Bortnikova, Kseniya; Havranek, Tomas; Irsova, Zuzana

Working Paper

Beauty and Professional Success: A Meta-Analysis

Beauty and Professional Success: A Meta-Analysis*

Kseniya Bortnikova[a], Tomas Havranek[a,b,c], and Zuzana Irsova[a]

[a]Charles University, Prague

[a]Centre for Economic Policy Research, London

[a]Meta-Research Innovation Center, Stanford

April 5, 2024

Abstract

Common wisdom suggests that beauty helps in the labor market. We show that two factors combine to explain away the mean beauty premium reported in the literature. First, correcting for publication bias reduces the premium by at least a third. Second, controlling for cognitive ability negates the premium for all occupations except sex workers, a point further underscored by the similarity of the beauty effect on earnings and productivity. The second factor implies a positive link, perhaps genetic, between beauty and intelligence. We find little evidence of substantial attenuation bias that could offset publication and omitted-variable biases. The empirical literature is inconsistent with discrimination based solely on tastes for beauty. To obtain these results we collect 1,159 estimates of the effect of beauty on earnings or productivity reported in 67 studies and codify 33 aspects that reflect estimation context, including the potential intensity of attenuation bias. We employ recently developed techniques to account for publication bias and model uncertainty.

Keywords: Beauty premium, productivity, meta-analysis, model uncertainty, publication bias

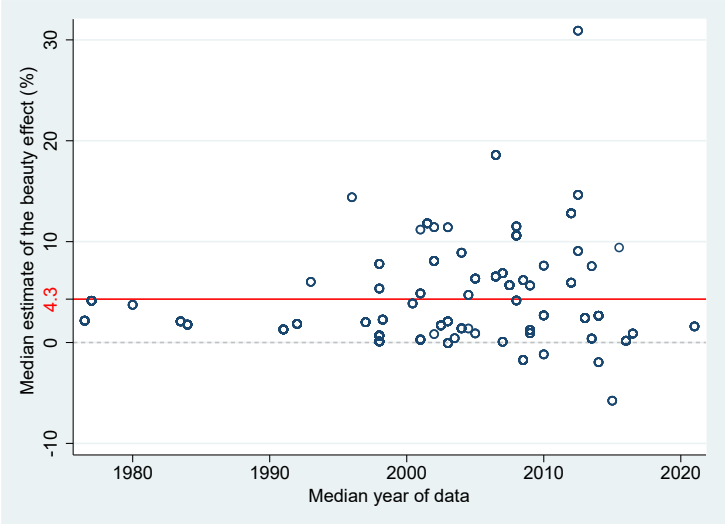
JEL Codes: C83, J24, J31

*Corresponding author: Zuzana Irsova, `irsova.com`. Contact email: `zuzana.irsova@ies-prague.org`.
Data and code are available in an online appendix at `meta-analysis.cz/beauty`.

1 Introduction

According to the authoritative survey by Hamermesh (2011, p. 55–56), people in the top third of looks earn about 5% more compared with average-looking people. We find a remarkably similar mean figure across 1,159 estimates reported in 67 studies published by 2024: moving up along the distribution of beauty by one standard deviation (e.g., from the 50th to the 84th percentile) is on average associated with an increase in earnings by 4.3%. But as Figure 1 shows, individual studies have yielded increasingly divergent results, from −5% to 30%. What explains the dispersion in results? Does the robust mean correlation imply that employers discriminate based on their taste for employees' beauty? The answers have consequences for anti-discriminatory legislation and compensations after accidents damaging looks. In his superb narrative survey Hamermesh (2011) does not make strong conclusions on these questions. After more than a decade, enough studies have been published to allow us to examine both questions formally using meta-analysis, the quantitative method of research synthesis.

Figure 1: Reported beauty premiums diverge



Notes: The horizontal axis shows the median year of data used in a study; the vertical axis shows the median estimate of the beauty premium or penalty reported in the study. All estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. The mean reported effect, denoted as a solid horizontal line, suggests a 4.3% increase in earnings or productivity following an increase in beauty from the 50th percentile to the 84th percentile.

Our main finding is that the correlation between beauty and earnings is not causal. The mean effect reported in the literature, 4.3%, is exaggerated by publication bias. Conservative corrections for publication bias reduce the mean effect to about 2.9%, while other techniques suggest a more aggressive reduction. Controlling for cognitive ability in the primary study or using difference-in-differences diminishes the remaining beauty premium corrected for publication bias to the vicinity of zero—with the exception of prostitutes for which the premium is still substantial. These findings suggest that beauty is not important in the labor market per se but via its correlation with productive characteristics. We corroborate this conclusion by showing that the effect of beauty is similar for earnings and (imperfect) measures of productivity, such as sales, research output, and study outcomes. If the beauty premium reported in the literature was due to taste-based discrimination, we would expect a larger correlation of beauty with earnings than with productivity. Finally, we control for several characteristics reflecting the potential extent of measurement error. We fail to find evidence of substantial attenuation bias that would call for correcting the causal estimate back upwards from about zero.

Two existing studies are especially relevant for our analysis. First, Nault *et al.* (2020) provide an excellent summary of previous meta-analyses related to the effect of beauty on various outcomes and personal characteristics. The meta-analyses, in line with our results and those of Hamermesh (2011), suggest a robust positive correlation between beauty and success. Nevertheless, the previous meta-analyses mostly focus on laboratory experiments in psychology with unclear external validity for the labor market and do not report economic effects (such as the percent increase in earnings following a one-standard-deviation increase in beauty) but standardized coefficients (such as correlation or Cohen's d), which complicates interpretation. Moreover, the meta-analyses neither correct the literature for publication bias nor try to establish a causal link between beauty and earnings. Nault *et al.* (2020) conclude their survey of meta-analyses by observing that, aside from earnings, beauty is also correlated with other characteristics of employees, indicating that the literature is more consistent with statistical than taste-based discrimination.

A second study intimately related to ours is Stinebrickner *et al.* (2019). Using unique data with detailed information on job tasks, the authors show that the beauty premium exists only in occupations where interpersonal interaction is important. Taken together, the results of

Stinebrickner *et al.* (2019), Nault *et al.* (2020), and our study present strong evidence against employer taste-based discrimination in relation to beauty. Our main contribution on top of these and other studies is threefold. First, we present the inaugural meta-analysis of the economics literature on the beauty premium. Second, we use recently developed techniques to correct the literature for publication bias. Third, using methods that address model uncertainty we trace the differences in results to differences in estimation context. Doing so allows us to gauge the effect of omitted variables, measurement error, and other identification issues. In turn, this model enables us to approximate a plausibly causal estimate of the mean beauty premium. In the process we also obtain indirect evidence that beauty is positively correlated with intelligence.

Publication bias describes a situation when reported results represent a systematically different subset of all results obtained by researchers. Ioannidis *et al.* (2017) and Bartos *et al.* (2024) show that the problem is ubiquitous in economics and that the typical economics estimate is exaggerated twofold due to the bias. Other high-quality recent papers document the extent of the problem in economics (Blanco-Perez & Brodeur, 2020; Brown *et al.*, 2024; Card *et al.*, 2018; DellaVigna & Linos, 2022; Elliott *et al.*, 2022; Imai *et al.*, 2020; Neisser, 2021; Stanley *et al.*, 2021; Vivalt, 2019; Xue *et al.*, 2020). It is important to note that publication bias does not imply cheating. Because negative or statistically insignificant estimates of the beauty premium are so unintuitive in most contexts, researchers may take them as evidence of model misspecification or other problems. In many cases they will be right to discard such estimates and try again. The problem is that no upper boundary exists that would mirror the lower bound at zero. Large imprecise estimates tend to be published, small imprecise estimates tend to be discarded. Hence a positive bias arises in the mean reported effect.

Most meta-analysis techniques correct for publication bias by exploiting the property of standard regression analysis: estimates should be independent of their standard errors. If a correlation exists, it is attributable to publication bias: large standard errors, given by noise in data or methods, must be compensated by large point estimates to produce statistical significance. It follows that more precise estimates are less likely to be biased. This is, however, a strong assumption, and a variety of mechanisms can produce a correlation between estimates and standard errors even in the absence of publication bias. For example, Keane & Neal (2023) show that, with instrumental variables, the correlation arises naturally. Method choices may

affect both estimates and standard errors systematically. Precision can be p-hacked (changed by changing specification in order to achieve statistical significance), which introduces reverse causality. To address these problems we use the novel estimator due to Irsova *et al.* (2023), which accounts for the potential endogeneity of the standard error and various forms of p-hacking. We also use recently developed nonlinear techniques by Andrews & Kasy (2019), Bom & Rachinger (2019), Furukawa (2020), and Ioannidis *et al.* (2017). Accounting for publication bias reduces the mean beauty premium from 4.3 to 2.9 or less.

Our other major contribution is a detailed examination of the link between estimation context and the beauty premiums reported in the literature. Because randomized controlled trials on the effect of beauty on earnings are infeasible, and convincing instruments are hard to come by (Hamermesh & Abrevaya, 2013), the bulk of the literature relies on ordinary least squares and tries to control for observable characteristics that might be correlated with earnings or beauty. A few studies, such as Mehic (2022), exploit the shift to online learning during the Covid-19 pandemic and employ the difference-in-differences method. We collect 33 variables that reflect the context in which the premiums are obtained: measurement of beauty (e.g. photo-rated vs. interviewer-rated), measurement of success (earnings or different proxies for productivity), data characteristics (e.g. dressy occupations vs. others), method choice (e.g. control for cognitive skills or social skills), and publication characteristics (e.g. publication status and journal impact factor).

Due to the large number of factors plausibly capturing estimation context and leading to different results, we face substantial model uncertainty: it is unclear ex ante which variables should be included in the final model. The natural response to model uncertainty is Bayesian model averaging (BMA, Fernandez *et al.*, 2001; Ley & Steel, 2009; Eicher *et al.*, 2011; Steel, 2020). BMA runs many regressions with different combinations of controls and then makes a weighted average over them with weights proportional to goodness of fit and parsimony. To account for potential collinearity we use the dilution prior (George, 2010), which gives less weight to models with a small determinant of the correlation matrix. Our results suggest that only three variables systematically explain the differences in the reported beauty premiums: 1) the standard error (a proxy for different amount of publication bias across studies), 2) a dummy for prostitutes, and 3) control for cognitive skills. Conditional on correcting for publication

bias, controlling for cognitive skills (or using difference-in-differences), and focusing on other occupations than prostitutes, the implied beauty effect is close to zero.

An important issue in the literature on the beauty premium is attenuation bias (Harper, 2000; Hamermesh & Abrevaya, 2013; Scholz & Sicinski, 2015). While Hamermesh (2011) shows that, across cultures, there is surprising agreement on what it means to be beautiful, some measurement error is inevitable. If the measurement error is classical, techniques such as ordinary least squares will yield beauty premiums that are biased towards zero. Two general strategies to combat the problem have been suggested in the literature. First, instrumental variables: Hamermesh & Crosnoe (2023) instrument children's looks by their mother's. Gu & Ji (2019) use the looks of other blood relatives. Hamermesh & Abrevaya (2013) use lagged attractiveness, and Pfann *et al.* (2000) use expenditure on beauty. While these instruments can also help with other endogeneity issues, we believe that they are most likely to be useful in attenuating attenuation bias. Second, some studies try to limit measurement error by employing a large number of raters or by using software rating. We find no evidence consistent with substantial attenuation bias: IV estimates tend to be similar to OLS estimates and it does not matter on average how many raters the study uses or whether it employs software rating.

As a byproduct, our results contribute to the literature on the relation between beauty and intelligence. Indeed, the bottom line of our analysis is that beauty is a proxy for productive characteristics, especially cognitive ability. Including a control for cognitive ability tends to substantially reduce the reported beauty premium, and the effect of beauty is similar for earnings and productivity. A positive genetic link between beauty and intelligence is biologically plausible in theory (Kanazawa & Kovar, 2004), though the empirical evidence has so far been inconclusive (Kanazawa, 2011; Mitchem *et al.*, 2015).

2 Data

We focus on economics studies conducted in field settings. Laboratory studies on the subject exist (most prominently the maze experiment by Mobius & Rosenblat, 2006), especially in the field of psychology, and have been covered by previous meta-analyses (Nault *et al.*, 2020). We believe these two parts of the literature are best analyzed separately. While laboratory studies on this topic are useful, it is not always clear how their findings translate to real-world behavior

in the labor market. Aside from external validity issues, a practical consideration is that most laboratory experiments do not report enough information that would allow us to convert their results to the percent increase in earnings or productivity following a one-standard-deviation increase in beauty.

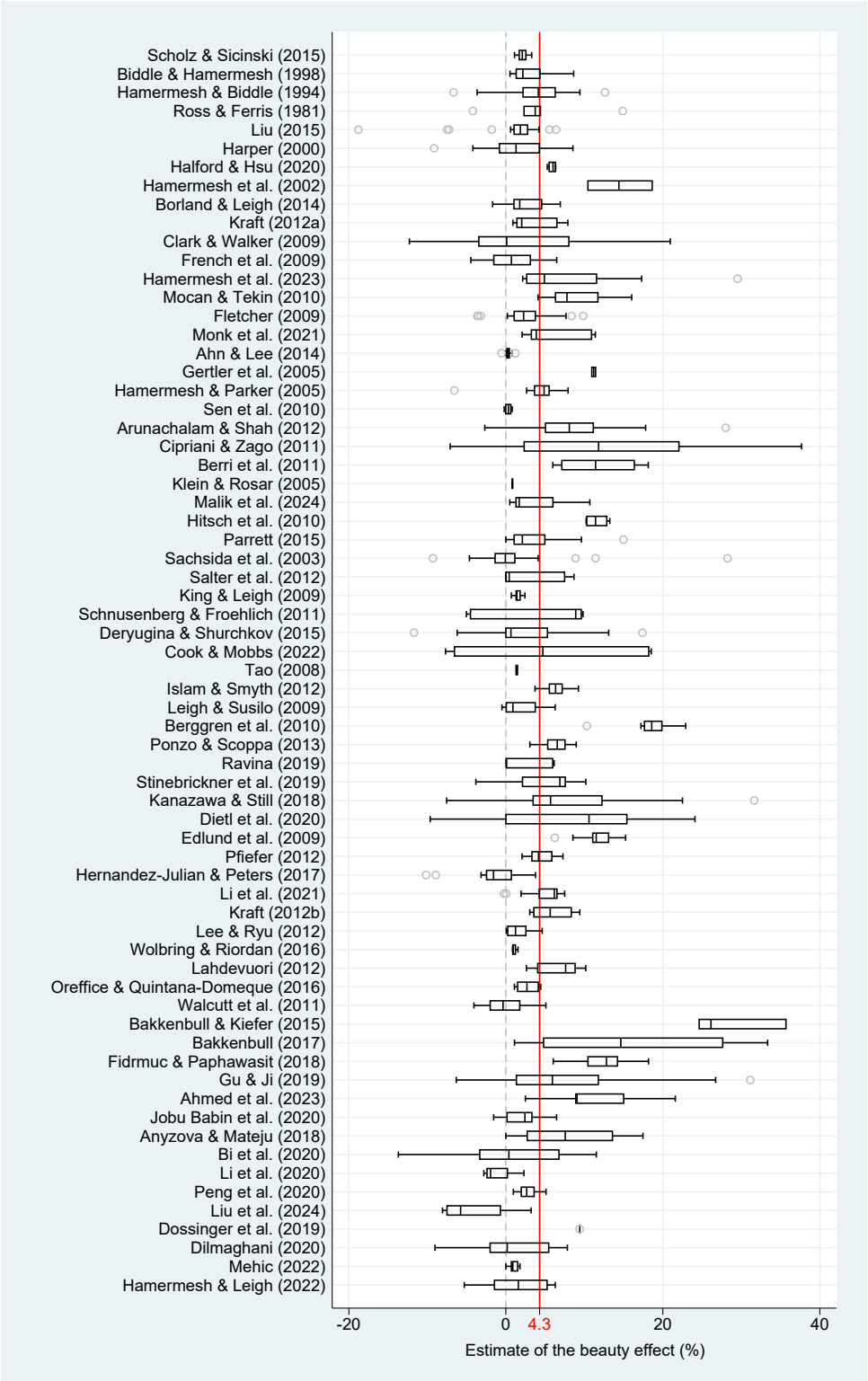
Table 1: The 67 studies included in the meta-analysis

Ahmed *et al.* (2023)	Gu & Ji (2019)	Liu (2015)
Ahn & Lee (2014)	Halford & Hsu (2020)	Liu *et al.* (2024)
Anyzova & Mateju (2018)	Hamermesh & Biddle (1994)	Malik *et al.* (2024)
Arunachalam & Shah (2012)	Hamermesh & Parker (2005)	Mehic (2022)
Bakkenbull & Kiefer (2015)	Hamermesh & Leigh (2022)	Mocan & Tekin (2010)
Bakkenbull (2017)	Hamermesh *et al.* (2002)	Monk *et al.* (2021)
Berggren *et al.* (2010)	Hamermesh & Crosnoe (2023)	Oreffice & Quintana-Domeque (2016)
Berri *et al.* (2011)	Harper (2000)	Parrett (2015)
Bi *et al.* (2020)	Hernandez-Julian & Peters (2017)	Peng *et al.* (2020)
Biddle & Hamermesh (1998)	Hitsch *et al.* (2010)	Pfeifer (2012)
Borland & Leigh (2014)	Islam & Smyth (2012)	Ponzo & Scoppa (2013)
Cipriani & Zago (2011)	Jobu Babin *et al.* (2020)	Ravina (2019)
Clark & Walker (2009)	Kanazawa & Still (2018)	Ross & Ferris (1981)
Cook & Mobbs (2023)	King & Leigh (2009)	Sachsida *et al.* (2003)
Deryugina & Shurchkov (2015)	Klein & Rosar (2005)	Salter *et al.* (2012)
Dietl *et al.* (2020)	Kraft (2012a)	Schnusenberg & Froehlich (2011)
Dilmaghani (2020)	Kraft (2012b)	Scholz & Sicinski (2015)
Dossinger *et al.* (2019)	Lahdevuori (2013)	Sen *et al.* (2010)
Edlund *et al.* (2009)	Lee & Ryu (2012)	Stinebrickner *et al.* (2019)
Fidrmuc & Paphawasit (2018)	Leigh & Susilo (2009)	Tao (2008)
Fletcher (2009)	Li *et al.* (2020)	Walcutt *et al.* (2011)
French *et al.* (2009)	Li *et al.* (2021)	Wolbring & Riordan (2016)
Gertler *et al.* (2005)		

Notes: Details on the literature search and criteria for inclusion are available in Appendix A. The last study was added on February 16, 2024.

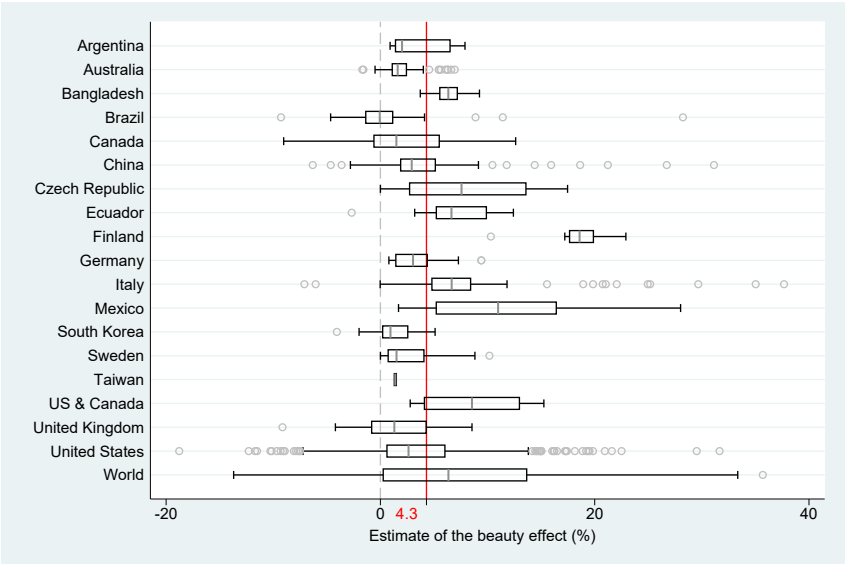
Figure A1 in Appendix A provides details on how we include individual empirical studies in the meta-analysis. We start with a Google Scholar search. We prefer Google Scholar to other databases because it goes through the full text of studies, not just the title, abstract, and keywords, as is the case for many other sources. After identifying potentially usable studies we also do "snowballing" by inspecting the studies frequently cited among the potentially usable ones. Snowballing reduces our dependence on Google Scholar. We use the following inclusion criteria: i) the study must report the effect of beauty on a continuous variable reflecting earnings or productivity in a field (real-world) setting; ii) the beauty measurement used in the study must focus on physiognomy (just the face), iii) the study must focus on the subject's own earnings or productivity, not e.g. the spouse's income, iv) the study must report statistics that allow us to convert the reported estimate to the percent increase in earnings or productivity following

Figure 2: Estimates vary both within and across studies



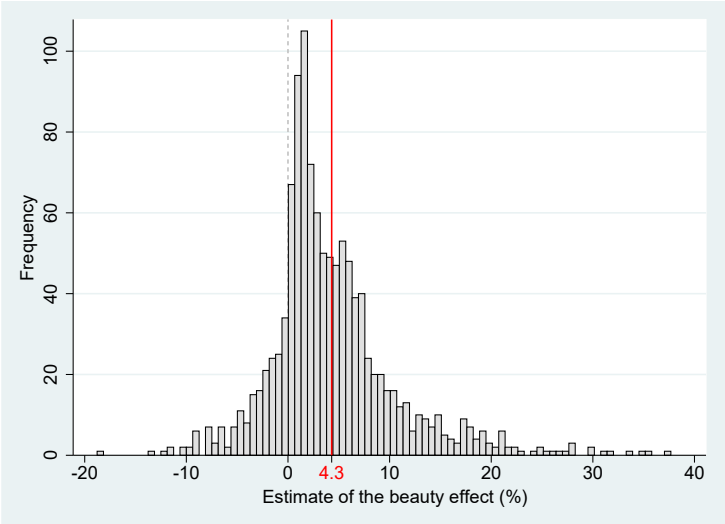
Notes: The figure shows a box plot of the estimated beauty effects. Studies are sorted by data age from oldest to newest. All estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. The length of each box represents the interquartile range (P25-P75), and the dividing line inside the box represents the median. The whiskers represent the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. The mean overall reported effect is denoted as a solid vertical line. For ease of exposition, extreme outliers are excluded from the figure but included in all statistical tests.

Figure 3: Cross-country heterogeneity in beauty premiums



Notes: The figure shows a box plot of the estimated beauty effects. All estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. The length of each box represents the interquartile range (P25-P75), and the dividing line inside the box represents the median. The whiskers represent the highest and lowest data points within 1.5 times the range between the upper and lower quartiles. The mean reported effect is denoted as a solid vertical line. For ease of exposition, extreme outliers are excluded from the figure but included in all statistical tests.

Figure 4: Small positive estimates are most common



Notes: The figure depicts a histogram of the estimated beauty premium or penalty reported in individual studies. All estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. The mean reported effect, denoted as a solid vertical line, suggests a 4.3% increase in earnings or productivity following an increase in beauty from the 50th percentile to the 84th percentile.

a one-standard-deviation increase in beauty; v) the study must report standard errors or other statistics from which standard errors can be computed.

The five criteria leave 67 studies listed in Table 1; we call them primary studies, and together they provide 1,159 estimates of the beauty effect. Each study typically reports many estimates: for example using OLS vs. IV, men vs. women, results for different occupations, etc. Later in the analysis, and most prominently in Section 4, we control for 33 characteristics that reflect the context in which the estimate was produced in the primary study. Figure 2 shows a box plot of the reported estimates. The studies in the figure are ranked by the age of the data they use from oldest to newest. There is no apparent trend in the findings. Most studies report at least some estimates that are close to the overall mean beauty premium, 4.3%. Many studies report much higher estimates, in several cases above 20%. Perhaps surprisingly, negative estimates of the beauty premium are not entirely rare. Overall we observe substantial variance in results both within and across studies.

Figure 3 shows that the estimated coefficients are typically positive but not huge across countries, with the notable exception of Finland. The figure does not suggest any systematic difference in the beauty premium across cultures or income levels. Figure 4 shows the histogram of the reported beauty effects. Two facts stand out in the figure. First, the distribution is asymmetrical: while many large positive outliers appear in the literature, few estimates are substantially negative. Second, the mode of the distribution is dominated by estimates that are just positive. Both observations are consistent with publication bias, the topic of the next section; but they could also be consistent with systematic heterogeneity.

Table 2 and Figure 5 give a general overview of the heterogeneity in the literature, more fully explored later in Section 4. It does not seem to matter much how beauty is measured. Self-rated measures of beauty tend to be associated with smaller beauty effects, but only a few studies use self-rating. Beauty penalties (comparisons of below-average and average looks) seem to be slightly smaller than beauty premiums (above-average vs. average looks). While in the main analysis we pool both types of estimates together, as a robustness check we also conduct the analysis separately for premiums and penalties. For comparability, we always recompute all estimates to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty.

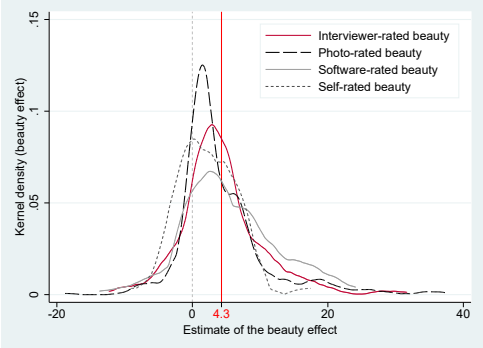
Table 2: Beauty effects in different contexts

			Unweighted			Weighted		
	Est.	Stud.	Mean	95% conf. int.		Mean	95% conf. int.	
Measurement of beauty								
Interviewer-rated beauty	526	24	4.33	3.80	4.86	5.87	5.31	6.42
Photo-rated beauty	481	37	4.28	3.71	4.85	4.82	4.16	5.47
Software-rated beauty	104	8	5.49	4.15	6.83	7.43	6.04	8.82
Self-rated beauty	56	6	2.04	0.79	3.29	2.79	1.32	4.27
Dummy beauty	460	23	3.63	3.05	4.22	5.02	4.39	5.65
Categorical beauty	699	52	4.70	4.24	5.16	5.29	4.76	5.81
Beauty premium	954	67	4.48	4.07	4.89	5.45	5.00	5.91
Beauty penalty	205	16	3.33	2.56	4.10	3.10	2.38	3.82
Measurement of success								
Earnings	688	43	4.33	3.92	4.75	6.15	5.65	6.66
Study outcomes	189	9	3.19	1.98	4.41	3.78	2.54	5.02
Teaching & research outcomes	139	8	4.95	4.05	5.85	3.38	2.36	4.40
Athletic success	33	2	6.28	3.03	9.54	5.36	2.27	8.46
Electoral success	37	4	7.00	4.26	9.75	5.67	3.09	8.24
Other outcomes	73	6	2.98	2.24	3.73	2.19	1.42	2.95
Data characteristics								
Male subjects	375	44	3.66	3.10	4.22	4.28	3.69	4.87
Female subjects	447	48	4.10	3.45	4.75	6.28	5.52	7.04
Mixed gender	337	36	5.21	4.57	5.85	4.86	4.18	5.54
High-skilled workers	335	27	4.30	3.78	4.81	4.00	3.42	4.58
Prostitutes	55	4	8.55	7.27	9.82	9.54	8.45	10.64
Other dressy occupations	138	15	4.92	3.89	5.96	5.37	4.31	6.43
Non-dressy occupations	966	51	3.94	3.55	4.34	4.82	4.36	5.28
Western culture	874	47	3.85	3.46	4.23	4.62	4.22	5.02
Other cultures	285	21	5.60	4.72	6.48	6.55	5.54	7.56
Panel data	998	55	3.86	3.47	4.25	4.82	4.38	5.25
Cross-section	161	13	6.87	5.90	7.83	6.87	5.87	7.88
Estimation technique								
Ordinary least squares	903	60	4.17	3.77	4.56	5.19	4.75	5.64
Instrumental variables	83	10	4.86	3.07	6.64	6.60	4.70	8.50
Difference-in-differences	12	2	1.24	0.63	1.84	1.13	0.55	1.71
Other method	161	13	4.84	3.78	5.90	4.91	3.82	6.00
Cognitive skill control	445	26	2.77	2.17	3.36	3.54	2.90	4.18
No cognitive skill control	714	53	5.22	4.78	5.66	5.94	5.43	6.45
Publication characteristics								
Published study	1,027	58	4.19	3.80	4.57	5.23	4.79	5.67
Unpublished study	132	9	4.99	3.96	6.03	5.06	4.11	6.02
High-quality peer review	151	7	5.40	4.56	6.25	7.53	6.55	8.50
All estimates	1,159	67	4.28	3.92	4.64	5.21	4.81	5.61

Notes: The table reports summary statistics of the estimated beauty effect for subsets of the literature. All estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. Est. = estimates. Stud. = studies. In the left-hand panel (unweighted) simple means and the corresponding 95% confidence intervals are reported; each estimate is assigned the same weight. In the right-hand panel (weighted) estimates are weighted by the inverse of the number of estimates reported per study, thus giving each study the same weight. High-quality peer review = 1 if the study was published in a top 5 journal in economics, the Review of Economics and Statistics, Journal of Labor Economics, or Journal of Public Economics. Details on the definition of subsamples are available in Table 4.

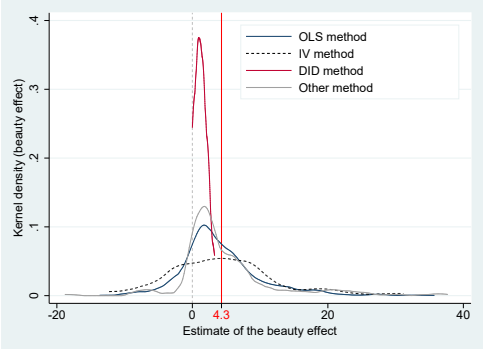
Figure 5: Selected patterns in the literature

(a) Beauty measurement

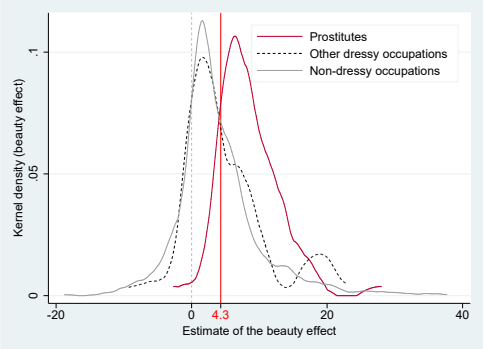


(b) Success measurement

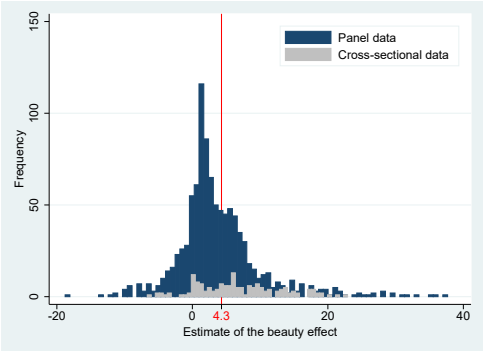
(c) Method choice



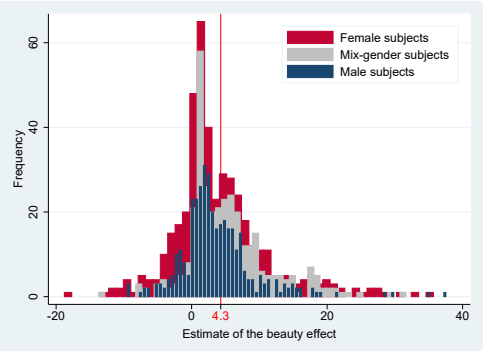
(d) Occupation



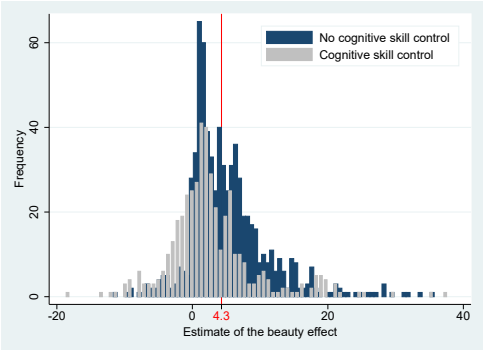
(e) Data dimension



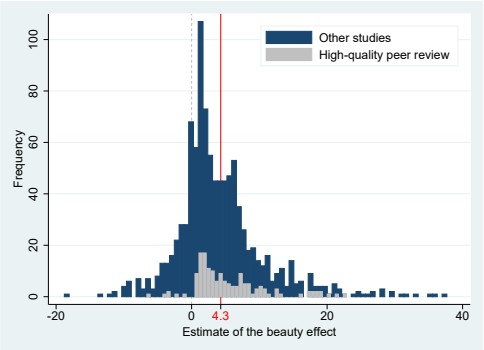
(f) Gender



(g) Ability control



(h) Journal quality



Notes: Estimates are recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty. The mean estimate is denoted as a solid vertical line.

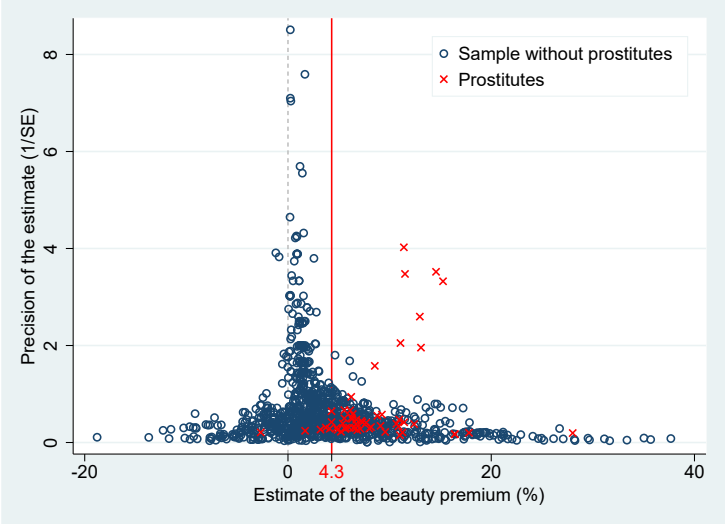
Regarding the measure of success, the response variable in primary studies, the literature can be divided into two groups: studies focusing on earnings and studies focusing on (imperfect) measures of productivity. Productivity in this context can be measured as sales, research outcomes, study outcomes, electoral success, etc. On average, the mean beauty premium for earnings is identical to the mean premium for productivity: 4.3%, which is consistent with no or little taste-based discrimination. Athletes and politicians seem to enjoy relatively large beauty premiums. Not surprisingly, the beauty premium is the largest for sex workers. Other data characteristics, such as gender and culture, do not seem to be associated with substantial systematic differences in the results.

Regarding estimation characteristics, two subsets of the literature stand out: estimates obtained using difference-in-differences and estimates obtained using OLS while controlling for cognitive ability. These two subsets tend to report beauty premiums much smaller than the rest of the literature. The finding, which as we will see will survive correction for publication bias and model uncertainty, suggests that the raw correlation between beauty and earnings is not causal but driven by a correlation between beauty and other productive characteristics, especially intelligence. Nevertheless, Table 2 and Figure 5 also suggest that studies published in top journals typically report relatively large estimates. In the following two sections we examine how these preliminary results are affected by an explicit treatment of publication bias and model uncertainty.

3 Publication Bias

Publication bias, broadly defined, is the difference between the mean reported result and the mean result originally obtained by researchers. The meta-analysis literature sometimes distinguishes between narrowly defined publication bias and p-hacking (Brodeur *et al.*, 2023). When the distinction is made, publication bias denotes the decision to report an estimate or hide it in a file drawer. P-hacking then denotes the process by which researchers adjust their model to make their estimates more publishable—for example, more statistically significant. Given that under extreme p-hacking no limits exist for the resulting estimates, no model can convincingly account for p-hacking. Most meta-analysis techniques were developed with narrowly defined publication bias in mind. Some of them, though, also address many plausible forms of

Figure 6: The funnel plot suggests publication bias



Notes: In the absence of publication bias the most precise estimates should be close to the mean estimate, denoted by the solid vertical line. Less precise estimates should be dispersed symmetrically around the mean effect. The figure indicates that small or negative imprecise estimates are less likely to be reported than similarly imprecise but large and positive estimates. Extreme outliers are excluded from the figure for ease of exposition but included in all statistical tests.

p-hacking, and both problems are often observationally equivalent. We use the term publication bias in its more general meaning, separating it from p-hacking only when necessary.

Figure 6 shows a visual test for publication bias, the so-called funnel plot (Duval & Tweedie, 2000; Stanley & Doucouliagos, 2010). It is a scatter plot of estimated beauty effects on the horizontal axis against their precision (1/SE) on the vertical axis. The most precise estimates at the top of the figure are close to the mean reported effect, while less precise estimates at the bottom are more widely dispersed. Crucially, imprecise estimates larger than the mean should be just as common as imprecise estimates smaller than the mean. A symmetrical inverted funnel shape should follow, and symmetry in the absence of publication bias is the key feature of the funnel plot. Figure 6 suggests that, for the literature on the beauty effect, the funnel plot is not symmetrical. Large imprecise estimates are much more common than small, and especially negative, imprecise estimates. (It is also apparent that the funnel for prostitutes is completely different from the rest of the data, which is why in Appendix B we conduct the analysis separately for the subsample of prostitutes and other occupations.) In other words, estimates are positively correlated with their standard errors. Because the correlation arises if

Table 3: Linear and nonlinear techniques detect publication bias

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.377***	0.208*	0.732***	0.758	0.656**
(*standard error*)	(0.118)	(0.119)	(0.164)	(0.489)	(0.272)
	[0.116, 0.634]			{-0.064, 2.160}	[-0.009, 1.268]
Effect beyond bias	2.865***	3.497***	2.243**	1.436	3.424***
(*constant*)	(0.464)	(0.446)	(0.871)	(1.866)	(1.156)
	[1.916, 3.781]			{-0.378, 3.251}	[0.776, 6.339]
First-stage robust F-stat				15.8	
Observations	1,159	1,159	1,159	1,159	1,159
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.720***			1.720***	P = 0.142
	(0.084)			(0.084)	(0.038)
	[1.262, 2.166]			[1.262, 2.166]	
Effect beyond bias	0.343	0.323**	0.055	0.343	0.493
	(0.118)	(0.147)	(1.276)	(0.118)	(0.410)
	[-0.039, 1.635]			[-0.039, 1.635]	
Observations	1,159	1,159	1,159	1,159	1,159

Notes: Panel A reports the results of regression $\widehat{b}_{ij} = b_0 + \beta \cdot SE(b_{ij}) + \epsilon_{ij}$, where $\widehat{b}_{ij}$ denotes the i-th beauty effect estimated in the j-th study, and $SE(b_{ij})$ denotes its standard error. FE = study-level fixed effects, BE = study-level between effects, MAIVE = Meta-Analysis Instrumental Variable Estimator (Irsova *et al.*, 2023) with the inverse of the square root of the sample size used as an instrument for the standard error. Weighted = the inverse of the number of estimates per study is used as the weight. In Panel B all models are weighted by inverse variance. The first specification reports a regression similar to those from the last column of Panel A but with inverse variance weights. WAAP = Weighted Average of the Adequately Powered estimates (Ioannidis *et al.*, 2017); Stem = the model by Furukawa (2020); Kink = the model by Bom & Rachinger (2019); Selection = the model by Andrews & Kasy (2019). P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). Standard errors, clustered at the study level, are reported in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018) are reported in square brackets. For MAIVE, in curly brackets we show the Anderson-Rubin 95% confidence interval recommended by Keane & Neal (2023). Separate results for the subsamples of prostitutes, other occupations, beauty premiums, and beauty penalties are available in Table B3 and Table B4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

researchers prefer positive or statistically significant estimates, it is synonymous with publication bias in most of the meta-analysis literature.

Specifications in Panel A of Table 3 present the corresponding statistical tests. Because of the relation to the funnel plot, these regressions are often called funnel-asymmetry tests (Card & Krueger, 1995; Egger *et al.*, 1997). In the first column we show a simple OLS regression and find a substantial correlation between estimates and standard errors. The intercept in the regression can be interpreted as the estimated beauty premium conditional on maximum precision (and therefore no publication bias), the top of the funnel, and thus the mean estimate corrected for publication bias (Stanley, 2008). We obtain a value of 2.9, which is 1/3 smaller than the uncorrected mean of 4.3. The corrected mean increases when we include study-level fixed effects (3.5). The fixed-effects estimator only captures decisions within studies, and can be thus interpreted as capturing p-hacking rather than strictly publication bias (Mathur, 2024). Publication bias,

narrowly defined, is captured by the between-effects estimator, which corresponds to selection across studies. The between-effects estimate for the corrected beauty premium is 2.2. Therefore it seems that in this literature publication bias can be more important than p-hacking. Note that techniques based on the funnel plot, unlike other methods reported later in Panel B, are robust to p-hacking on the reported point estimates: even if estimates are artificially large to offset large standard errors, funnel-based techniques are virtually unaffected because they focus on the most precise estimates.

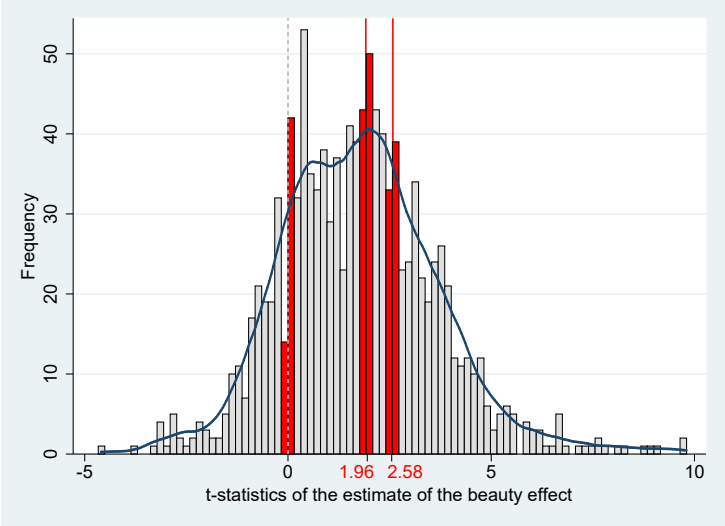
But, in contrast to the common meta-analysis assumption employed in the above-mentioned funnel asymmetry tests, in the literature on the beauty effect some of the correlation between estimates and standard errors can plausibly be unrelated to publication bias. First, Keane & Neal (2023) show that for IV estimates the correlation arises by construction. Second, some method choices can influence both estimates and standard errors. For example, compared to OLS, IV estimates can bring larger point estimates (because they address attenuation bias) but also larger standard errors (because IV estimation tends to be generally less precise). Third, if researchers p-hack standard errors (e.g. by changes in clustering), the correlation resulting from this form of p-hacking is not associated with any bias in the mean reported estimate, and funnel asymmetry tests introduce a downward bias that did not exist before.

In other words, we face a classical endogeneity problem in our meta-analysis specifications. Irsova *et al.* (2023) present a simple solution: the meta-analysis instrumental variable estimator (MAIVE). MAIVE uses the inverse of the square root of the sample size used in the primary study as an instrument for the reported standard error. Sample size is a strong instrument by virtue of the definition of the standard error, which is a function of the sample size. While it does not completely eliminate the endogeneity problem, it alleviates it substantially: researchers find it more difficult to artificially increase sample size than to artificially decrease the standard error; the problem identified by Keane & Neal (2023) does not apply to sample size; and the choice of methods (such as IV vs. OLS) is often unrelated to sample size. The fourth column in Table 3 reports the results of MAIVE. The Anderson-Rubin confidence intervals recommended by Andrews *et al.* (2019) and Keane & Neal (2023) suggest marginal statistical insignificance of publication bias at the 5% level. The point estimate is large, and leads to a large correction in the mean beauty premium to 1.4, with 3.3 as the upper bound of the 95% confidence interval.

The last column of Panel A shows a specification weighted by the inverse of the number of estimates reported per study: in other words, each study now has the same weight. The corrected mean beauty premium is similar to the one previously reported for study-level fixed effects. In Panel B we show models weighted by inverse variance, which is common in meta-analysis but which we avoided in Panel A due to the apparent endogeneity of the standard error in this literature. The first specification is the funnel-asymmetry test similar to those in Panel A but weighted by inverse variance. The remaining models present recently developed nonlinear techniques for publication bias correction. While they relax the assumption that publication bias is a linear function of the standard error, they do not allow for any p-hacking: put differently, these techniques assume that each reported estimate is individually unbiased. All techniques in Panel B yield very small, indeed almost zero corrected beauty premiums, and by comparing the first column in Panel A with the first column in Panel B it seems that the reason for the difference is inverse variance weighting, an inherent part of the nonlinear models in Panel B. Robustness checks reported in Table B3 and Table B4 in Appendix B show that excluding estimates for sex workers or beauty penalties does not change the results.

Another piece of evidence indicates that inverse variance weights are unsuitable for the beauty premium literature. Table B1, reported in Appendix B, shows that, based on the Andrews & Kasy (2019) model, estimates and standard errors are correlated even after correction for publication bias. While the result may indicate that any of the assumptions of the Andrews & Kasy (2019) model are not met, the assumption of no relation between estimates and standard errors in the absence of publication bias is by far the most important assumption (Kranz & Putz, 2022). For this reason we prefer the results reported in Panel A of Table B3 to the precision-weighted specifications reported in Panel B. To be on the safe side, for the representative estimate we choose the median value reported in Panel A: 2.9 corresponding to the simple OLS. This correction for publication bias is conservative given all the estimates in Panel B but especially given the MAIVE estimator that we would otherwise prefer. The advantage of the simple OLS correction is that it can be easily incorporated into the analysis of heterogeneity and model uncertainty in the next section. It is important to keep in mind, however, that the effect of publication bias is probably larger, and the mean corrected beauty effect is much smaller than commonly thought.

Figure 7: Bias caused by selection for a positive sign



Notes: The figure shows the distribution of t-statistics of the reported estimates of the beauty effect. The vertical lines represent the value of 0 associated with a sign change, the critical value of 1.96 associated with significance at the 5% level, and the critical value of 2.58 associated with significance at the 1% level. Bins just below and above the corresponding threshold are highlighted. The zero threshold matters most. We exclude estimates with extremely large t-statistics from the figure for ease of exposition but include them in all statistical tests. The corresponding caliper tests are reported in Table B2.

Figure 7 gives intuition on the sources of publication bias. The figure is a histogram of reported t-statistics, and important thresholds (0, 1.96, 2.58) are highlighted. In all three cases, estimates just above the threshold are more common in the literature, which is again consistent with publication bias. But the jumps at 1.96 and 2.58 are relatively small compared to the jump at zero: estimates that are just positive are much more likely to be reported than estimates that are just negative. Caliper tests presented in Table B2 in Appendix B corroborate these observations and suggest that publication bias is driven by the preference for positive estimates, while the preference for statistically significant estimates plays a relatively less important part.

4 Heterogeneity

This section has two goals. First, we examine whether our findings regarding publication bias are robust to the inclusion of controls reflecting study design. The approach is complementary to that of the MAIVE method introduced in the previous section: MAIVE uses the inverse of the square root of sample size for the reported standard error; now we explicitly control for

as many observable data and method choices as possible. The approach of the current section will dominate MAIVE if sample size is correlated with method choices that influence both the reported estimates and their standard errors. Second, we examine why, on top of differences in the propensity for publication bias, the estimates reported in the literature vary so much.

To achieve both goals we need to capture the main observable differences in estimation context. The corresponding variables are defined and summarized in Table 4. For ease of exposition we divide them into five groups: measurement of beauty, measurement of success, data characteristics, estimation technique, and publication characteristics. The table shows stylized facts regarding the literature. For example, only a few studies use self-rating or computer algorithms to generate beauty ratings; most studies rely on interviewers (45% of the estimates) or humans evaluating photos (42%). Only about 18% of the studies focus explicitly on beauty penalty as opposed to beauty premium. Most studies (59%) focus on earnings, while the rest rely on various, though often imperfect, proxies for productivity. The subjects often recruit from occupations for which beauty should not intuitively represent an important productive factor (83%). Quasi-experimental estimation techniques are quite rare in the literature because of the paucity of convincing instruments and natural experiments—with the exception of the switch to online learning during the Covid-19 pandemic, which allows for difference-in-differences estimation. A substantial number of studies (38%) find a way to control for a proxy for cognitive ability, such as IQ. Almost all studies in our sample are published in peer-reviewed journals (89%).

Table 4: Description and summary statistics of variables reflecting context

Variable	Description	Mean	SD	WM
Beauty effect	Reported estimate recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty.	4.28	6.28	5.21
Standard error (SE)	Standard error of the estimate (the variable is important for gauging publication bias).	3.75	4.27	4.05
Measurement of beauty				
Interviewer-rated beauty	=1 if a rater assesses the beauty of a subject in person.	0.45	0.50	0.34
Photo-rated beauty	=1 if a rater assesses the beauty of a subject based on a photo.	0.42	0.49	0.53
Software-rated beauty	=1 if a software tool (e.g., a symmetry assessment algorithm) assesses the beauty of a subject.	0.09	0.29	0.10
Self-rated beauty	=1 if subjects self-rate their beauty (reference category).	0.05	0.21	0.06
Dummy beauty	=1 if the beauty variable is a dummy (such as "attractive") and compared to a baseline (mean).	0.40	0.49	0.30

Continued on next page

Table 4: Description and summary statistics of variables reflecting context (continued)

Variable	Description	Mean	SD	WM
Categorical beauty	=1 if the beauty variable included in the regression is defined on a scale, e.g. from 1 to 10 (reference category).	0.60	0.49	0.70
Beauty penalty	=1 if the original estimate concerns the effect of below-average looks (e.g. by focusing on a dummy variable "unattractive").	0.18	0.38	0.10
Number of raters	Logarithm of the average number of raters per subject. When software rating is used, the variable is set to sample maximum.	1.85	1.52	2.09
Measurement of success				
Earnings	=1 if the success measure concerns earnings.	0.59	0.49	0.61
Study outcomes	=1 if the success measure concerns performance at school.	0.16	0.37	0.12
Teaching & research outcomes	=1 if the success measure concerns academic performance (such as citations).	0.12	0.33	0.12
Athletic success	=1 if the success measure concerns athletic performance (such as points or TV viewership).	0.03	0.17	0.02
Electoral success	=1 if the success measure concerns electoral success (such as votes).	0.03	0.18	0.06
Other outcomes	=1 if the success measure concerns other issues related to performance, such as sales or analysts' forecast error (reference category).	0.06	0.24	0.07
Data characteristics				
Male subjects	=1 if the subjects are men.	0.32	0.47	0.29
Female subjects	=1 if the subjects are women.	0.39	0.49	0.36
Mix-gender subjects	=1 if the sample includes both men and women (reference category).	0.29	0.45	0.35
Subjects' age	Logarithm of the average age of the subject.	3.40	0.45	3.51
High-skilled workers	=1 if the study focuses on college-educated workers.	0.29	0.45	0.36
Prostitutes	=1 if the study focuses on sex workers.	0.05	0.21	0.06
Other dressy occupations	=1 if the study focuses on occupations where appearance matters: executives, politicians, lawyers, consultants, salesmen (excluding sex workers).	0.12	0.32	0.19
Non-dressy occupations	=1 if the study focuses on occupations where appearance should not matter much for productivity, such as teachers, athletes, analysts, and general population (reference category).	0.83	0.37	0.75
Western culture	=1 if the study focuses on people (both raters and subjects) from the West.	0.75	0.43	0.69
Panel data	=1 if panel data or pooled cross-sections are used in the study.	0.86	0.35	0.81
Cross-section	=1 if purely cross-sectional data are used in the study (reference category).	0.14	0.35	0.19
Data year	Logarithm of the average year of the data used to estimate the beauty normalized by the year of the oldest data in our sample.	3.34	0.61	3.45
Estimation technique				
OLS method	=1 if the ordinary least squares method is used for estimation.	0.78	0.42	0.79
IV method	=1 if instrumental variable methods are used for estimation.	0.07	0.26	0.07
DID method	=1 if the difference-in-differences method is used for estimation.	0.01	0.10	0.01

Continued on next page

Table 4: Description and summary statistics of variables reflecting context (continued)

Variable	Description	Mean	SD	WM
Other method	=1 if other methods (maximum likelihood, quantile regression, ridge regression, tobit, propensity score matching) are used for estimation (reference category for estimation methods).	0.14	0.35	0.13
Age control	=1 if the study controls for subjects' age or experience.	0.88	0.33	0.81
Education control	=1 if the study controls for subjects' education.	0.66	0.47	0.58
Ethnicity control	=1 if the study controls for subjects' ethnicity or race.	0.55	0.50	0.42
Cognitive ability control	=1 if the study controls for subjects' cognitive skills (e.g. IQ).	0.38	0.49	0.31
Non-cognitive ability control	=1 if the study controls for subjects' non-cognitive skills such as measures of communication skills, confidence, leadership skills, or another indicator (such as "Big Five" personality traits).	0.21	0.41	0.22
Physicality control	=1 if the study controls for subjects' physicality using weight, height, or body mass index.	0.27	0.44	0.25
Publication characteristics				
Publication year	The logarithm of the year when the study first appeared in Google Scholar normalized by the year of the earliest publication in our sample.	2.83	0.74	2.90
Published study	=1 if the study was published in a peer-reviewed journal.	0.89	0.32	0.87
Impact factor	Journal Citation Reports impact factor (Clarivate, 2023).	2.88	2.82	2.65
Citations	Logarithm of the number of per-year citations received since the study first appeared in Google Scholar.	1.42	1.06	1.41

Notes: SD = standard deviation, WM = mean weighted by the inverse of the number of estimates reported per study.

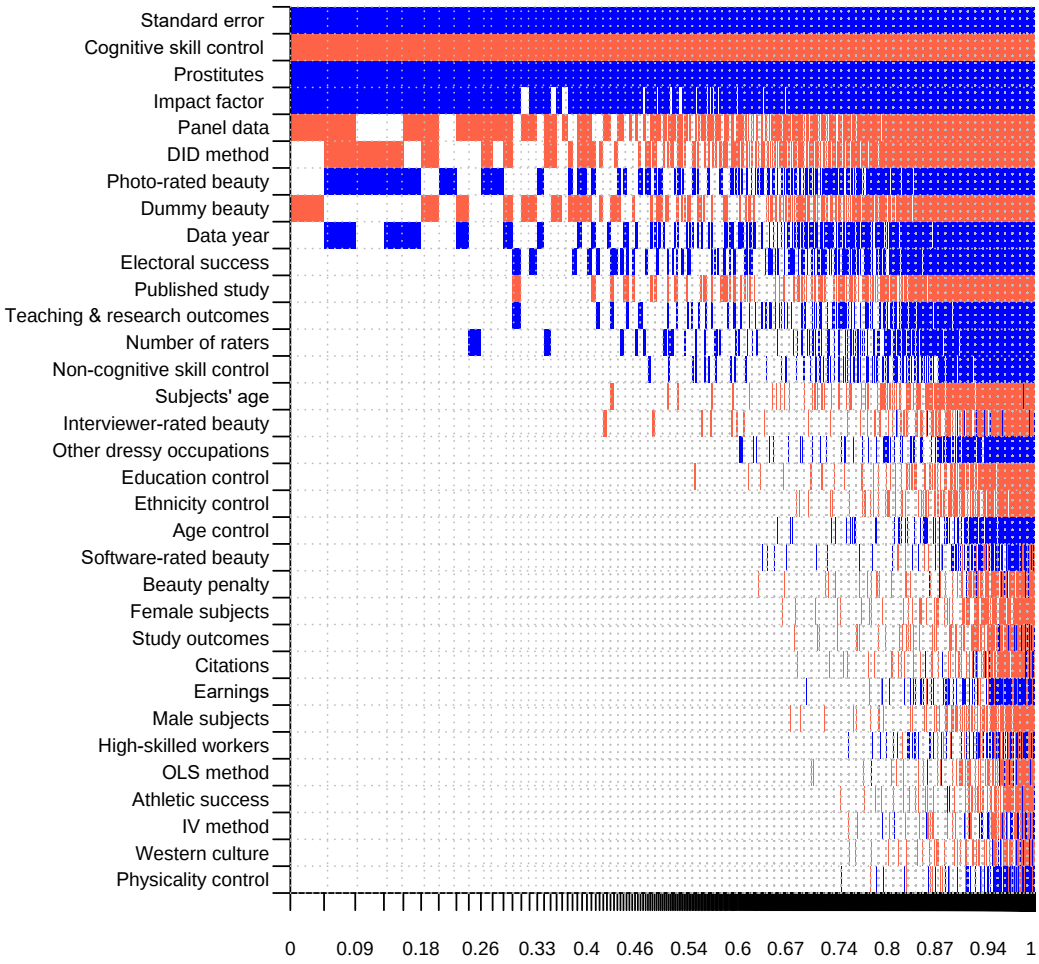
After excluding baseline categories, we are left with 33 explanatory variables. All of them can potentially affect the reported estimates of beauty effects, but probably only few will prove systematically important in practice. We thus face substantial model uncertainty: including all the variables into one regression would result in exceedingly imprecise estimates even for the most important variables. As Steel (2020) notes, the natural response to model uncertainty is Bayesian model averaging (BMA). BMA exploits the Markov chain Monte Carlo algorithm (see, e.g., Feldkircher & Zeugner, 2009) which allows us to avoid estimating all the 2^{33} potential models and to concentrate on the most important portion of the model mass. For our baseline estimation we choose the unit information prior recommended by Eicher *et al.* (2011), which gives the prior that each coefficient is zero the same weight as one data point. Additionally we use the dilution model prior developed by George (2010), which discounts models with substantial collinearity. In effect, BMA weights individual models by measures related to model

fit and parsimony. For each variable the sum of the weights of the models in which the variable is included is denoted by posterior inclusion probability (PIP). Variables with a high PIP are effective in explaining the differences in the beauty effects reported in the literature.

Figure 8 presents a graphical summary of Bayesian model averaging results. On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. In other words, the variables shown at the top are the ones most useful in explaining differences in the reported beauty effects. The horizontal axis shows the values of the cumulative posterior model probability, the model weight used in BMA. The models on the left display the best combination of data fit and parsimony. Blue color (darker in grayscale) means that the estimated parameter of the corresponding explanatory variable is positive. Red color (lighter in grayscale) means the estimated parameter is negative. No color means the corresponding explanatory variable is excluded. The figure shows that only 4 variables out of the 33 that we consider are robustly associated with the reported beauty effects: standard error (proxy for publication bias), cognitive skill control, a dummy variable for prostitutes, and the impact factor of the journal in which the study was published.

Table 5 shows the numerical results of Bayesian model averaging and a simple stepwise regression provided as a frequentist robustness check. The posterior mean in BMA denotes the partial derivative of the reported beauty premium with respect to the corresponding study characteristic. For example, including a control for cognitive skills typically reduces the beauty premium by 2.3 percentage points compared to the case in which cognitive skills are ignored in primary studies. The corresponding variable also has a high posterior inclusion probability, almost 100%. The same is true for the standard error: even when we explicitly control for heterogeneity we obtain strong evidence for publication bias. We also find that, unsurprisingly, sex workers enjoy substantially larger beauty premiums (by about 5 percentage points) than other occupations. The BMA results also suggest that studies published in better journals (as measured by the impact factor) tend to publish larger beauty premiums. Nevertheless, the latter finding does not survive the frequentist check reported in the right-hand part of Table 5. In contrast, the stepwise regression shows statistical significance for the variable related to difference-in-differences, for which BMA finds a large coefficient estimate (-2.5) but an inclusion probability slightly below 50%.

Figure 8: Model inclusion in Bayesian model averaging



Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. In other words, the variables shown at the top are the ones most useful in explaining differences in the reported beauty effects. The horizontal axis shows the values of cumulative posterior model probability. The models on the left display the best combination of data fit and parsimony. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table 5. All variables are described in Table 4. Technical details and diagnostics of the BMA exercise are available in Table B5 and Figure B1.

Table 5: Why reported beauty premiums vary

Response variable: Beauty premium	Bayesian model averaging (baseline model)			Stepwise regression (frequentist check)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Constant	2.759	NA	1.000	3.582	0.562	0.000
Standard error	0.435	0.042	1.000	0.426	0.112	0.000
Measurement of beauty						
Interviewer-rated beauty	-0.033	0.204	0.036			
Photo-rated beauty	0.591	0.745	0.424			
Software-rated beauty	0.009	0.149	0.014			
Dummy beauty	-0.485	0.664	0.387			
Beauty penalty	-0.007	0.089	0.012			
Number of raters	0.051	0.140	0.134			
Measurement of success						
Earnings	0.006	0.098	0.010			
Study outcomes	-0.006	0.114	0.011			
Teaching & research	0.238	0.645	0.141			
Athletic success	-0.004	0.104	0.007			
Electoral success	0.670	1.361	0.224			
Data characteristics						
Male subjects	-0.005	0.064	0.010			
Female subjects	-0.006	0.069	0.012			
Subjects' age	-0.075	0.332	0.061			
High-skilled workers	0.002	0.064	0.009			
Prostitutes	4.793	1.058	0.999	4.386	1.199	0.001
Other dressy occupations	0.035	0.227	0.032			
Western culture	-0.001	0.039	0.006			
Panel data	-1.131	1.016	0.606			
Data year	0.337	0.504	0.346			
Estimation technique						
OLS method	-0.002	0.050	0.008			
IV method	-0.001	0.065	0.007			
DID method	-2.484	2.895	0.470	-5.078	2.061	0.016
Age control	0.015	0.135	0.018			
Education control	-0.017	0.125	0.026			
Ethnicity control	-0.012	0.105	0.019			
Cognitive skill control	-2.285	0.447	1.000	-2.750	0.691	0.000
Non-cognitive skill control	0.071	0.291	0.070			
Physicality control	0.000	0.036	0.006			
Publication characteristics						
Published study	-0.313	0.747	0.173			
Impact factor	0.265	0.123	0.908			
Citations	-0.002	0.035	0.011			
Studies	67			67		
Observations	1,159			1,159		

Notes: The posterior mean in BMA denotes the partial derivative of the reported beauty premium with respect to the corresponding study characteristic. For example, including a control for cognitive skills typically reduces the beauty premium by 2.3 percentage points. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. BMA employs the unit information prior recommended by (Eicher *et al.*, 2011) and the dilution prior suggested by George (2010), which accounts for collinearity. The frequentist check (stepwise regression) is estimated using the 5% significance threshold and standard errors clustered at the study level. All variables are described in Table 4. Technical details and diagnostics of the BMA exercise are available in Table B5 and Figure B1.

The analysis of heterogeneity yields three key takeaways. First, the finding of publication bias is robust to explicit control for observables. Second, most aspects related to estimation context are not systematically associated with the reported beauty effects. We find no evidence that variables potentially associated with the extent of attenuation bias (algorithmic beauty rating, number of raters, IV estimation) affect the results much. Similarly, it does not seem to matter whether researchers consider the effect of beauty on earnings or on proxies for productivity. Third, much of the beauty effect beyond publication bias is due to a correlation between beauty and cognitive ability. The third point is best illustrated by computing the mean beauty premium conditional on correcting for publication bias and controlling for cognitive ability either via an explicit inclusion of the control or via difference-in-differences. When we plug in the corresponding variables to the results of the frequentist check in Table 5, we obtain an implied beauty premium of -0.1 with the 95% confidence interval $(-2.2, 2.0)$. The point estimate from BMA is similar but accompanied by an uninformatively wide credible interval due to the large number of other variables that must be set to their sample means.

In Appendix B we report robustness checks that employ different priors for BMA (Table B6) and that use a subsample of reported estimates without beauty penalties (Table B8). We also present another version of the frequentist check: instead of the stepwise regression, in Table B6 we run OLS that only includes variables with a posterior inclusion probability above 0.5. Our main results are not affected by these changes. The only plausible scenario that could produce a non-negligible beauty premium beyond publication bias and after controlling for cognitive ability is one in which we put great weight on results published in journals with a high impact factor—perhaps as a proxy for unobserved aspects of study quality. The corresponding regression estimates for the impact factor variable, while statistically insignificant in all the frequentist models we run, suggest an increase in the premium of about one percentage point associated with an increase in the impact factor of 4.

5 Conclusion

We collect 1,159 estimates from 67 studies examining the effect of beauty on earnings or proxies for productivity. The mean reported effect, recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty, equals 4.3%.

After correction for publication bias, the mean beauty premium reduces to 2.9%. The premium falls to near zero when studies control for cognitive ability. Moreover, even before any adjustment for publication bias and omitted variables, the effect of beauty is similar for earnings and productivity. Hence we find no evidence for discrimination based on tastes for beauty. Beauty seems to be correlated with productive characteristics and does not affect earnings causally.

Three qualifications of our results are in order. First, random errors in beauty measurement can bias the reported estimates downwards. While we cannot fully correct for what almost certainly is an attenuation bias in the literature, we use three strategies to gauge the extent of the problem: comparison of OLS and IV estimates, comparison of human and software rating, and comparison based on the number of raters. More raters or software rating should plausibly diminish measurement error. Because the IV estimates in the literature are unlikely to help with omitted variables or reverse causality, under the assumption of a classical measurement error the difference between OLS and IV serves as a proxy for attenuation bias (see Havranek *et al.* 2024 for more details on this approach in the context of the elasticity of substitution between skilled and unskilled labor). We find no evidence that IV estimates differ systematically from OLS estimates, and it does not seem to matter how many raters are employed or whether software rating is used. We thus fail to identify any substantial attenuation bias.

Second, our dataset includes both beauty premiums (comparisons of average and above-average looks) and beauty penalties (comparisons of average and below-average looks). We recompute all estimates to an effect corresponding to a one-standard-deviation increase in beauty and pool them together in our baseline analysis. The pooling assumes linearity in the effects of beauty, which is a strong assumption. So, as a robustness check, we also focus separately on beauty premiums and beauty penalties. We fail to find a systematic difference between the two. The indifference result is unexpected because previous research suggests that beauty penalties are likely to be larger than beauty premiums (Hamermesh, 2011).

Third, for most primary studies we do not have crisp data on job tasks that would allow us to cleanly separate occupations where beauty is likely to be a genuinely productive factor. Stinebrickner *et al.* (2019) have such data and find no beauty effects for jobs where employees do not come into personal contact with customers. In a meta-analysis setting we can separate beauty premium estimates for occupations where looks are likely to be especially important

(lawyers, politicians, salespeople, etc.) from those where looks are unlikely to matter much (analysts, researchers, craftsmen, etc.). We put sex workers aside as a special category where beauty is a key productive characteristic. Our results show that sex workers enjoy beauty premiums clearly much larger than other occupations, but we fail to find systematic differences among the remaining categories. For politicians the raw reported correlation between beauty and success is almost as high as for sex workers, but the effect mostly disappears after correcting for publication bias and controlling for cognitive ability.

References

Ahmed, S., M. Ranta, E. Vahamaa, & S. Vahamaa (2023): "Facial attractiveness and CEO compensation: Evidence from the banking industry." *Journal of Economics and Business* **123(C)**: p. 106095.

Ahn, S. C. & Y. H. Lee (2014): "Beauty And Productivity: The Case Of The Ladies Professional Golf Association." *Contemporary Economic Policy* **32(1)**: pp. 155–168.

Andrews, I. & M. Kasy (2019): "Identification of and correction for publication bias." *American Economic Review* **109(8)**: pp. 2766–2794.

Andrews, I., J. H. Stock, & L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice." *Annual Review of Economics* **11(1)**: pp. 727–753.

Anyzova, P. & P. Mateju (2018): "Beauty still matters: The role of attractiveness in labour market outcomes." *International Sociology* **33(3)**: pp. 269–291.

Arunachalam, R. & M. Shah (2012): "The Prostitute's Allure: The Return to Beauty in Commercial Sex Work." *The B.E. Journal of Economic Analysis & Policy* **12(1)**: pp. 1–27.

Bakkenbull, L.-B. (2017): "The Impact of Attractiveness on Athletic Performance of Tennis Players." *International Journal of Social Science Studies* **5(3)**: pp. 12–20.

Bakkenbull, L.-B. & S. Kiefer (2015): "Are Attractive Female Tennis Players More Successful? An Empirical Analysis." *Kyklos* **68(4)**: pp. 443–458.

Bartos, F., M. Maier, E.-J. Wagenmakers, F. Nippold, H. Doucouliagos, J. P. A. Ioannidis, W. M. Otte, M. Sladekova, T. K. Deresssa, S. B. Bruns, D. Fanelli, & T. Stanley (2024): "Footprint of publication selection bias on meta-analyses in medicine, environmental sciences, psychology, and economics." *Research Synthesis Methods* **(forthcoming)**.

Berggren, N., H. Jordahl, & P. Poutvaara (2010): "The looks of a winner: Beauty and electoral success." *Journal of Public Economics* **94(1-2)**: pp. 8–15.

Berri, D. J., R. Simmons, J. Van Gilder, & L. O'Neill (2011): "What does it mean to find the face of the franchise? Physical attractiveness and the evaluation of athletic performance." *Economics Letters* **111(3)**: pp. 200–202.

Bi, W., H. Chan, & B. Torgler (2020): ""Beauty" premium for social scientists but "unattractiveness" premium for natural scientists in the public speaking market." *Humanities and Social Sciences Communications* **7(1)**: pp. 1–9.

Biddle, J. E. & D. S. Hamermesh (1998): "Beauty, Productivity, and Discrimination: Lawyers' Looks and Lucre." *Journal of Labor Economics* **16(1)**: pp. 172–201.

Blanco-Perez, C. & A. Brodeur (2020): "Publication Bias and Editorial Statement on Negative Findings." *Economic Journal* **130(629)**: pp. 1226–1247.

Bom, P. R. D. & H. Rachinger (2019): "A kinked meta-regression model for publication bias correction." *Research Synthesis Methods* **10(4)**: pp. 497–514.

Borland, J. & A. Leigh (2014): "Unpacking the Beauty Premium: What Channels Does It Operate Through, and Has It Changed Over Time?" *The Economic Record* **90(288)**: pp. 17–32.

Brodeur, A., S. Carrell, D. Figlio, & L. Lusher (2023): "Unpacking p-hacking and publication bias." *American Economic Review* **113(11)**: pp. 2974–3002.

Brown, A. L., T. Imai, F. Vieider, & C. Camerer (2024): "Meta-Analysis of Empirical Estimates of Loss-Aversion." *Journal of Economic Literature* **(forthcoming)**.

Card, D., J. Kluve, & A. Weber (2018): "What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations." *Journal of the Euro-*

pean Economic Association **16(3)**: pp. 894–931.

Card, D. & A. B. Krueger (1995): "Time-series minimum-wage studies: A meta-analysis." *American Economic Review* **85(2)**: pp. 238–243.

Cipriani, G. P. & A. Zago (2011): "Productivity or Discrimination? Beauty and the Exams." *Oxford Bulletin of Economics and Statistics* **73(3)**: pp. 428–447.

Clark, C. & D. Walker (2009): "Does adolescent attractiveness matter? Academic performance, college attendance, and criminal & delinquent behavior." *Southern Business and Economic Journal* **32(3–4)**: pp. 57–78.

Cook, D. O. & S. Mobbs (2023): "CEO Selection and Executive Appearance." *Journal of Financial and Quantitative Analysis* **58(4)**: pp. 1582–1611.

DellaVigna, S. & E. Linos (2022): "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica* **90(1)**: pp. 81–116.

Deryugina, T. & O. Shurchkov (2015): "Does Beauty Matter In Undergraduate Education?" *Economic Inquiry* **53(2)**: pp. 940–961.

Dietl, H., A. Ozdemir, & A. Rendall (2020): "The role of facial attractiveness in tennis TV-viewership." *Sport Management Review* **23(3)**: pp. 521–535.

Dilmaghani, M. (2020): "Beauty perks: Physical appearance, earnings, and fringe benefits." *Economics and Human Biology* **38(C)**: p. 100889.

Dossinger, K., C. Wanberg, Y. Choi, & L. Leslie (2019): "The beauty premium: The role of organizational sponsorship in the relationship between physical attractiveness and early career salaries." *Journal of Vocational Behavior* **112**: pp. 109–121.

Duval, S. & R. Tweedie (2000): "Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis." *Biometrics* **56(2)**: pp. 455–463.

Edlund, L., J. Engelberg, & C. Parsons (2009): "The wages of sin." *Discussion paper 0809-16*, Department of Economics at Columbia University, New York.

Egger, M., G. D. Smith, M. Schneider, & C. Minder (1997): "Bias in meta-analysis detected by a simple, graphical test." *BMJ* **315(7109)**: pp. 629–634.

Eicher, T. S., C. Papageorgiou, & A. E. Raftery (2011): "Default priors and predictive performance in Bayesian model averaging, with application to growth determinants." *Journal of Applied Econometrics* **26(1)**: pp. 30–55.

Elliott, G., N. Kudrin, & K. Wuthrich (2022): "Detecting p-hacking." *Econometrica* **90(2)**: pp. 887–906.

Feldkircher, M. & S. Zeugner (2009): "Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian Model Averaging." *IMF Working Papers 09/202/2009*, International Monetary Fund.

Fernandez, C., E. Ley, & M. F. J. Steel (2001): "Benchmark priors for Bayesian Model Averaging." *Journal of Econometrics* **100(2)**: pp. 381–427.

Fidrmuc, J. & B. Paphawasit (2018): "Beautiful minds: Physical attractiveness and research productivity in economics." *Technical report*, Presented at the 2nd IZA/HSE Workshop, Moscow, July 2018.

Fletcher, J. M. (2009): "Beauty vs. brains: Early labor market outcomes of high school graduates." *Economics Letters* **105(3)**: pp. 321–325.

French, M. T., P. K. Robins, J. F. Homer, & L. M. Tapsell (2009): "Effects of physical attractiveness, personality, and grooming on academic performance in high school." *Labour Economics* **16(4)**: pp. 373–382.

Furukawa, C. (2020): "Publication bias under aggregation frictions: Theory, evidence, and a new correction method." *Working paper*, MIT.

George, E. I. (2010): "Dilution priors: Compensating for model space redundancy." In "IMS Collections Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown," volume 6, p. 158–165. Institute of Mathematical Statistics.

Gerber, A. & N. Malhotra (2008): "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* **3(3)**: pp. 313–326.

Gertler, P., M. Shah, & S. M. Bertozzi (2005): "Risky Business: The Market for Unprotected Commercial Sex." *Journal of Political Economy* **113(3)**: pp. 518–550.

Gu, T. & Y. Ji (2019): "Beauty premium in China's labor market: Is discrimination the main reason?" *China Economic Review* **57(C)**: p. 101335.

Halford, J. & H.-C. Hsu (2020): "CEO attractiveness and firm value." *The Financial Review* **55(4)**: pp. 529–556.

Hamermesh, D. & J. Abrevaya (2013): "Beauty is the promise of happiness?" *European Economic Review* **64(C)**: pp. 351–368.

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

Hamermesh, D. S. & J. E. Biddle (1994): "Beauty and the Labor Market." *American Economic Review* **84(5)**: pp. 1174–1194.

Hamermesh, D. S. & A. K. Leigh (2022): ""Beauty too rich for use": Billionaires' assets and attractiveness." *Labour Economics* **76**: p. 102153.

Hamermesh, D. S., X. Meng, & J. Zhang (2002): "Dress for success–does primping pay?" *Labour Economics* **9(3)**: pp. 361–373.

Hamermesh, D. S. & A. Parker (2005): "Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity." *Economics of Education Review* **24(4)**: pp. 369–376.

Hamermesh, D.S.and Gordon, R. & R. Crosnoe (2023): "O Youth and Beauty:" Children's Looks and Children's Cognitive Development." *Journal of Economic Behavior & Organization* **212(C)**: pp. 275–289.

Harper, B. (2000): "Beauty, Stature and the Labour Market: A British Cohort Study." *Oxford Bulletin of Economics and Statistics* **62(s1)**: pp. 771–800.

Havranek, T., Z. Irsova, L. Laslopova, & O. Zeynalova (2024): "Publication and Attenuation Biases in Measuring Skill Substitution." *The Review of Economics and Statistics* **(forthcoming)**.

Havranek, T., T. D. Stanley, H. Doucouliagos, P. Bom, J. Geyer-Klingeberg, I. Iwasaki, W. R. Reed, K. Rost, & R. C. M. van Aert (2020): "Reporting Guidelines for Meta-Analysis in Economics." *Journal of Economic Surveys* **34(3)**: pp. 469–475.

Hernandez-Julian, R. & C. Peters (2017): "Student Appearance and Academic Performance." *Journal of Human Capital* **11(2)**: pp. 247–262.

Hitsch, G., A. Hortacsu, & D. Ariely (2010): "What makes you click? Mate preferences in online dating." *Quantitative Marketing and Economics* **8(C)**: pp. 393–427.

Imai, T., T. A. Rutter, & C. F. Camerer (2020): "Meta-Analysis of Present-Bias Estimation Using Convex Time Budgets." *Economic Journal* **(forthcoming)**.

Ioannidis, J. P., T. D. Stanley, & H. Doucouliagos (2017): "The Power of Bias in Economics Research." *Economic Journal* **127(605)**: pp. F236–F265.

Irsova, Z., P. R. D. Bom, T. Havranek, & H. Rachinger (2023): "Spurious Precision in Meta-Analysis." *MetaArXiv Preprint 3qp2w*.

Islam, A. & R. Smyth (2012): "The Economic Returns to Good Looks and Risky Sex in the Bangladesh Commercial Sex Market." *The B.E. Journal of Economic Analysis & Policy* **12(1)**: pp. 1–25.

Jobu Babin, J., A. Hussey, A. Nikolsko-Rzhevskyy, & D. A. Taylor (2020): "Beauty Premiums Among Academics." *Economics of Education Review* **78(C)**: p. 102019.

Kanazawa, S. (2011): "Intelligence and physical attractiveness." *Intelligence* **39(1)**: pp. 7–14.

Kanazawa, S. & J. L. Kovar (2004): "Why beautiful people are more intelligent." *Intelligence* **32**: pp. 227–243.

Kanazawa, S. & M. Still (2018): "Is There Really a Beauty Premium or an Ugliness Penalty on Earnings?" *Journal of Business and Psychology* **33**: pp. 249–262.

Keane, M. & T. Neal (2023): "Instrument strength in IV estimation and inference: A guide to theory and practice." *Journal of Econometrics* **235(2)**: pp. 1625–1653.

King, A. & A. Leigh (2009): "Beautiful Politicians." *Kyklos* **62(4)**: pp. 579–593.

Klein, M. & U. Rosar (2005): "Physische Attraktivitat und Wahlerfolg: Eine Empirische Analyse am Beispiel der Wahlkreiskandidaten bei der Bundestagswahl 2002." *Politische Vierteljahresheft* **46(C)**: pp. 266–290.

Kraft, P. (2012a): "The role of beauty in the labor market: The signaling effect of beauty." *Dissertation chapter 2*, Center for Economic Research and Graduate Education, Prague.

Kraft, P. (2012b): "The role of beauty in the labor market: Attractive compensation, or compensation for being attractive? Evidence from German CEOs." *Dissertation chapter 3*, Center for Economic Research and Graduate Education, Prague.

Kranz, S. & P. Putz (2022): "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics: Comment." *American Economic Review* **112(9)**: pp. 3124–3136.

Lahdevuori, S. (2013): *CEO appearance, compensation, and firm performance - Evidence from Sweden.* Master thesis, School of Business, Aalto University, Aalto, Finland.

Lee, S. & K. Ryu (2012): "Plastic Surgery: Investment in Human Capital or Consumption?" *Journal of Human Capital* **6(3)**: pp. 224–250.

Leigh, A. & T. Susilo (2009): "Is voting skin-deep? Estimating the effect of candidate ballot photographs on election outcomes." *Journal of Economic Psychology* **30(C)**: pp. 61–70.

Ley, E. & M. F. Steel (2009): "On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression." *Applied Econometrics* **24**: pp. 651–674.

Li, C., A.-P. Lin, H. Lu, & K. J. Veenstra (2020): "Gender and beauty in the financial analyst profession: Evidence from the United States and China." *Review of Accounting Studies* **25**: pp. 1230–1262.

Li, M., M. Triana, S. Byun, & O. Chapa (2021): "Pay for beauty? A contingent perspective of CEO facial attractiveness on CEO compensation." *Human Resource Management* **60(6)**: pp. 843–862.

Liu, X. (2015): "Three Essays on Labor Economics. Physical Attractiveness and Earnings: Evidence from a Longitudinal Survey." *Dissertation chapter 1*, Department of Economics, The University of Arizona, United States, pp. 16–51.

Liu, Y., H. Lu, & K. Veenstra (2024): "Beauty and Accounting Academic Career." *Journal of Accounting, Auditing & Finance* **(forthcoming)**.

Malik, N., P. Singh, & K. Srinivasan (2024): "When Does Beauty Pay? A Large-Scale Image Based Appearance Analysis on Career Transitions." *Information Systems Research* **(forthcoming)**.

Mathur, M. B. (2024): "P-hacking in meta-analyses: A formalization and new meta-analytic methods." *Research Synthesis Methods* **(forthcoming)**.

Mehic, A. (2022): "Student beauty and grades under in-person and remote teaching." *Economics Letters* **219(C)**: p. 110782.

Mitchem, D. G., B. P. Zietsch, M. J. Wright, N. G. Martin, J. K. Hewitt, & M. C. Keller (2015): "No relationship between intelligence and facial attractiveness in a large, genetically informative sample." *Evolution and Human Behavior* **36(3)**: pp. 240–247.

Mobius, M. M. & T. S. Rosenblat (2006): "Why Beauty Matters." *American Economic Review* **96(1)**: pp. 222–235.

Mocan, N. & E. Tekin (2010): "Ugly Criminals." *The Review of Economics and Statistics* **92(1)**: pp. 15–30.

Monk, E., M. Esposito, & H. Lee (2021): "Beholding Inequality: Race, Gender, and Returns to Physical Attractiveness in the United States." *American Journal of Sociology* **127(1)**: pp. 194–241.

Nault, K. A., M. Pitesa, & S. Thau (2020): "The Attractiveness Advantage At Work: A Cross-Disciplinary Integrative Review." *Academy of Management Annals* **14(2)**: pp. 1103–1139.

Neisser, C. (2021): "The Elasticity of Taxable Income: A Meta-Regression Analysis." *Economic Journal* **131(640)**: pp. 3365–3391.

Oreffice, S. & C. Quintana-Domeque (2016): "Beauty, body size and wages: Evidence from a unique data set." *Economics & Human Biology* **22(C)**: pp. 24–34.

Parrett, M. (2015): "Beauty and the feast: Examining the effect of beauty on earnings using restaurant tipping data." *Journal of Economic Psychology* **49(C)**: pp. 34–46.

Peng, L., X. Wang, & S. Ying (2020): "The heterogeneity of beauty premium in China: Evidence from CFPS." *Economic Modelling* **90(C)**: pp. 386–396.

Pfann, G. A., J. E. Biddle, D. S. Hamermesh, & C. M. Bosman (2000): "Business success and businesses' beauty capital." *Economics Letters* **67(2)**: pp. 201–207.

Pfeifer, C. (2012): "Physical attractiveness, employment and earnings." *Applied Economics Letters* **19(6)**: pp. 505–510.

Ponzo, M. & V. Scoppa (2013): "Professors' Beauty, Ability, and Teaching Evaluations in Italy." *The B.E. Journal of Economic Analysis & Policy* **13(2)**: pp. 811–835.

Ravina, E. (2019): "Love & loans: The effect of beauty and personal characteristics in credit markets." *working paper*, (previous versions of the working paper from 2008, New York University), Northwestern University, Unted States.

Roodman, D., J. G. MacKinnon, M. O. Nielsen, & M. D. Webb (2018): "Fast and wild: Bootstrap inference in Stata using boottest." *Queen's Economics Department Working Paper 1406*, Department of Economics, Queen's University, Canada: Kingston.

Ross, J. & K. R. Ferris (1981): "Interpersonal Attraction and Organizational Outcomes: A Field Examination." *Administrative Science Quarterly* **26(4)**: pp. 617–632.

Sachsida, A., A. C. Dornelles, & C. W. Mesquita (2003): "Beauty and the Labor Market – Study one Specific Occupation." *Technical report*, Mestrado em Economia de Empresas, Catholic University of Brasilia.

Salter, S. P., F. G. Mixon, & E. W. King (2012): "Broker beauty and boon: a study of physical attractiveness and its effect on real estate brokers' income and productivity." *Applied Financial Economics* **22(10)**: pp. 811–825.

Schnusenberg, O. & C. Froehlich (2011): "Hot and easy in Florida: The case of economics professors." *Research in Higher Education Journal* **10(C)**: p. 10628.

Scholz, J. K. & K. Sicinski (2015): "Facial Attractiveness and Lifetime Earnings: Evidence from a Cohort Study." *The Review of Economics and Statistics* **97(1)**: pp. 14–28.

Sen, A., M. Voia, & F. Woolley (2010): "Hot or Not: How appearance affects earnings and productivity in academia." *Carleton Economic Papers 10-07*, Carleton University. Current version under review in Journal of Human Capital under: Does hotness pay? The relationship between appearance, productivity and earnings for Ontario economics professors.

Stanley, T. & H. Doucouliagos (2010): "Picture This: A Simple Graph That Reveals Much Ado About Research." *Journal of Economic Surveys* **24(1)**: pp. 170–191.

Stanley, T. D. (2008): "Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection." *Oxford Bulletin of Economics and Statistics* **70(1)**: pp. 103–127.

Stanley, T. D., H. Doucouliagos, J. P. A. Ioannidis, & E. C. Carter (2021): "Detecting publication selection bias through excess statistical significance." *Research Synthesis Methods* **12(6)**: pp. 776–795.

Steel, M. F. J. (2020): "Model Averaging and its Use in Economics." *Journal of Economic Literature* **58(3)**: pp. 644–719.

Stinebrickner, R., T. Stinebrickner, & P. Sullivan (2019): "Beauty, Job Tasks, and Wages: A New

Conclusion about Employer Taste-Based Discrimination." *The Review of Economics and Statistics* **101(4)**: pp. 602–615.

Tao, H.-L. (2008): "Attractive Physical Appearance vs. Good Academic Characteristics: Which Generates More Earnings?" *Kyklos* **61(1)**: pp. 114–133.

Vivalt, E. (2019): "Specification Searching and Significance Inflation Across Time, Methods and Disciplines." *Oxford Bulletin of Economics and Statistics* **81(4)**: pp. 797–816.

Walcutt, B., L. Patterson, & S. Seo (2011): "Beauty Premium and Grade Point Average: A Study of Business Students at a Korean University." *Business Studies Journal* **3(1)**: pp. 51–68.

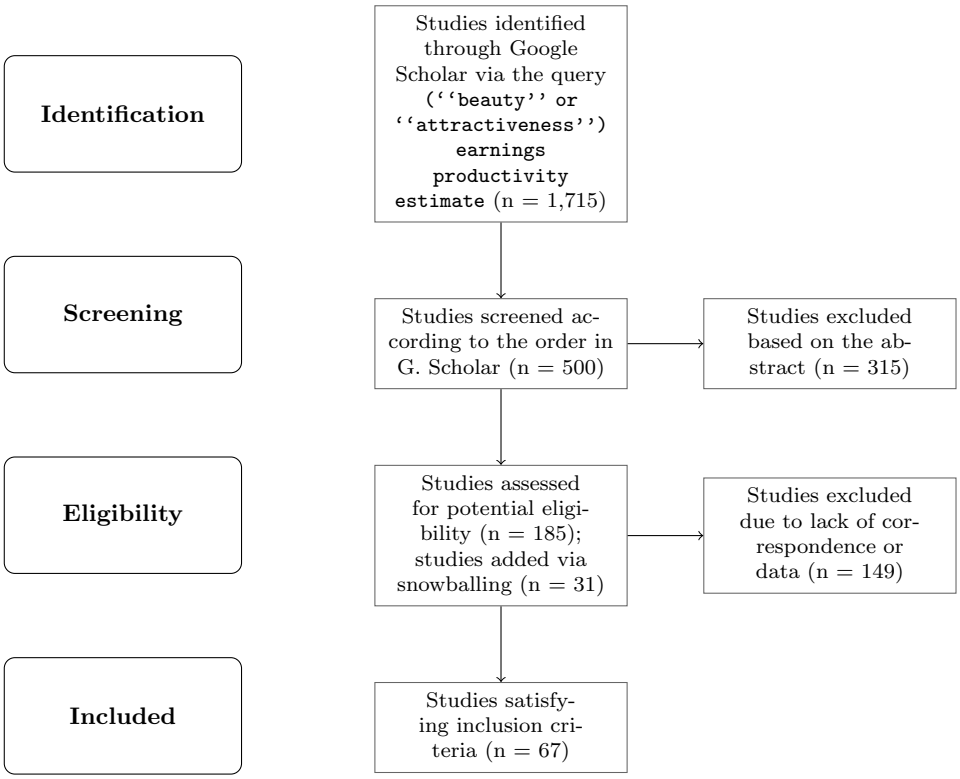
Wolbring, T. & P. Riordan (2016): "How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching." *Social Social Research* **57**: pp. 253–272.

Xue, X., W. R. Reed, & A. Menclova (2020): "Social capital and health: a meta-analysis." *Journal of Health Economics* **72(C)**: p. 102317.

Appendices

A Details on the Literature Search

Figure A1: PRISMA flow diagram



Notes: Preferred reporting items for systematic reviews and meta-analyses (PRISMA) is an evidence-based set of items for reporting in systematic reviews and meta-analyses. More details on PRISMA and reporting standards in the context of economics meta-analyses are provided by Havranek *et al.* (2020). Snowballing: we download the references of the potentially eligible studies identified in step "Screening" and inspect the 100 studies most commonly cited among the 185 studies. If, based on the abstract, these commonly cited studies show any promise of containing empirical estimates of the beauty effect, we add them to the set of potentially eligible studies. Snowballing yields 31 additional studies. Criteria for inclusion: i) the study must report the effect of beauty on a continuous variable reflecting earnings or productivity in a field (real-world) setting; ii) the beauty measurement used in the study must focus on physiognomy (just the face), iii) the study must focus on the subject's earnings or productivity, not e.g. the income of the spouse, iv) the study must report statistics that allow us to convert the reported estimate to the percent increase in earnings or productivity following a one-standard-deviation increase in beauty; v) the study must report standard errors or other statistics from which standard errors can be computed. The literature search was terminated on February 16, 2024. The dataset, together with R and Stata codes, is available at meta-analysis.cz/beauty.

B Additional Results (for Online Publication)

Table B1: Specification test for the Andrews & Kasy (2019) model

	All	Premium	Penalty	Prostitutes	No prostitutes
Correlation	0.550 [0.43, 0.613]	0.535 [0.402, 0.611]	0.665 [0.543, 0.725]	-0.217 [-0.462, 0.043]	0.568 [0.448, 0.624]
Observations	1,159	954	205	55	1,104

Notes: Following Kranz & Putz (2022), the table shows, for selected subsets of the literature, the correlation coefficient between the logarithm of the absolute value of the beauty effect and the logarithm of the corresponding standard error, weighted by the inverse publication probability estimated by the Andrews & Kasy (2019) model. If the assumptions of the model hold, the correlation is zero. Bootstrapped 95% confidence interval in parentheses.

Table B2: Caliper tests suggest selection for positive estimates

	t-statistic = 0	t-statistic = 1.96	t-statistic = 2.58
Caliper 0.05	0.119 (0.109) N = 21	0.176** (0.078) N = 37	0.133 (0.089) N = 30
Caliper 0.1	0.196*** (0.069) N = 46	0.074 (0.064) N = 61	0.064 (0.067) N = 55
Caliper 0.15	0.222*** (0.062) N = 54	0.054 (0.055) N = 83	0.059 (0.061) N = 68
Caliper 0.2	0.205*** (0.052) N = 78	0.019 (0.049) N = 106	0.024 (0.055) N = 82
Caliper 0.25	0.173*** (0.048) N = 98	0.04 (0.043) N = 137	0 (0.05) N = 102

Notes: The table reports results for caliper tests introduced by Gerber & Malhotra (2008). The tests compare the relative frequency of estimates above and below an important threshold for the t-statistic; the rows show results for different caliper widths. A test statistic of 0.176, for example, means that 67.6% estimates are just above the threshold and 32.4% estimates are just below the threshold. N = number of observations. Standard errors are reported in parentheses and clustered at the study level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B3: Publication bias tests separately for beauty premiums and penalties

Part 1. Subsample of penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.165**	-0.354	0.702***	0.578	0.354***
(*standard error*)	(0.0678)	(0.235)	(0.168)	(1.285)	(0.131)
	[0.021, 0.406]			{-1.941, 3.096}	[0.028, 1.395]
Effect beyond bias	2.647***	4.801***	0.447	0.933	2.655**
(*constant*)	(0.733)	(0.977)	(0.877)	(5.565)	(1.153)
	[1.225, 4.378]			{-3.133, 4.998}	[0.038, 5.782]
First-stage F-stat				0.8	
Observations	205	205	205	205	205
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.097***			1.097***	P = 0.369
	(0.164)			(0.12)	(0.019)
	[0.709, 1.489]			[0.709, 1.489]	
Effect beyond bias	0.139***	0.244***	0.245	0.139	0.236***
	(0.0531)	(0.021)	(0.323)	(0.097)	(0.067)
	[-0.405, 1.425]			[-0.405, 1.425]	
Observations	205	205	205	205	205

Part 2. Sample without penalties

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.437***	0.282*	0.745***	0.772*	0.666**
(*standard error*)	(0.139)	(0.160)	(0.172)	(0.440)	(0.281)
	[0.125, 0.745]			{0.032, 1.949}	[-0.044, 1.295]
Effect beyond bias	2.881***	3.448***	2.228**	1.652	3.470***
(*constant*)	(0.555)	(0.585)	(0.914)	(1.580)	(1.202)
	[1.745, 4.058]			{0.068, 4.171}	[0.845, 6.305]
First-stage robust F-stat				17.3	
Observations	954	954	954	954	954
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.881***			1.881***	P = 0.300
	(0.246)			(0.246)	(0.037)
	[1.351, 2.378]			[1.351, 2.378]	
Effect beyond bias	0.346	0.38*	0.013	0.346	0.200
	(0.294)	(0.229)	(1.323)	(0.294)	(0.828)
	[-0.051, 2.285]			[-0.051, 2.285]	
Observations	954	954	954	954	954

Notes: Part 1 only includes estimates that measure the effect of below-average looks (as always, recomputed to represent the percent increase in earnings or productivity following a one-standard-deviation increase in beauty). Panel A reports the results of regression $\widehat{b}_{ij} = b_0 + \beta \cdot SE(b_{ij}) + \epsilon_{ij}$, where $\widehat{b}_{ij}$ denotes the i-th beauty effect estimated in the j-th study, and $SE(b_{ij})$ denotes its standard error. FE = study-level fixed effects, BE = study-level between effects, MAIVE = Meta-Analysis Instrumental Variable Estimator (Irsova *et al.*, 2023) with the inverse of the square root of the sample size used as an instrument for the standard error. Weighted = the inverse of the number of estimates per study is used as the weight. In Panel B all models are weighted by inverse variance. The first specification reports a regression similar to those from the last column of Panel A but with inverse variance weights. WAAP = Weighted Average of the Adequately Powered estimates (Ioannidis *et al.*, 2017); Stem = the model by Furukawa (2020); Kink = the model by Bom & Rachinger (2019); Selection = the model by Andrews & Kasy (2019). P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). Standard errors, clustered at the study level, are reported in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018) are reported in square brackets. For MAIVE, in curly brackets we show the Anderson-Rubin 95% confidence interval recommended by Keane & Neal (2023). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B4: Publication bias tests separately for prostitutes and other occupations

Part 1. Subsample of prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.148	1.467	-0.756	-1.567	-0.468**
(*standard error*)	(0.776)	(0.782)	(1.032)	(0.967)	(0.210)
	[-2.956, 2.671]			{-7.405, -0.514}	[-3.545, 0.383]
Effect beyond bias	8.136***	4.488	11.260*	12.880***	11.760***
(*constant*)	(2.767)	(2.164)	(2.686)	(1.392)	(0.414)
	[1.525, 15.17]			{1.110, 17.173 }	[5.245, 13.97]
First-stage robust F-stat				15.8	
Observations	55	55	55	55	55
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	-2.154***			-2.126	P = 1.215
	(0.745)			(1.878)	(0.091)
	[-4.057, -0.568]			[-4.326, -0.755]	
Effect beyond bias	13.290***	12.168***	11.205***	12.214***	8.490***
	(0.298)	(0.372)	(0.905)	(0.349)	(1.691)
	[-22.03, 16.87]			[-22.49, 17.96]	
Observations	55	55	55	55	55

Part 2. Sample without prostitutes

Panel A	OLS	FE	BE	MAIVE	Weighted
Publication bias	0.391***	0.204*	0.800***	0.875*	0.718***
(*standard error*)	(0.117)	(0.119)	(0.161)	(0.450)	(0.271)
	[0.120, 0.642]			{0.118, 2.078}	[0.061, 1.331]
Effect beyond bias	2.581***	3.289***	1.603*	0.741	2.617**
(*constant*)	(0.447)	(0.453)	(0.875)	(1.728)	(1.061)
	[1.660, 3.501]			{0.048, 3.696}	[0.333, 5.113]
First-stage robust F-stat				16.2	
Observations	1,104	1,104	1,104	1,104	1,104
Panel B	Precision-weighted	WAAP	Stem	Kink	Selection
Publication bias	1.610***			1.610***	P = 0.307
	(0.202)			(0.202)	(0.039)
	[1.184, 2.051]			[1.184, 2.051]	
Effect beyond bias	0.152	0.229***	0.008	0.152	0.669***
	(0.118)	(0.07)	(0.798)	(0.118)	(0.231)
	[-0.050, 0.963]			[-0.050, 0.963]	
Observations	1,104	1,104	1,104	1,104	1,104

Notes: Part 1 only includes estimates that measure the beauty effect among prostitutes. Panel A reports the results of regression $\widehat{b}_{ij} = b_0 + \beta \cdot SE(b_{ij}) + \epsilon_{ij}$, where $\widehat{b}_{ij}$ denotes the i-th beauty effect estimated in the j-th study, and $SE(b_{ij})$ denotes its standard error. FE = study-level fixed effects, BE = study-level between effects, MAIVE = Meta-Analysis Instrumental Variable Estimator (Irsova *et al.*, 2023) with the inverse of the square root of the sample size used as an instrument for the standard error. Weighted = the inverse of the number of estimates per study is used as the weight. In Panel B all models are weighted by inverse variance. The first specification reports a regression similar to those from the last column of Panel A but with inverse variance weights. WAAP = Weighted Average of the Adequately Powered estimates (Ioannidis *et al.*, 2017); Stem = the model by Furukawa (2020); Kink = the model by Bom & Rachinger (2019); Selection = the model by Andrews & Kasy (2019). P denotes the probability that estimates insignificant at the 5% level are published relative to the probability that significant estimates are published (normalized at 1). Standard errors, clustered at the study level, are reported in parentheses. 95% confidence intervals from wild bootstrap (Roodman *et al.*, 2018) are reported in square brackets. For MAIVE, in curly brackets we show the Anderson-Rubin 95% confidence interval recommended by Keane & Neal (2023). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table B5: Diagnostics of the baseline BMA estimation (UIP and dilution priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
7.4321	$3 \cdot 10^5$	$1 \cdot 10^5$	1.82 mins	53,604
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$8.6 \cdot 10^9$	0.06%	100%	0.9967	1,159
Model prior	*g-prior*	*Shrinkage-stats*		
Dilution / 16.5	UIP	Av = 0.9991		

Notes: We employ the combination of the unit information prior recommended by (Eicher *et al.*, 2011) and dilution prior suggested by George (2010), which accounts for collinearity.

Figure B1: Model size and convergence of the baseline BMA estimation

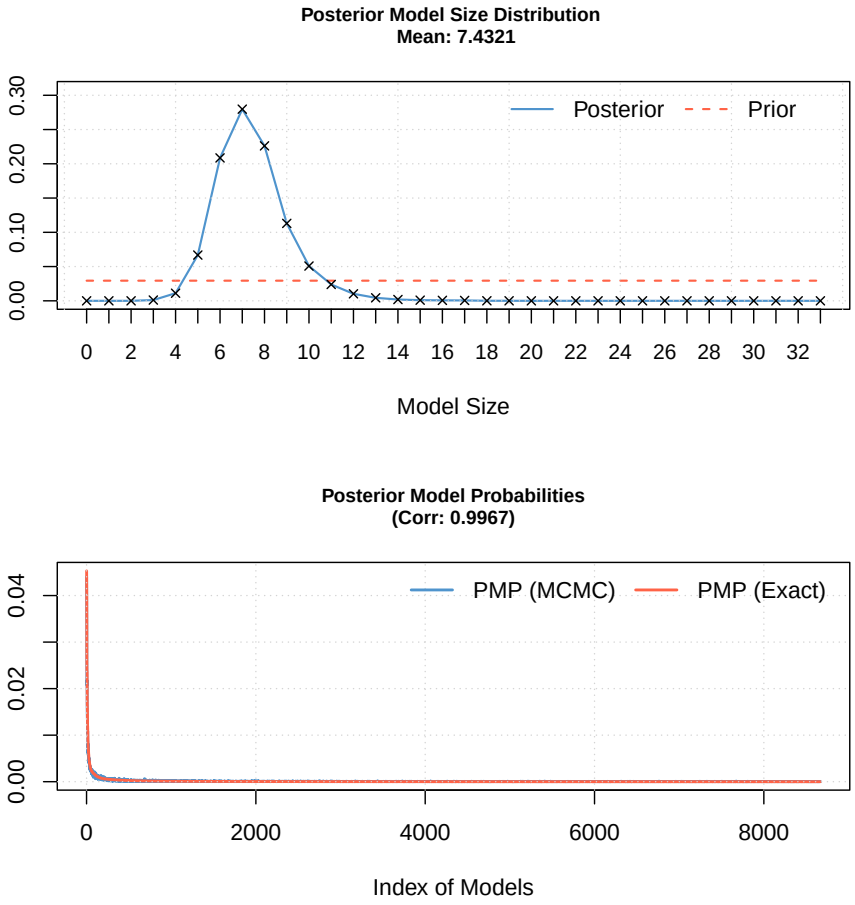
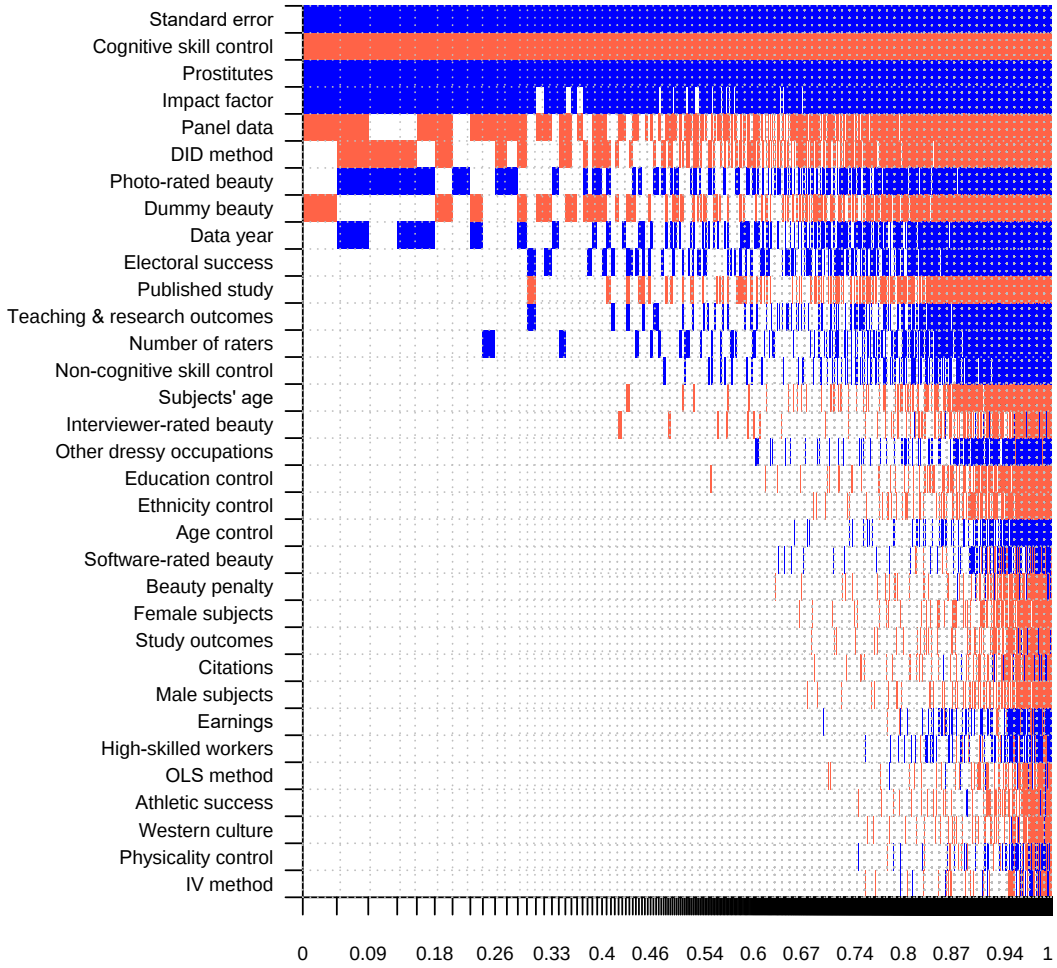


Figure B2: Model inclusion in BMA (BRIC and random priors)



Notes: On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table B6. All variables are described in Table 4.

Table B6: Why reported beauty premiums vary (robustness checks)

Response variable: Beauty premium	Bayesian model averaging (BRIC and random priors)			Ordinary least squares (only for PIP > 0.5)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Constant	2.756	NA	1.000	4.326	1.728	0.012
Standard error	0.435	0.042	1.000	0.422	0.112	0.000
Measurement of beauty						
Interviewer-rated beauty	-0.033	0.201	0.035			
Photo-rated beauty	0.592	0.745	0.425			
Software-rated beauty	0.008	0.140	0.014			
Dummy beauty	-0.485	0.663	0.386			
Beauty penalty	-0.008	0.093	0.013			
Number of raters	0.051	0.139	0.134			
Measurement of success						
Earnings	0.005	0.078	0.010			
Study outcomes	-0.007	0.101	0.011			
Teaching & research outcomes	0.236	0.641	0.140			
Athletic success	-0.004	0.107	0.007			
Electoral success	0.669	1.358	0.224			
Data characteristics						
Male subjects	-0.005	0.064	0.010			
Female subjects	-0.006	0.069	0.012			
Subjects' age	-0.074	0.331	0.060			
High-skilled workers	0.002	0.062	0.008			
Prostitutes	4.793	1.060	0.999	4.170	1.652	0.012
Other dressy occupations	0.035	0.224	0.032			
Western culture	-0.001	0.039	0.006			
Panel data	-1.131	1.017	0.605	-1.795	1.739	0.302
Data year	0.336	0.503	0.345			
Estimation technique						
OLS method	-0.002	0.048	0.007			
IV method	-0.001	0.059	0.005			
DID method	-2.481	2.895	0.469			
Age control	0.014	0.129	0.017			
Education control	-0.016	0.121	0.025			
Ethnicity control	-0.012	0.106	0.020			
Cognitive skill control	-2.285	0.447	1.000	-2.194	0.707	0.002
Non-cognitive skill control	0.072	0.291	0.070			
Physicality control	0.000	0.036	0.006			
Publication characteristics						
Published study	-0.311	0.745	0.172			
Impact factor	0.265	0.123	0.907	0.194	0.185	0.293
Citations	-0.002	0.034	0.011			
Studies	67			67		
Observations	1,159			1,159		

Notes: The posterior mean in BMA denotes the partial derivative of the reported beauty premium with respect to the corresponding study characteristic. For example, including a control for cognitive skills typically reduces the beauty premium by 2.3 percentage points. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. BMA employs the BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009). The frequentist check only includes variables with PIP above 0.5; standard errors clustered at the study level. All variables are described in Table 4. Technical details and diagnostics of the BMA exercise are available in Table B7 and Figure B3.

Table B7: Diagnostics of the alternative BMA estimation (BRIC and random priors)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
7.3807	$3 \cdot 10^5$	$1 \cdot 10^5$	1.51 mins	53,598
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$8.6 \cdot 10^9$	0.06%	100%	0.9947	1,159
Model prior	*g-prior*	*Shrinkage-stats*		
Random / 16.5	BRIC	Av = 0.9991		

Notes: The specification uses the BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009).

Figure B3: Model size and convergence of the alternative BMA estimation (BRIC and random priors)

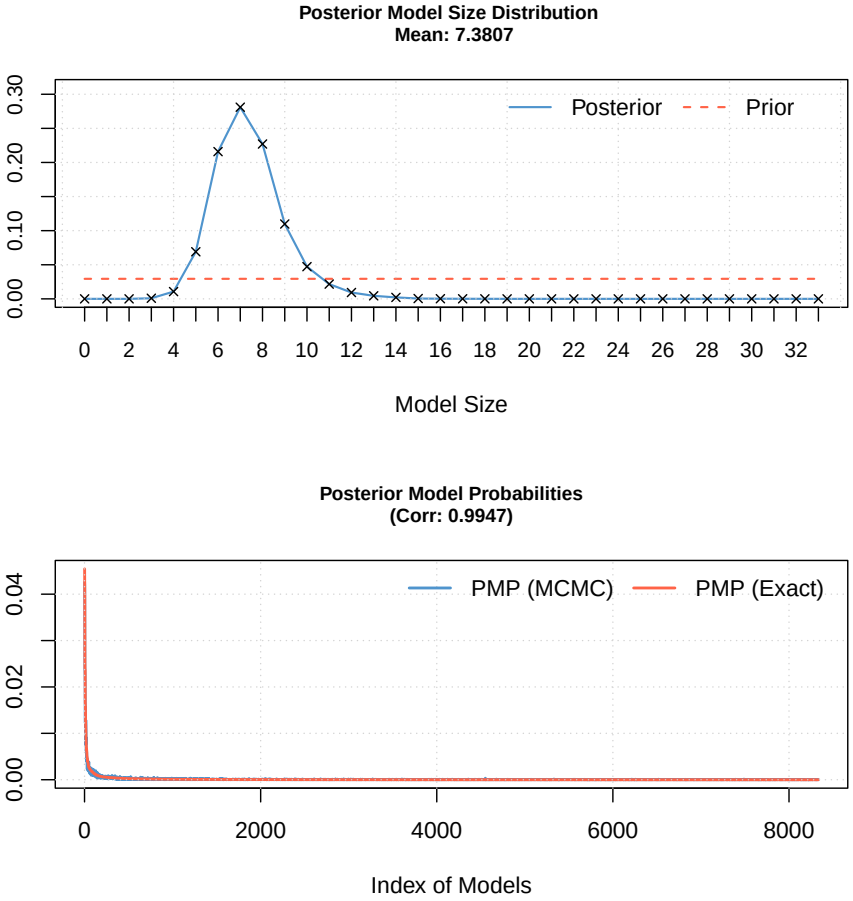
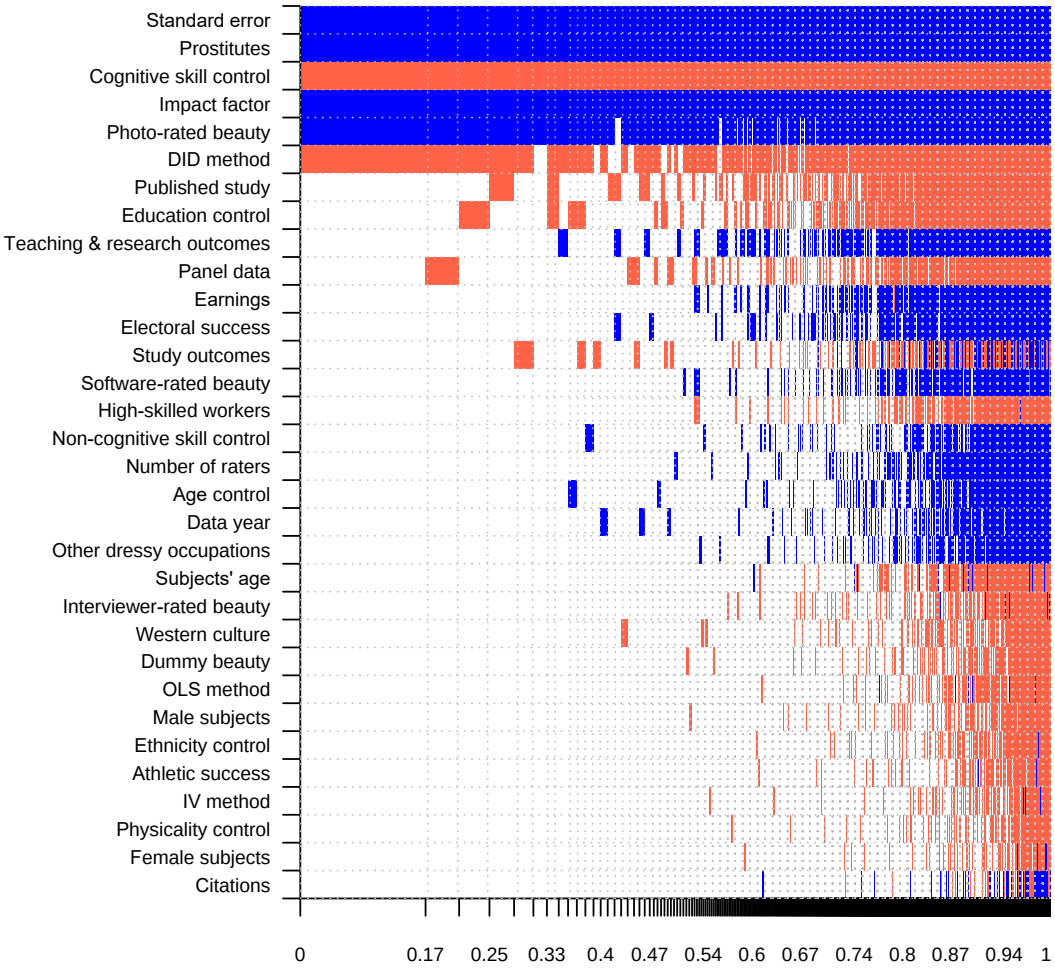


Figure B4: Model inclusion in BMA (beauty penalties excluded)



Notes: We exclude estimates that focus on the effect of below-average looks. On the vertical axis the explanatory variables are ranked according to their posterior inclusion probabilities from the highest at the top to the lowest at the bottom. The horizontal axis shows the values of cumulative posterior model probability. Blue color (darker in grayscale) = the estimated parameter of a corresponding explanatory variable is positive. Red color (lighter in grayscale) = the estimated parameter of a corresponding explanatory variable is negative. No color = the corresponding explanatory variable is not included in the model. Numerical results are reported in Table B8. All variables are described in Table 4.

Table B8: Why reported beauty premiums vary (penalties excluded)

Response variable: Beauty premium	Bayesian model averaging (UIP and dilution priors)			Ordinary least squares (only for PIP > 0.5)		
	P. mean	P. SD	PIP	Mean	SE	p-value
Constant	2.223	NA	1.000	1.695	0.995	0.088
Standard error	0.498	0.046	1.000	0.502	0.135	0.000
Measurement of beauty						
Interviewer-rated beauty	-0.044	0.285	0.036			
Photo-rated beauty	1.854	0.795	0.904	2.199	0.969	0.023
Software-rated beauty	0.247	0.902	0.089			
Dummy beauty	-0.019	0.143	0.026			
Number of raters	0.025	0.113	0.062			
Measurement of success						
Earnings	0.341	0.950	0.156			
Study outcomes	-0.096	0.800	0.140			
Teaching & research outcomes	0.648	1.317	0.252			
Athletic success	-0.022	0.275	0.014			
Electoral success	0.521	1.425	0.146			
Data characteristics						
Male subjects	-0.011	0.096	0.020			
Female subjects	-0.002	0.043	0.008			
Subjects' age	-0.049	0.298	0.039			
High-skilled workers	-0.115	0.489	0.067			
Prostitutes	6.320	0.979	1.000	6.521	1.871	0.000
Other dressy occupations	0.061	0.330	0.045			
Western culture	-0.029	0.175	0.035			
Panel data	-0.299	0.659	0.198			
Data year	0.044	0.198	0.061			
Estimation technique						
OLS method	-0.020	0.181	0.023			
IV method	-0.013	0.158	0.013			
DID method	-4.882	2.838	0.809	-6.361	2.472	0.010
Age control	0.078	0.348	0.061			
Education control	-0.306	0.562	0.265			
Ethnicity control	-0.010	0.103	0.018			
Cognitive skill control	-3.261	0.483	1.000	-3.542	0.799	0.000
Non-cognitive skill control	0.065	0.282	0.063			
Physicality control	-0.006	0.074	0.012			
Publication characteristics						
Published study	-0.519	0.935	0.272			
Impact factor	0.381	0.098	0.997	0.347	0.193	0.072
Citations	0.000	0.024	0.008			
Studies	67			67		
Observations	954			954		

Notes: We exclude estimates that focus on the effect of below-average looks. The posterior mean in BMA denotes the partial derivative of the reported beauty premium with respect to the corresponding study characteristic. For example, including a control for cognitive skills typically reduces the beauty premium by 2.3 percentage points. P. mean = posterior mean, P. SD = posterior standard deviation, PIP = posterior inclusion probability, SE = standard error. BMA employs the BRIC g-prior suggested by Fernandez *et al.* (2001) and the beta-binomial model prior according to Ley & Steel (2009). The frequentist check only includes variables with PIP above 0.5; standard errors clustered at the study level. All variables are described in Table 4. Technical details and diagnostics of the BMA exercise are available in Table B9 and Figure B5.

Table B9: Diagnostics of the BMA estimation (beauty penalties excluded)

Mean no. regressors	*Draws*	*Burn-ins*	*Time*	*No. models visited*
8.3506	$3 \cdot 10^5$	$1 \cdot 10^5$	1.71 mins	49,500
Modelspace	*Visited*	*Topmodels*	*Corr PMP*	*No. obs.*
$4.3 \cdot 10^9$	0.12%	100%	0.9992	954
Model prior	*g-prior*	*Shrinkage-stats*		
Dilution / 16	UIP	Av = 0.999		

Notes: We employ the combination of the unit information prior recommended by (Eicher *et al.*, 2011) and dilution prior suggested by George (2010), which accounts for collinearity.

Figure B5: Model size and convergence of the BMA estimation (beauty penalties excluded)