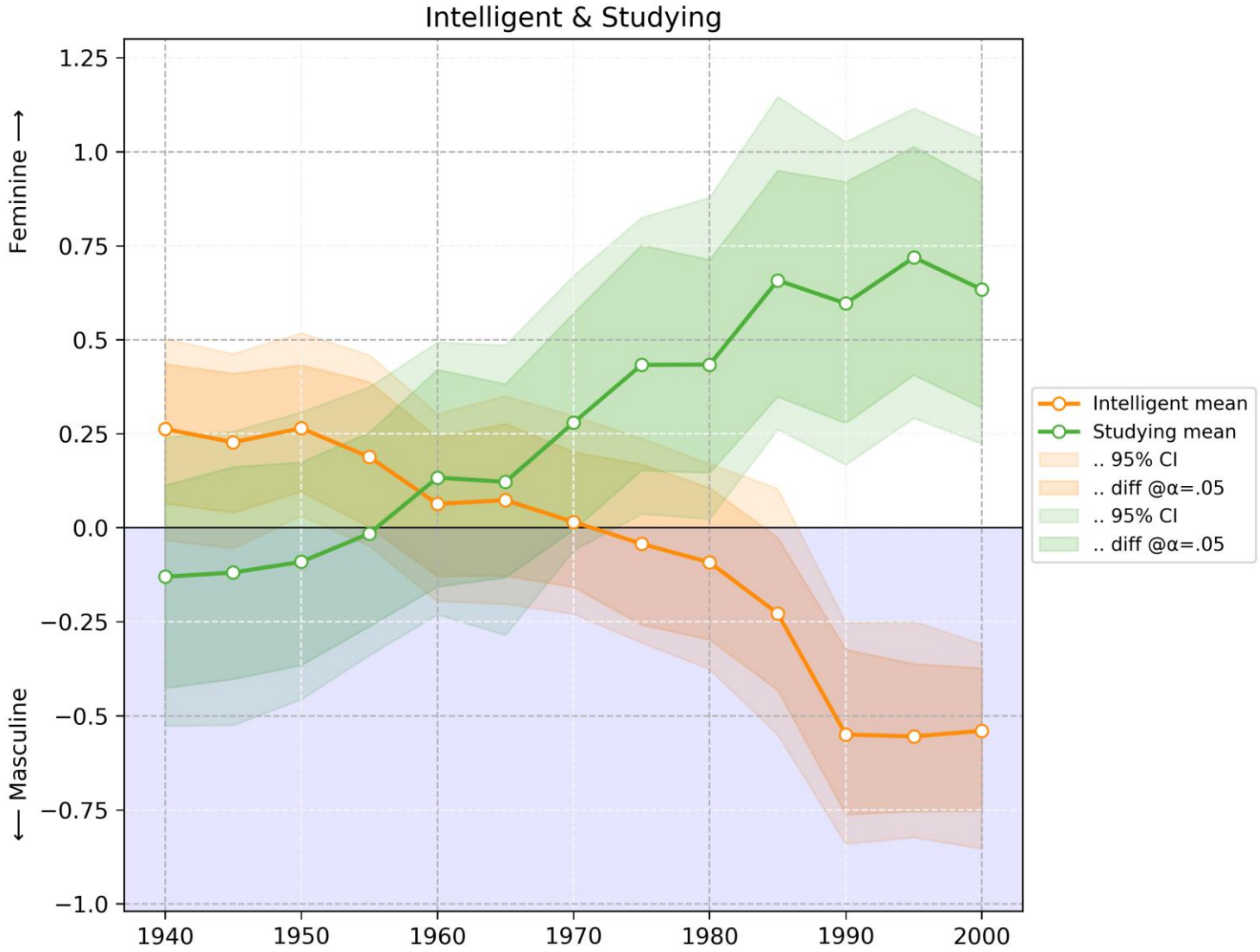


Figure 8. Gendered associations (y) of intelligence and school effort (studying) across time (x). Lighter CI is for α = .05 comparisons against 0. Darker CI is for approx. α = .05 comparisons against a different time point, or across the two scales.



SUPPLEMENTARY MATERIALS

Appendix A. Filtering Keywords for Inclusion in Scales

After assembling the initial term lists as described in the main text, we filtered the candidate terms according to two criteria designed to ascertain their quality. The first is a *frequency criterion*, which ascertains that the terms in our scales are frequent enough to produce robust embedding estimates. We set the minimum frequency threshold to 100, keeping only those terms that appeared at least 100 times in *each* of our 13 corpus time windows. (Note that, due to the power-law distribution of word frequencies, this restriction is more constraining than it may first appear).

The second is an *intended usage criterion* designed to account for polysemy and the difference between words' definitions and empirical usage. For this criterion, we hand-validated most of the keywords that passed the frequency test. For the hand-validation, we constructed a codebook based on the dictionary definition of each keyword, where we indicated which of its meanings fits our scale and which does not. Two undergraduate research assistants then drew a random sample of corpus passages containing the candidate keyword, read those passages, and used the codebook to code whether the terms were being used in intended ways. They also flagged any unexpected patterns of word usage.

To determine where to place the rejection threshold for the intended usage criterion, we estimated the across-time gender projections of all the candidate terms, and examined whether words with imperfect hand-validation scores still followed the same trends as other terms in their scale. We found that the terms' gender projection trends generally retained a high correlation with the trends for the other terms in their scale as long as their intended usage percentage was above 25%. We thus set 25% as our cutoff point for the second criterion. Using a higher

threshold results in smaller term lists and thus wider margins of errors around our estimates, but otherwise does not substantively affect the results. Most of the terms in our final scales easily cleared the threshold for this criterion: the mean and median percentages of times our included keywords were used in intended ways were 75% and 82%, respectively.

We will now discuss our reasoning behind each set of keywords.

"Socio-behavioral skills" and "problem behaviors." To select initial terms for these scales, we began with concepts measured on the Early Child Longitudinal Study – Kindergarten (ECLS-K), which is widely analyzed in academic work on gender differences in academic performance (DiPrete and Buchmann 2013; Owens 2016). Initial terms for (i) socio-behavioral skills included attributes directly relevant to learning ("attentive," "listens," "careful," "meticulous"), and attributes more generally associated with good academic performance ("polite," "courteous"). Initial terms for (ii) problem behaviors captured emotional attributes ("anger," "violent," "rebellious," "uncontrollable"), actions ("disrupted," "interrupted," "attack," "fight"), and classifications of deviance ("rebel," "bully"). These behaviors are negatively associated with academic performance, motivation, and orientation towards school (DiPrete and Buchmann 2013).

"Schooling" and "school effort." The impetus behind our (iii) general "schooling" scale comes from accounts arguing that closeness to academic institutions may be stereotyped as feminine. When assembling this scale's initial term list, we aimed for breadth, and thus included social roles like "student," abstract concepts like "educational" or "academic", physical spaces and items like "classroom" and "blackboard," educational institutions like "school" and "college," and important features of schooling like "classes," "grades" and "graduation." We excluded terms for teachers and teaching because the ratio of female to male teachers increased

during our period, potentially confounding those terms. Our general scale for schooling also excluded terms directly related to studying or schoolwork, which we instead used to construct our fourth scale: (iv) "school effort." Our initial term list for school effort consisted of words directly related to studying ("study," "studying," "studious") and effort ("effortful"). We also included terms closely related to effortful academic activity ("homework," "textbook").

"Intelligence" and "unintelligence." Our initial term list for (v) intelligence included adjectives describing above-average intelligence ("smart," "clever," "intelligent") and exceptional intelligence ("brilliant," "genius"). Initial terms for (vi) unintelligence include demeaning terms ("stupid," "dumb," "idiotic") and terms related to foolishness ("fool," "fools"). In Appendix B, we establish that these scales are robust to alternate choices of keywords.

Appendix tables A1 through A6 list the full initial term lists for each scale, indicate which terms were included in the final scale, and which criteria led to the rejection of the terms that did not make it into that scale.

[Tables A1-A6 about here]

References

DiPrete, Thomas A., and Claudia Buchmann. 2013. *The Rise of Women: The Growing Gender Gap in Education and What It Means for American Schools*. Russell Sage Foundation.

Owens, Jayanti. 2016. "Early Childhood Behavior Problems and the Gender Gap in Educational Attainment in the United States." *Sociology of Education* 89(3):236–58. doi: 10.1177/0038040716650926.

Appendix B. Alternate Keyword Selections for "Intelligence" and "Unintelligence" Scales

In this appendix, we ascertain that our findings for the "intelligence" and "unintelligence" scales are robust to alternate choices of keywords.

Intelligence

Prior scholarship has noted that brilliance and genius appear to have stronger masculine associations than more 'ordinary' high intelligence (e.g., Eagly et al. 2020). We thus examine whether our intelligence scale would still yield the same results if we split the scale into two: a "brilliant" scale consisting of terms for brilliance and genius, and a "smart" scale consisting of the remaining terms. We then calculated the gender projections for these scales.

[Figure B1 about here]

The results are depicted in Figure B1. As the figure indicates, the "brilliant" scale indeed appears consistently more masculine than the "smart" scale, as we expected from prior scholarship. Despite this difference, the figure also shows that both scales exhibited parallel changes across the time periods we examined, becoming significantly more masculine across time (for brilliant, delta = -0.855, $p < 0.01$; for smart, delta = -0.78, $p < 0.01$). Across the decades, the "brilliant" and "smart" scales were correlated with one another at $r = 0.98$ ($p < 0.01$). They also both retained significant negative correlations with school effort ($r = -0.92$, $p < 0.01$ and $r = -0.93$, $p < 0.01$, respectively). Results for both scales thus point to the same conclusion: between 1940 and 2000, intelligence—whether conceptualized as brilliance or as ordinary smarts—appears to have become significantly more masculine in its gender associations.

Unintelligence

The word "fool" may have specifically masculine connotations. To ascertain that this is not unduly influencing our results, we split our unintelligence scale into a "foolish" scale consisting

of the keywords “fool”, “fools” and “foolish”, and an “unintelligent-but-not-foolish” scale consisting of the remaining keywords (“idiot, idiotic, stupid, stupidly, dumb, dummy”). There may also be terms for unintelligence with specifically feminine connotations which we did not include in our original unintelligence scale. These terms often appear to be ways of dismissing a woman as insufficiently serious—e.g., “frivolous” or “silly”. We thus constructed a “frivolous” scale consisting of the terms “frivolous”, “silly”, "irrational", "unreasonable", "crazy", "insane", "hysterical", and "senseless."

Results for the “foolish”, “unintelligent-but-not-foolish” and “frivolous” scales are depicted in Figure B2. As expected, in 1940, the “frivolous” scale was more feminine than the other two scales (mean = 1.145), “unintelligent-but-not-foolish” was in the middle (mean = 0.665), and finally “foolish” was the most masculine (mean = 0.075). The figure further shows that, between 1940 and 2000, the three scales moved in parallel to become significantly more masculine (respectively: delta = -0.783, $p < 0.01$; delta = -1.112, $p < 0.01$, and delta = -0.849, $p < 0.05$). Indeed, the three scales retain approximately the same relative gendering in 2000 as they did in 1940. Across the decades, the “foolish”, “unintelligent-but-not-foolish” and “frivolous” scales are correlated with one another at $r \geq 0.94$ ($p < 0.01$), and with the “intelligence” scale at $r \geq 0.97$ ($p < 0.01$). All three trends thus point to the same substantive conclusion as our primary analyses: by 2000, unintelligence—however conceptualized—appears to have become significantly more associated with masculinity than it was in 1940.

[Figure B2 about here]

Conclusions

In this appendix, we constructed five alternate scales for “intelligence” and “unintelligence” that we designed to be either more feminine or more masculine than our original scales. We found

that all five scales exhibited the same over-time trajectories as one another, moving in parallel towards masculinity. Our results thus appear reassuringly robust to specific choices made during scale construction. This suggests that our findings indeed correspond to broad latent shifts in the gendered meanings of intelligence and unintelligence rather than more surface shifts in specific keywords.

Appendix C. Part-of-speech Composition of the Keyword Scales

Our scales are composed of words belonging to different parts of speech (POS)—nouns, adjectives, verbs, and adverbs. If some parts of speech became systematically more feminine or masculine over time, then the differences in the POS composition of our scales could potentially provide a counter-explanation for our results. In this appendix, we examine and rule out this possibility. Our analyses below consist of three steps. First, we construct scales from large dictionaries of nouns, adjectives, verbs, and adverbs to examine whether they indeed differ in their changing gender associations, and to estimate the average change in gender associations for each part of speech. Second, we take the four scales that showed statistically significant changes in our primary analyses (schooling, school effort, intelligence, and unintelligence), calculate their part-of-speech composition, and estimate the predicted changes in their gendering based solely on their part-of-speech composition. Third, we compare these predictions to the actual observed changes for each of these scales to examine what proportion (if any) of the observed change in each scale could be explained by its part-of-speech composition. To preview, we find that part of speech composition can only explain between 0% and 13.8% of the observed change in the gender associations of our scales. Moreover, it is no better than chance at predicting whether a given scale would become more feminine or masculine over time. We thus conclude that our results are not an artifact of the part-of-speech composition of our scales.

Changing associations of each part of speech

To arrive at the average gendering changes for each part of speech, we began with dictionaries of adjectives, nouns, adverbs and verbs from the Moby Project (Ward 1996). From this we created four keyword scales composed of words that can be used *exclusively* as adjectives, nouns, adverbs, and verbs (and not interchangeably as two or more parts of speech). We then filtered these keyword scales to retain only keywords that occurred at least 100 times in each of our 13 corpus windows. This left four POS scales consisting of 1427 adjectives, 2030 nouns, 874 verbs, 494 adverbs, respectively.

Next, we calculated across-time projections for these four POS scales to estimate the average changes in gendering for each part of speech between 1940 and 2000 (see Appendix Table C1). As these results indicate, only the adjectives scale exhibited a statistically significant change in gendering between these corpus windows (delta = 0.204, $p < 0.05$), and even this change was relatively minor in magnitude. The other three parts of speech exhibited even smaller changes (nouns: -0.078, *n.s.*; verbs: 0.049, *n.s.*, adverbs: 0.174, *n.s.*).

[Appendix Table C1 about here]

Predictions based on scales' POS composition

Next, we took each of the four keyword scales that exhibited a statistically significant change in our primary analyses (schooling, school effort, intelligence, and unintelligence), and hand-tagged their part-of-speech compositions. The results of this tagging are shown in Appendix Table C2.

[Appendix Table C2 about here]

We then used the part-of-speech composition of the four keyword scales (Table C2) and the average gendering changes exhibited by each part of speech (Table C1) to arrive at predictions regarding the expected change in the gendering of each keyword scale. Appendix

Table C2 contains this predicted change in gendering for each of the four keyword scales. For example, in our primary empirical analyses, we observed that, between 1940 and 2000, the Schooling scale moved 0.817 standard deviations towards femininity. As Table C2 indicates, the Schooling scale consists of 16.7% adjectives, 79.2% nouns, 4.2% verbs, and 0% adverbs. The "predicted change" column in table C3 thus indicates that its predicted change in gendering based on this part-of-speech composition is (-0.140 * 0.167) + (–0.072 * 0.792) + (0.014 * 0.042) + (-0.085 * 0) = -0.080. In other words, based on its POS composition, Schooling could be expected to move 0.080 s.d. towards masculinity between 1940 and 2000.

Appendix Table C3 then contrasts this predicted change in gendering with the actual change in gendering we observed in our primary analyses. For example, the observed change in gendering for the schooling scale was +0.817 (i.e., it moved 0.817 s.d. towards femininity). Note that this observed change is in the *opposite* direction from the predicted change we calculated above (-0.080). The part of speech composition of the Schooling scale therefore cannot explain *any* of the observed change in this scale's gendering.

[Appendix Table C3 about here]

As the "% Explained" column of Table C3 indicates, part of speech composition cannot explain any of the change in gendering in the schooling and school effort scales and can potentially explain only 13.8% and 10.1% of the gendering change in the intelligence and unintelligence scales, respectively. Part of speech composition thus fails to explain 100%, 100%, 86.2%, and 89.9% of the gendering change in these respective scales. Thus, only a minor proportion of the significant change in gender associations we report in our primary analyses could be potentially attributed to the part-of-speech composition of our scales. Indeed, the predicted changes based on POS composition only matched the observed direction of the

changes for 2 of the 4 scales (intelligence and unintelligence), and pointed in the opposite direction for the other two scales (schooling and school effort). The POS composition of our four primary scales is thus no better than chance at predicting whether those scales would feminize or masculinize. We thus conclude that our results are not driven by the distribution of adjectives, nouns, verbs, or adverbs that compose our scales.

References

Ward, Grady. 1996. "Moby Lexical Resources." *Speech at Carnegie Mellon University*. Retrieved December 17, 2021 (http://www.speech.cs.cmu.edu/comp.speech/Section1/Lexical/moby.html).

APPENDIX TABLES

Appendix Table A1. All candidate keywords we considered for Schooling scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall
academic	✓	✓	✓	classmates	X		X
education	✓	✓	✓	schoolmates	X		X
educational	✓	✓	✓	sophomore	X	✓	X
class	✓	✓	✓	diploma	X	✓	X
classes	✓	✓	✓	diplomas	X		X
school	✓	✓	✓	graduation	✓	✓	✓
schools	✓	✓	✓	graduate	✓	✓	✓
schooling	✓	✓	✓	graduates	✓	✓	✓
semester	X		X	graduated	✓	✓	✓
semesters	X		X	educated	✓	✓	✓
schoolroom	X		X	grade	✓	✓	✓
schoolyard	X		X	grades	✓	✓	✓
blackboard	✓	✓	✓	exam	X	✓	X
blackboards	X		X	exams	X		X
classroom	✓	✓	✓	quiz	X	✓	X
classrooms	✓	✓	✓	quizzes	X		X
student	✓	✓	✓	college	✓	✓	✓
students	✓	✓	✓	colleges	✓	✓	✓
classmate	X		X	university	✓	✓	✓
schoolmate	X		X	universities	✓	✓	✓

Note: ✓ = yes, X = no. Blank entries in the "Usage" criterion mean we did not hand-validate that keyword.

Appendix Table A2. All candidate keywords we considered for Socio-Behavioral Skills scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall
attentive	✓	✓	✓	conscientious	✓		✓
attentively	X		X	conscientiously	X		X
attention	✓	✓	✓	polite	✓	✓	✓
listening	✓	✓	✓	politeness	✓		✓
listens	✓	✓	✓	politely	✓		✓
listened	✓	✓	✓	careful	✓	✓	✓
courteous	✓	✓	✓	carefully	✓		✓
courteously	X		X	meticulous	✓		✓
cooperative	✓	X	X	meticulously	X		X
cooperatively	X		X				

Note: ✓ = yes, X = no. Blank entries in the "Usage" criterion mean we did not hand-validate that keyword.

Appendix Table A3. All candidate keywords we considered for Problem Behaviors scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

	Count	Usage	Overall		Count	Usage	Overall		Count	Usage	Overall
rough	✓	X	X	attacking	✓		✓	disobeys	X		X
anger	✓		✓	roughhouse	X		X	disobeyed	X		X
angry	✓		✓	roughhousing	X		X	disobeying	X		X
aggression	✓		✓	quarrelsome	X		X	disobedience	X		X
aggressive	✓	✓	✓	quarrel	✓		✓	disrupt	X		X
aggressiveness	X		X	quarrels	X		X	disrupting	X		X
aggressively	X		X	quarreling	X		X	disrupts	X		X
violent	✓		✓	interrupt	✓	✓	✓	disruption	X		X
violence	✓		✓	interrupted	✓		✓	disrupted	✓		✓
violently	✓		✓	interrupting	✓		✓	disrupter	X		X
destroys	✓	✓	✓	undisciplined	X		X	impulsive	✓		✓
destruction	✓		✓	unruly	X		X	impulsively	X		X
destructive	✓		✓	unruliest	X		X	rebel	✓		✓
fight	✓		✓	disorderly	X		X	rebels	✓		✓
fights	✓	✓	✓	disorderliest	X		X	rebelled	X		X
fighting	✓	✓	✓	stubborn	✓		✓	rebelling	X		X
fought	✓	✓	✓	stubbornly	✓		✓	rebellious	✓		✓
combative	X		X	stubbornest	X		X	rebelliously	X		X
combatively	X		X	stubbornness	X		X	rowdy	X		X
conflict	✓	✓	✓	misbehave	X		X	rowdily	X		X
conflicts	✓		✓	misbehaved	X		X	rowdiest	X		X
conflictual	X		X	misbehaves	X		X	uncontrollable	✓		✓
conflictually	X		X	misbehaving	X		X	unmanageable	X		X
confrontation	X		X	misbehavior	X		X	bully	✓	✓	✓
confrontations	X		X	misbehaviors	X		X	bullies	X		X
confrontationally	X		X	disobedient	X		X	bullied	X		X
attack	✓		✓	disobediently	X		X	bullying	X		X
attacks	✓		✓	disobey	X		X				

Note: ✓ = yes, X = no. Blank entries in "Usage" mean we did not hand-validate that keyword.

Appendix Table A4. All candidate keywords we considered for School Effort scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall
homework	X	✓	X	studies	✓	✓	✓
homeworks	X		X	studied	✓	✓	✓
textbook	✓	✓	✓	studious	X		X
textbooks	✓	✓	✓	effortfully	X		X
study	✓	✓	✓	effortful	X		X
studying	✓	✓	✓				

Note: ✓ = yes, X = no. Blank entries in "Usage" mean we did not hand-validate that keyword.

Appendix Table A5. All candidate keywords we considered for Intelligence scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall
genius	✓	✓	✓	intelligently	X		X
geniuses	X		X	cleverly	✓	✓	✓
brilliant	✓	✓	✓	smarter	✓	✓	✓
brilliance	✓	✓	✓	cleverer	X		X
smartest	X		X	savvier	X		X
ingenious	✓		✓	intelligence	✓	✓	✓
ingeniously	X		X	smarts	X		X
cleverest	X		X	aptitude	X		X
savviest	X		X	intellect	✓	✓	✓
brainiest	X		X	acumen	X		X
polymath	X		X	precocious	X		X
polymaths	X		X	precocity	X		X
smart	✓	✓	✓	brainy	X		X
intelligent	✓	✓	✓	brainier	X		X
clever	✓		✓	intellectual	✓		✓
savvy	X		X	intellectually	✓		✓
insightful	X		X				

Note: ✓ = yes, X = no. Blank entries in the "Usage" criterion mean we did not hand-validate that keyword.

Appendix Table A6. All candidate keywords we considered for Unintelligence scale. Columns show whether they satisfied the word "Count" criterion and the hand-coded "Usage" criterion. "Overall" column indicates whether the keywords entered the final scale.

Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall	Keyword	Count	Usage	Overall
stupid	✓	✓	✓	dimwits	X		X	thickheaded	X		X
stupidly	✓		✓	blockhead	X		X	thickheads	X		X
stupidity	✓		✓	blockheaded	X		X	featherhead	X		X
stupidest	X		X	blockheads	X		X	featherheaded	X		X
dumb	✓	✓	✓	dolt	X		X	featherheads	X		X
dumbest	X		X	dolts	X		X	fathead	X		X
idiotic	✓		✓	oaf	X		X	fatheads	X		X
idiot	✓	✓	✓	oafs	X		X	fatheaded	X		X
idiocy	X		X	simpleton	X		X	jackass	X		X
idiotically	X		X	simpletons	X		X	jackasses	X		X
fool	✓	✓	✓	knucklehead	X		X	cretin	X		X
fools	✓	✓	✓	knuckleheaded	X		X	cretins	X		X
foolish	✓	✓	✓	knuckleheads	X		X	halfwit	X		X
dummies	X		X	bonehead	X		X	halfwitted	X		X
dummy	✓		✓	boneheads	X		X	halfwits	X		X
dunce	X		X	boneheaded	X		X	nitwit	X		X
dunces	X		X	ignoramus	X		X	nitwits	X		X
moron	X		X	ignoramuses	X		X	nitwitted	X		X
morons	X		X	numskull	X		X	nincompoop	X		X
imbecile	X		X	numskulls	X		X	nincompoops	X		X
imbeciles	X		X	numbskull	X		X	harebrain	X		X
dimwit	X		X	numbskulls	X		X	harebrained	X		X
dimwitted	X		X	thickhead	X		X				

Note: ✓ = yes, X = no. Blank entries in the "Usage" criterion mean we did not hand-validate that keyword.

Appendix Table C1: Average gendering of each part of speech in COHA and the change in this gendering between 1940 and 2000.

Year	Adjectives	Nouns	Verbs	Adverbs
1940	0.344	-0.005	0.035	0.259
1945	0.353	-0.003	0.054	0.285
1950	0.378	0.002	0.041	0.280
1955	0.330	-0.013	0.019	0.214
1960	0.310	-0.017	0.033	0.201
1965	0.319	-0.006	0.052	0.180
1970	0.307	-0.027	0.076	0.213
1975	0.315	-0.028	0.114	0.240
1980	0.318	-0.023	0.083	0.263
1985	0.272	-0.039	0.034	0.241
1990	0.199	-0.053	0.038	0.173
1995	0.193	-0.065	0.025	0.182
2000	0.204	-0.078	0.049	0.174
Change	-0.140*	-0.072	+0.014	-0.085

Note: *: $p < 0.05$, **: $p < 0.01$.

Appendix Table C2: Part-of-speech composition of each keyword scale

Keyword Scale	Adjectives	Nouns	Verbs	Adverbs	Predicted change
Schooling	16.7%	79.2%	4.2%	0%	-0.080
School Effort	0%	33.3%	66.7%	0%	-0.015
Intelligence	53.9%	30.8%	0%	15.4%	-0.111
Unintelligence	45.5%	45.5%	0%	9.1%	-0.104

Appendix Table C3: Percentage of the observed change in each scale that could be explained by its part-of-speech composition

Keyword Scale	Predictions based on scale POS		Observed results		% Explained	% Unexplained
	Predicted change	Direction	Change	Direction		
Schooling	-0.080	More masculine	+0.817	More feminine	0%	100%
School Effort	-0.015	More masculine	+0.764	More feminine	0%	100%
Intelligence	-0.111	More masculine	-0.803	More masculine	13.8%	86.2%
Unintelligence	-0.104	More masculine	-1.033	More masculine	10.1%	89.9%

APPENDIX FIGURES

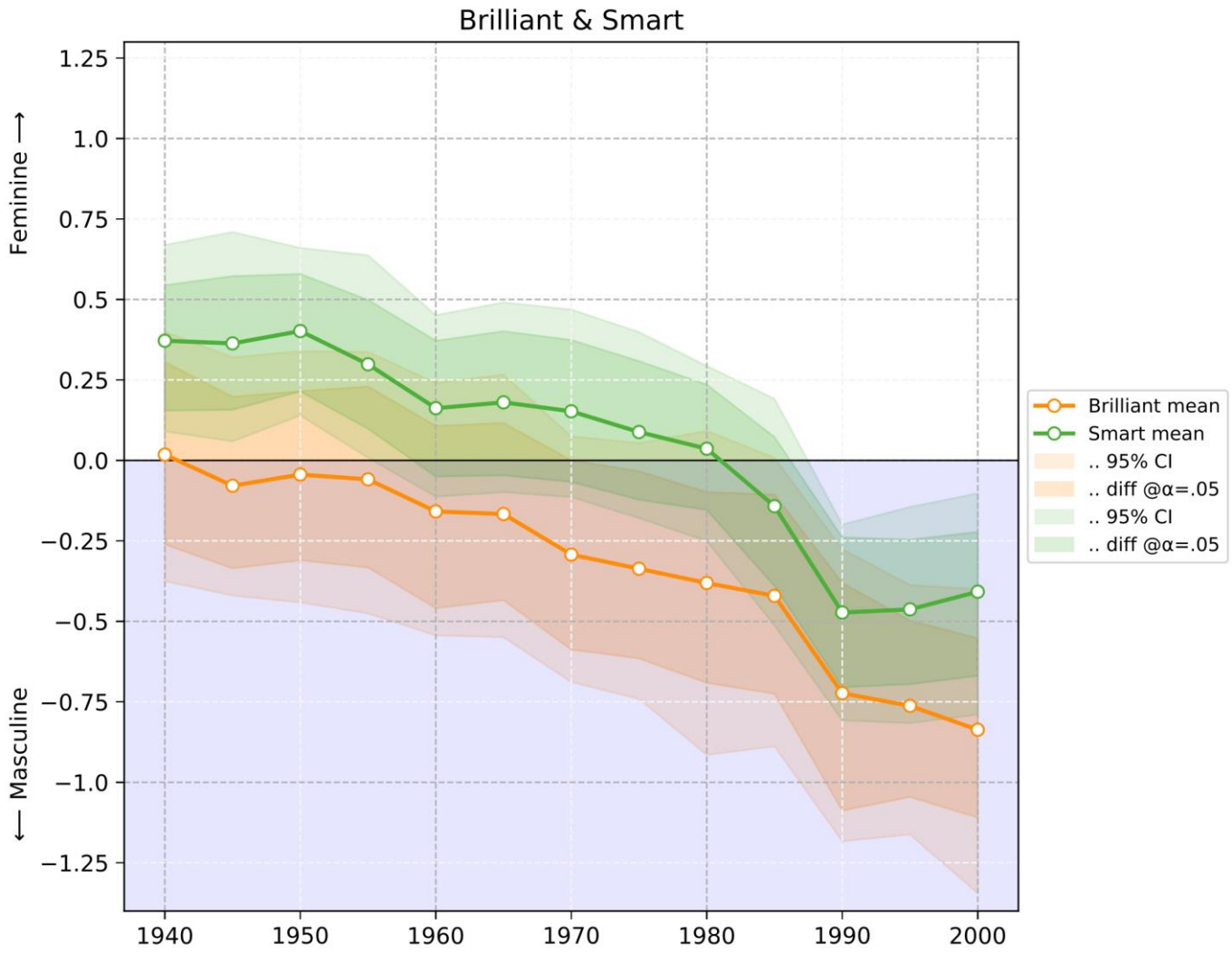


Figure B1. Trends in the gendering of scales for brilliance (orange) versus regular intelligence (green).

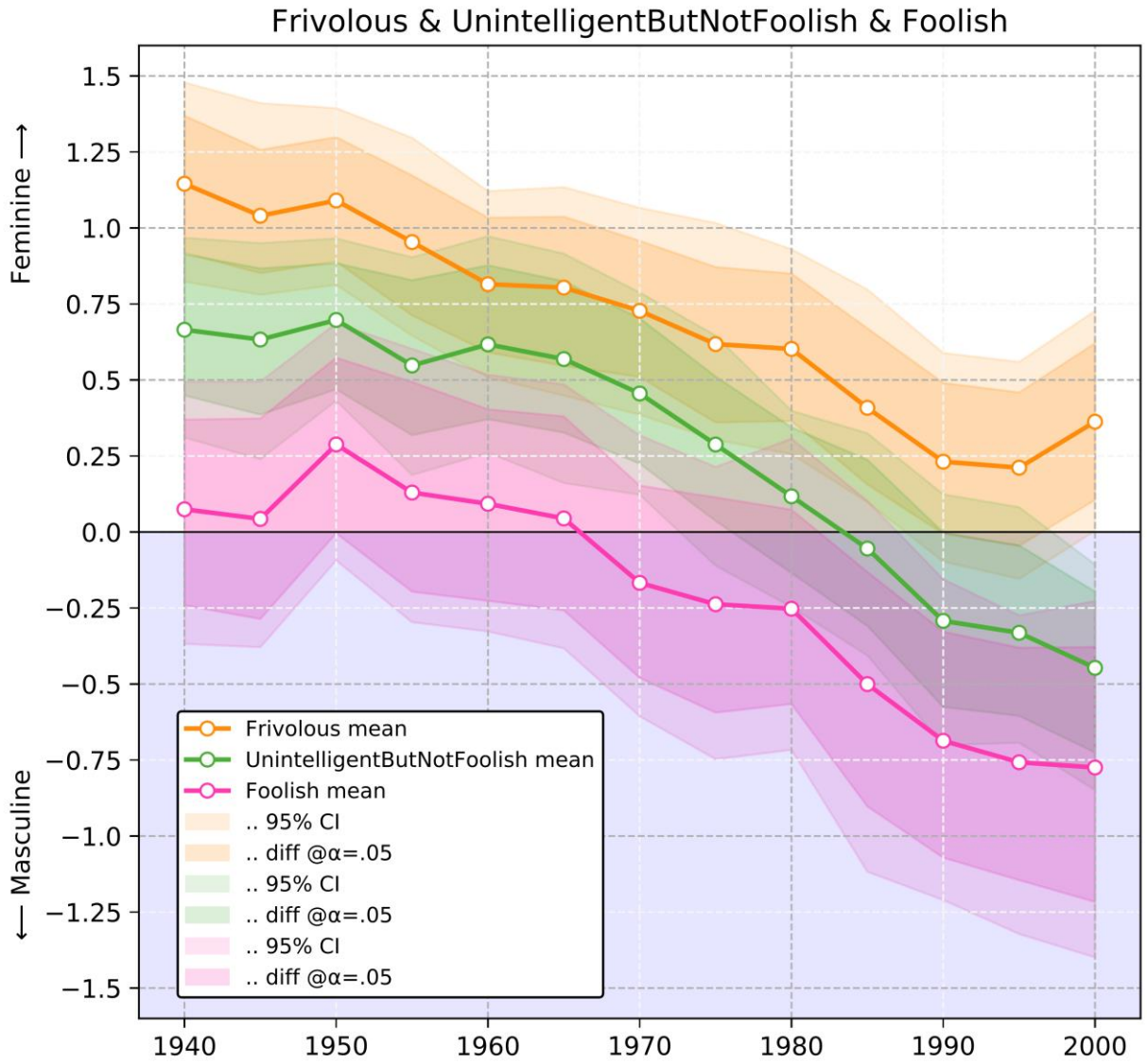


Figure B2. Trends in the gendering of foolishness (pink), unintelligence-but-not-foolishness (green), and frivolousness (orange).