SUPPORTING INFORMATION

Extracting Multistage Screening Rules from Online Dating Activity Data

Elizabeth Bruch
Department of Sociology & Complex Systems
University of Michigan

Fred Feinberg
Ross School of Business & Department of Statistics
University of Michigan

Kee Yeun Lee
Department of Management & Marketing
Hong Kong Polytechnic University

August 2016

CONTENTS OF SUPPORTING TEXT

List of Figures

List of Tables

S1. METHODOLOGICAL DETAILS

Splines and Spline Bases

Figure S1a depicts a trilinear spline utility function associated with a particular person's reaction, say, to the Age covariate, with a decreasing response to Age throughout the observable range of the data (X axis standardized for clarity to -3 to 3). There are two "knots", located at -1.0 and at 1.5; that is, bounding approximately the 16th to 93rd percentiles of the "X" distribution. Figure S-1b depicts the spline basis for Figure S-1a.

Model Estimation

We present a general strategy for estimating discrete choice models that can identify both slopes and knots for continuous attributes, and also allow for multiple decision stages (i.e., browsing and writing) and multiple observations per stage (i.e., multiple instances of browsing and writing for each user).[1,2] We break the statistical model into two parts: a heterogeneous regression model (e.g., one with a given knot configuration) and an efficient method to search among knot configurations. We then use a combination of stochastic and gradient-based methods to iterate between estimating the two-stage, latent class models for a given set of knots, and exploring the space of possible knots. Estimation leverages specific capabilities of commercially available software packages, Latent Gold for heterogeneous (latent class) regression conditional on a set of knots, and Matlab to efficiently search the space of knot configurations. An open-source R package, `StagedChoiceSplineMix`, is available from CRAN (https://cran.r-project.org/). The code implements our entire model, albeit substantially less efficiently compared with the joint commercial solution.

In any latent class, the task is to estimate all parameters $\{\beta_{0i}, \beta_{1ik}, \beta_{2ik}, \delta_{1ik}, \delta_{2ik}, \gamma_{il}\}$, for both the Browsing and Writing stages. These are conditional on the observed outcomes, which are related probabilistically to $\{V_{ij}^{B}, V_{ij}^{W}\}$ by the binary logit formulas in equations 2 and 3 from the main article text. The general strategy for estimation starts, as suggested by Harrell (2001), by dividing up all to-be-splined covariates in one of two ways: (a) according to some 'natural' breaking points suggested by the data (e.g., years for Age); or (b) using some convenient quantiles (e.g., 10% or 5%). In our data specifically, except for Height, which has a natural range

[1] Note that it is critical for researchers to determine which covariates are substantively relevant and explanatorily significant using lower-dimensional (perhaps polynomial) models and interactions before attempting to run spline versions of the model. The key advantages of splines are as laid out in the paper: avoiding high-order asymptotics and identifying abrupt changes in utility; they are not to determine which variables should be included in the model in the first place.

[2] Our approach is not intended to solve other well-known modeling issues, like multicollinearity. More generally, researchers need to resolve such data input issues using well-known processing and testing methods like factor or principal components analysis, or simply avoiding variables that are either strongly substantively or empirically correlated. At the same time, this modeling framework is not meant to be causal; mild to moderate correlations in a reduced-form framework are not threats to validity.

that accords to 12 units, we use 20-iles for our other variables, Age and BMI. So, we are left with all our to-be-splined covariates re-coded as in Figure S-1b, as per our potential knot configurations.

The key task is to explore those knot configurations. To make the issue concrete, for our application, we have 19 knots (corresponding to 20-iles) for Age and BMI, and 11 for Height. For each spline with K potential knots, the number of possible trilinear spline configurations is therefore $K(K\text{-}1)/2$, since the two knots must be in order, and not the same. For our specific values (19 and 11), this yields 19*18*19*18*11*10/8 = 1,608,255 possible configurations of 6 knots. Since we have two stages, Browsing and Writing, the total number of configurations is 1,608,255 squared, or 2.6×10^{12}, for 12 knots. Each additional latent class increases the size of the search space by this quantity; in our final model, with five latent classes (and 60 knots), the number of knot configurations is therefore $[2.6 \times 10^{12}]^5 = 1.2 \times 10^{62}$. Globally exhaustive search is well beyond any present or near-term computational method. Instead, we use a search strategy that takes advantage of the discretized nature of the splined-variables:

Step 1) For each to-be-splined covariate, choose knots that are approximately 1/3 and 2/3 of the way "across". For our applications, this would be knots 7 and 13 for the 19-knotted covariates, and knots 4 and 8 for the 11-knotted one.

Step 2) Estimate a homogeneous model (i.e., one latent class) for the model including all desired splined variables in both Browsing and Writing, as well as all additional categorical covariates. These will serve as starting values for the latent class model.

Step 3) Given an initial knot configuration (for our final model consisting of 60 knots), *randomly move* 40% of the knots either one unit to the left, or one to the right, then re-estimate the entire model using the homogeneous coefficients as starting values. Perform this for 500 generated random configurations, and choose the two best of these (judging by log-likelihood) as updated candidate models. [The 40% figure is a 'tunable' value left to the researcher's discretion and experience, based on convergence speed. The same is true for running 500 models on each pass.]

Step 4) Repeat Step 3 using each of the two candidate models; that is, generate 500 randomly re-knotted (with 40% of the knots changing) models for each, which are then estimated. The two best of *that* process are retained. In each case, check to ensure that the knot configuration is legitimate, that is, the left knot is smaller than the right knot.

Step 5) Step 4 is repeated until the changes in log-likelihood are relatively small, or the process is no longer yielding better-fitting models. Because something on the order of 5000 models have been, and the knot configuration space is on the order of 10^{62}, further search is needed to ensure a good fit.

Step 6) Retain the best-fitting model thus far. Then set up every model that consists of *exactly one* of the 12 knots (per latent class) moving up one or down one, yielding 24 models per latent class. For a 5-class model, this requires 120 new models to be run. Retain the best of these as the new starting point.

Step 7) Repeat Step 6 until there is no further improvement (or improvements on the order of rounding error in calculation, roughly a million times smaller than the log-likelihood itself).

The procedure above will terminate in a *local maximum*, but, as discussed above, it is impossible to guarantee a global one. To ensure that we have avoided 'bad' local maxima, we re-run the entire procedure using several new starting values for the knot configuration – preferably ones fairly far from the original "equal spacing" and from any already-obtained local maximum – and end the procedure when many of the starting values converge to approximately the same point.

S2. DATA

There are several advantages to using online dating data to study mate choice. First, since we know who is active on the site at any given time, the choice set can be observed and explicitly represented for all members of the site.[3] Second, all overtures are observed regardless of whether or not they are reciprocated. Third, because all members of the site are theoretically just a click away, the online dating environment (at least in theory) represents a frictionless environment. After all, the whole purpose of online dating is to reduce search costs in the marriage and dating market.

A potential disadvantage of online data, of course, is that we do not observe if two users who met online actually date, cohabit, or marry. Given that more homogamous matches tend to persist over time, we would expect that the preferences estimated from online dating data would imply more heterogamous matches than what prevailing marriages would suggest (Schwartz 2010). Overall, these data should be viewed as representing preferences at an early stage of mate selection. It is likely that preferences for more superficial attributes are most salient in the early stages of mate choice, when people have had less opportunity to bond on more idiosyncratic traits or chemistry.

Upon joining a dating service, users must fill out a profile providing answers to a number of survey questions, as well as fill out several free text fields based on prompts (e.g., "About Me"). The survey questions include measures of a range of demographic attributes (income, marital status, whether they have children, education, race, religion, and age) as well as measures of cultural interests or expectations, for example what kinds of food or music they enjoy, whether they attend church frequently, and whether they are open to a long-distance relationship. Many users also include one or more photos in their profile. Once they have completed their profile, users can search for, browse, and write to potential partners.

Users typically begin by searching for mates based on a specified age range and geographic region. This query returns a list of "short profiles" containing information on potential partners' age, user name, a brief description, and a photo if available. Users can then

[3] Of course, this does not include choices available outside the online dating context. For example, someone may simultaneously be using an online dating site and also going to bars to meet people in real life. But we can evaluate the *relative* desirability of potential mates available online. In other words, if we observe the i^{th} user sending a note to the j^{th} potential match, we can reasonably assume that user i preferred potential match j over other online matches who were not written to.

decide to "browse" potential mates by clicking on their short profile to access the complete profile containing the full set of profile attributes as well as essay questions, larger versions of the main photo, and additional photos if available. Based on the full profile, users can then decide to write to a potential mate. The data provided a complete moment-by-moment description of users' activities, including which profiles he or she browsed, whether or not the photos were viewed, and whether the user sent a first contact message.

We have made anonymized spline variables from the online dating data available for readers to examine through OpenICPSR: https://www.openicpsr.org/openicpsr/project/100228/view. Note that these files are intended to provide readers with a sense of the format and structure of these activity data; they are not appropriate for further research purposes.

Descriptive Statistics

Table S-1 lists the attributes used in the analysis taken from the user registration data. The profile data contain a variety of attributes, including users' age, education, income, height, weight, and marital status. Our analysis focuses on attributes that are revealed on the profile website and have been shown in prior studies to matter in mate choice. These include three continuous attributes: height, body mass index (BMI), and age. We also include categorical predictors for marital status (never married vs. separated, widowed, or divorced), children (kids vs. no kids), smoking (smoker vs. nonsmoker), and education (less than college, college degree, or a post-college degree).

The average age of both men and women on our site is approximately 35. Reported height is 70.3" for men, and 64.5" for women, which closely match US averages. Roughly 70% have never been married, with women slightly more likely to have cohabiting children (28% vs. 24% for men). The sample is more educated than the general population, with almost half (48%) completing college, which may reflect the online pool. Site behavior mirrors findings from other dating sites, e.g., women browse more profiles (134.8 vs. 121.5) but send far fewer messages (12.0 vs. 21.4) over the 3-month observation period.

New female subscribers (1159) outpaced male (696) during the observation window. This prevalence of female users on the site is consistent with a larger body of data (from the Census and other dating sites) documenting an excess of single women in the New York metro area.

S3. ADDITIONAL RESULTS

We compare the fit of models with and without latent classes, as well as models that allow for conventional representation of continuous covariates. Table 1 summarizes these results, and reports goodness of fit statistics for all models.[4] Based on both the BIC statistic and L^2, the latent class models fit the data far better than the homogeneous models. We found that model fit

[4] The L^2 – or "likelihood ratio chi-squared" – statistic assesses how well the observed variation in browsing and writing decisions is explained by a given model vs. one making perfect predictions. The BIC statistic is an alternative measure of goodness of fit that penalizes less parsimonious models. For both statistics, smaller values correspond to superior fit.

improved substantially with each additional class. But beyond 5 classes, despite having a slightly improved fit, the models for both men and women led to degenerate solutions where the added class was extremely small, and the coefficient estimates had large standard errors. Full estimation results appear in Tables S-3 (continuous covariates) and S-4 (categorical covariates).

Response Profiles over Range of Covariates

Figure S-2 displays plots for how having children affects people's likelihood of being browsed or written to on the site. The top pane reports results for men, the bottom for women. Figures S-3 and S-4 report men's and women's relative risk of browsing or writing to a person of a given educational attainment for users with a college or graduate degree.

As a supplement, below these appear plots of the total range in log-odds for each of the continuous attributes over the bulk of the data, and we also highlight categorical attributes that play a disproportionate role in the mate choice decision process. Figures S-5 and S-6 present box-and-whisker plots that show the total change in log-odds observed over the observed range of continuous covariates on the probability of browsing or writing a potential mate. In all plots, the horizontal line represents the median log-odds, and the shaded area captures the 25^{th} to 75^{th} percentile. The whiskers capture the furthest observations that still fall within +/- 1.5 times the interquartile range of the graphed quantity. All observations outside of the whiskers are outliers. To ensure that these plots cover a meaningful range of attribute values and do not focus our attention on anomalous outliers, we again remove observations at the very extremes (top and bottom 1%). We construct these plots so that the Y-axis is constant within a given attribute, so that we can compare its relative effect at the browsing and writing stages. But we allowed the Y-axes to differ across attributes to avoid plots (mainly for BMI and Height, which have much smaller ranges than age) that are too compressed to read.

Demographic Variation across Classes

The latent class model assigns people to classes probabilistically; each user's overall likelihood is a weighted combination of the likelihoods across all five classes, and all users are incorporated into the estimation of coefficients for each class. However, for presentation purposes we assign each user to his or her *modal* class, that is, the class with the highest assignment probability. In the vast majority of cases, these modal probabilities are greater than 0.9, suggesting that the model does a good job of distinguishing heterogeneous behavior among users. The average attributes associated with women and men by modal class are presented in Tables S-2a and S-2b, respectively.

Table S-2a shows that while the average age of women in most classes are in the early to mid-thirties, women in Class 3 tend to be substantially older. The average age in this class is around 40. Moreover, women in this class are more likely to have children living at home (38%), less likely to be never married (60%), and are more likely to have a high school education (38%). This is also our smallest class, with 12% of our users having this modal assignment. Class 1 women also are less educated than the other classes; only 3 percent of these women report having a post-college degree. However, this is also our youngest class (average is 29.89), so it's possible that some of these women have not completed their education. Class 5 women have the highest rates of educational attainment. Ninety six percent of the women assigned to this class have at least a college degree.

Table S-2b reports attributes for men by modal class assignment. We see that Class 1 is our oldest class, and also our heaviest with an average BMI of 25.19 (typically BMIs of between 18.5-24.9 are considered "normal range," so this is slightly overweight). Class 1 men are also most likely to be never married (76%). This is the smallest of the men's classes, with only 8 percent of the men having this modal assignment. Class 2 contains the greatest number of divorced fathers; 36 percent of these men report having children at home at least part of the time, and 41 percent report having been married in the past. Class 3 men are both less educated, on average, and more likely to smoke, compared to the other classes. Finally, the Class 5 men appear to be mostly young professionals: none of these men report children living at home, and almost all of them (98%) have at least a college degree. These attribute differences are helpful in interpreting the variations in behavior observed in browsing and writing.

While the summary statistics are useful in capturing average differences across classes, they conceal a great deal of within-class variation. Figure S-7 presents boxplots of the distribution of Age, BMI, and height within classes for men and women. The median values for each variable are shown as a horizontal bar. The only median value that substantially deviates from the means reported in Table S-2 is age for the Class 2 men. While the average value for these men is approximately 35, we see that half the men in this class are in their early 20s. The other striking finding from these graphs is the huge range of ages in women's Class 3. Recall that these women are, on average, substantially older than other women in our sample and tend to pursue men who are a good deal older than they are. The boxplots reveal that this class also contains a nontrivial proportion of younger women as well, albeit ones who also appear to pursue older men. The overall takeaway from these graphs is that—while differences do exist across classes of men and women—there is a great deal of overlap in their values.

S4. COMPARISON TO CONVENTIONAL MODELING APPROACHES

While Table 1 reveals that the two-stage spline models with unobserved heterogeneity fit the data better than other specifications, it does not reveal how the substantive story told in the splines differs from more conventional approaches. In this final set of analyses, we compare the substantive content conveyed in Figure 3 of the paper with what we would learn from a more conventional decision model. Since unobserved heterogeneity is standard in most statistical software packages, an appropriate comparison is between our model and a single-stage choice model with a polynomial representation of nonlinearity plus unobserved heterogeneity, as opposed to homogeneous analogs, which typically fit real-world choice data poorly.

Figures S-8a through S-8d contrast what a conventional model infers about how men and women respond to age, BMI, and height differences with what we learn from the multistage spline models. The y-axis is the log-odds of browsing or writing. The colors identifying latent classes in Figures 8b and 8d correspond to the colors identifying classes in Figure 3. Since these are marginal effects, we focus on the shape of the response function across values of age, BMI, and height differences. In examining results from the conventional model (8a and 8c), we see that while different rules apply at different stages—and there is clear heterogeneity in behavior across classes—it is impossible to link class-specific behavior at each stage. In contrast to Figures Sb and 8bd, we see that the cubic functions smooth out all sharp cutoffs, making it difficult to identify potential "rules" people are using to select mates. But most critically, because the whole

range of data—not just local information—drives the shape of the cubic, we observe a number of substantively erroneous results from the conventional models. For example, in Figure S-8a, the red line in Panel B suggests that one class of women is most likely to write to men who are younger than they are. Similarly in Panel F, the blue line implies that one class of women pursues men who are around 5 inches shorter than they are. We also find odd maxima in the results for men, shown in Figure S-8b. The red line in Panel C suggests that there is a class of men who prefer women who are 8-10% heavier than they are. None of these findings manifest in the spline results shown in Figures 8b and 8d (as well as Figure 3 in the main paper). They are also at odds with underlying patterns in the data, both ours and those of prior inquiry in the area (e.g., women who actually prefer shorter men). Rather, they are an apparent artifact of the cubic "needing to get the asymptotics correct," at the expense of correctly identifying other, substantively salient features of the response curve, such as the modally optimal height, BMI, or age within-class.

S5. R CODE, SIMULATED DATA, MODEL COMPARISONS

While the models reported in the body text were—due to data size—estimated using a combination of Latent Gold and Matlab, we also provide R code that implements the model. Five R functions (functions.r) and one simulated data set (simdata.csv) were used for the simulation study. The entire package, `StagedChoiceSplineMix`, can be downloaded from the CRAN archive (https://cran.r-project.org/).

1. *R* functions

`StagedChoiceSplineMix`

Description

The function performs iterations between an EM algorithm for a mixture of two-stage logistic regressions with fixed knots and knot movements. The function generates candidate knots for each splined variable. Four additional functions are used within the function. Brief explanations of these four functions are as follows.

1) `gen.init`: generates initial values for mixtures of logistic regressions. This function is used if starting points for parameters are not specified by the user or when the EM algorithm needs to be initialized due to errors.

2) `twostglogitregmixEM`: performs an EM algorithm for mixtures of two-stage logistic regressions.

3) `move.knot`: generates a new set of knots for the following iteration. Please refer to the supporting information for the precise rule used.

4) `bs.se`: performs a parametric bootstrapping standard error approximation for mixtures of two-stage logistic regressions.

Usage

```
StagedChoiceSplineMix (data, M, sp.cols, num.knots, sp.knots,
betab, betaw, lambda k, nst, epsilon, maxit, maxrestarts, maxer)
```

Arguments

data:

raw data for `StagedChoiceSplineMix`

format

- 1st column: id
- 2nd column: the 1st stage binary variable (browsing)
- 3rd column: the 2nd stage binary variable (writing) conditional on the 1st stage binary variable. It should be left blank (or NA) if 1st stage variable is equal to 0
- The rest of columns: covariates including splined variables

M: number of iterations (def: 100)

sp.cols: vector of column numbers of splined variables in a data set (if sp.col is 0, twostglogitregmixEM function should be used)

num.knots: vector of numbers of knot candidates for splined variables. (def: a vector all of whose entries are "19")

sp.knots: list of knot configurations. For each splined variable, a knot configuration is a k by 4 matrix whose rows represent latent classes and columns represent knots [browsing knot 1, browsing knot 2, writing knot 1, writing knot 2]. (def: approximately 1/3 and 2/3 of knot candidates for knot 1 and knot 2 respectively)

betab: matrix of starting points for betab (browsing parameters). If not given, `gen.init` generates starting points.

betaw: matrix of starting points for betaw (writing parameters). If not given, `gen.init` generates starting points.

lambda: vector of starting points for lambda (membership proportion). If not given, `gen.init` generates starting points.

k: number of latent classes (def: 2)

nst: number of random multiple starting points to try given a knot configuration. For each knot configuration, the output with the largest log-likelihood is stored among nst trials. (def: 20)

epsilon: stopping tolerance for the EM algorithm. (def:1e-06)

maxit: maximum number of the EM iterations allowed. If convergence is not declared before maxit, the EM algorithm stops with an error message and generates new starting points. (def: 500)

maxrestarts: maximum number of restarts (due to a singularity problem) allowed in the EM iterations. If convergence is not declared before maxrestarts, the algorithm stops with an error message and generates new starting points. (def: 100)

maxer: maximum number of errors allowed within a given knot configuration. If convergence is not declared before maxer, it tries a new knot configuration. (def: 20)

Value

`StagedChoiceSplineMix` returns a list of the following items:

best: best output, i.e., that giving the largest log-likelihood among M outputs.

- best$loglik: log-likelihood
- best$betab: parameter estimates of betab
- best$betaw: parameter estimates of betaw
- best$lambda: parameter estimates of lambda
- best$sp.knots: knot configuration

loglike: vector of log-likelihoods for M outputs.

Example

```
## read five R functions in functions.r

## required package: plyr
#install.packages("plyr")
library(plyr)

## set a random seed
seed<-66
set.seed(seed)

## read data (simulated data)
data<-read.csv("simdata.csv")

## number of latent classes
k<-3

## starting points: true parameters used in the data generation (optional)
betab<-matrix(c(1.5,2.0,0.3,-0.2,-0.1,0.6,-2.5,0.5,1,3,-1,1,2,-0.5,-1.5),5,k,byrow=T)
betaw<-matrix(c(-2.0,-1,0.3,-0.3,-0.2,0.2,-3,-2,0,3,2,-1.5,3,2,-1),5,k,byrow=T)
lambda<-c(0.3,0.3,0.4)

## number of random multiple starting points to try given a knot configuration
```

```
nst<-20

## vector of the columns of spline variables in the data set (required)
sp.cols<-5

## vector of the numbers of candidate knots for splined variables (optional)
num.knots<-19

## true knot configuration used in the data generation
sp1.knots<-matrix(c(8,14,4,11,5,15,5,12,5,13,7,14),3,4,byrow=T)

## list of knot configuration of spline variables (optional)
sp.knots<-list(sp1.knots)

## run "StagedChoiceSplineMix"
out<-StagedChoiceSplineMix(data=data, M=100, sp.cols=sp.cols, num.knots, sp.knots, betab,
betaw, lambda, k=k, nst=nst, epsilon=1e-06, maxit=500, maxrestarts=100, maxer=20)

## output
out$loglike  # vector of  M log-likelihoods
out$best$loglik # log-likelihood of the best output
out$best$betab # betab estimates of the best output
out$best$betaw # betaw estimates of the best output
out$best$lambda # lambda estimates of the best output
out$best$sp.knots # knot configuration of the best output

## parametric bootstrapping to calculate standard error (best output)
out.se<-bs.se(output=out$best, B=100)  # B: number of bootstrap samples
out.se$betab.se
out.se$betaw.se
out.se$lambda.se
```

2. Simulated data

The simulated data (simdata.csv) is generated using information below:

User identifier: *userid*

- Number of users: 700
- Number of observations per user: 200

Dependent variables:

- *browsed*: 1st stage binary dependent variable
- *wrote*: 2nd stage binary dependent variable

Covariates:

- *x1* (discrete variable): random draws from a binomial distribution with 70% success probability
- *sp1* (splined variable): random draws from a uniform distribution between -2 and 2

Number of latent classes: 3

True parameters:

- Betab (browsing)

	Class1	Class 2	Class 3
Intercept	1.5	2	0.3
x1	-0.2	-0.1	0.6
sp1	-2.5	0.5	1
sp1 knot1	3	-1	1
sp1 knot2	2	-0.5	-1.5

- Betaw (writing)

	Class1	Class 2	Class 3
Intercept	-2	-1	0.3
x1	-0.3	-0.2	0.2
sp1	-3	-2	0
sp1 knot1	3	2	-1.5
sp1 knot2	3	2	-1

- Lambda (membership proportion)

Class 1	Class 2	Class 3
0.3	0.3	0.4

Knot configuration:

- 19 candidate knots for *sp1* (20-iles)

	Browsing		Writing	
	knot 1	knot 2	knot 1	knot 2
Class 1	8	14	4	11
Class 2	5	15	5	12
Class 3	5	13	7	14

Comparing Model Fits

Below is a comparison of fits achieved using the commercial package (Latent Gold) used for estimation in the paper vs. those using the purpose-built R code.

These tables suggest that both programs recover the true parameters accurately. Although the parameter estimates are very similar, minor differences in standard errors originate from the different calculation methods used. While LG calculates standard errors analytically (via the Newton-Raphson method), standard errors in the R implementation are calculated by parametric bootsrapping approximation (Efron and Tibshirani, 1993), using 100 samples.

Membership proportion (lambda)

	Parameter estimates			Standard errors	
Class	TRUE	LG	R	LG	R
1	0.300	0.326	0.326	0.016	0.018
2	0.300	0.284	0.284	0.016	0.015
3	0.400	0.390	0.390	0.017	0.017

Regression parameters (betab and betaw)

			Parameter estimates			Standard errors	
Stage	Covariates	Class	TRUE	LG	R	LG	R
browsing (betab)	Intercept	1	1.500	1.626	1.626	0.098	0.101
		2	2.000	2.041	2.041	0.079	0.087
		3	0.300	0.215	0.215	0.069	0.076
	x1	1	-0.200	-0.258	-0.258	0.061	0.068
		2	-0.100	-0.081	-0.081	0.024	0.025
		3	0.600	0.559	0.558	0.026	0.026
	sp1	1	-2.500	-2.430	-2.430	0.121	0.122
		2	0.500	0.535	0.535	0.056	0.064
		3	1.000	0.921	0.921	0.049	0.052
	sp1_knot1	1	3.000	2.872	2.872	0.168	0.165
		2	-1.000	-1.057	-1.057	0.070	0.080
		3	1.000	1.090	1.090	0.070	0.076
	sp1_knot2	1	2.000	2.136	2.136	0.230	0.236
		2	-0.500	-0.422	-0.421	0.064	0.058
		3	-1.500	-1.524	-1.524	0.086	0.083

			Parameter estimates			Standard errors	
writing (betaw)	Intercept	1	-2.000	-2.114	-2.114	0.187	0.240
		2	-1.000	-1.120	-1.120	0.117	0.113
		3	0.300	0.341	0.341	0.054	0.043
	x1	1	-0.300	-0.255	-0.255	0.036	0.044
		2	-0.200	-0.207	-0.207	0.035	0.032
		3	0.200	0.186	0.186	0.030	0.026
	sp1	1	-3.000	-3.078	-3.078	0.138	0.180
		2	-2.000	-2.101	-2.101	0.092	0.091
		3	0.000	0.020	0.020	0.046	0.033
	sp1_knot1	1	3.000	3.043	3.043	0.159	0.202
		2	2.000	2.105	2.105	0.116	0.114
		3	-1.500	-1.520	-1.520	0.070	0.054
	sp1_knot2	1	3.000	3.095	3.095	0.113	0.105
		2	2.000	2.097	2.098	0.102	0.099
		3	-1.000	-1.166	-1.166	0.113	0.110

Comparison of knot configurations and One-knot-deviated log-likelihoods

This table lists the LLs of all configurations when one of the knots is moved (either increased or decreased by one). The LLs from LG are identical to those from R. Baseline (true knot configuration) yields the best (i.e., closest to zero) LL, suggesting that the true knot configuration can be recovered well even for larger data sets, given sufficient run time.

True knot configuration

	Browsing		Writing	
	knot 1	knot 2	knot 1	knot 2
Class 1	8	14	4	11
Class 2	5	15	5	12
Class 3	5	13	7	14

Baseline LL at the true knot configuration: -95678.15

LL of one-knot-deviated configurations

			1 down		1 up	
Class	Stage	Knot	LG	R	LG	R
1	browsing	1	-95697.4	-95697.4	-95681.9	-95681.9
		2	-95681.2	-95681.2	-95685.8	-95685.8
	writing	1	-95708.8	-95708.8	-95700.1	-95700.1
		2	-95715.7	-95715.7	-95716.1	-95716.1
2	browsing	1	-95685.8	-95685.8	-95682.9	-95682.9
		2	-95679.1	-95679.1	-95681.2	-95681.2
	writing	1	-95690.4	-95690.4	-95693.8	-95693.8
		2	-95694.7	-95694.7	-95688.8	-95688.8
3	browsing	1	-95683.4	-95683.4	-95691.6	-95691.6
		2	-95681.8	-95681.8	-95686.2	-95686.2
	writing	1	-95686.7	-95686.7	-95691.5	-95691.5
		2	-95686.5	-95686.5	-95680.1	-95680.1

S6. REFERENCES

Harrell, F. 2001. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Schwartz, C. 2010. "Pathways to Educational homogamy in Marital and Cohabiting Unions." *Demography* 47:735-53.

List of Figures

FIGURE S-1a: Trilinear Spline Function

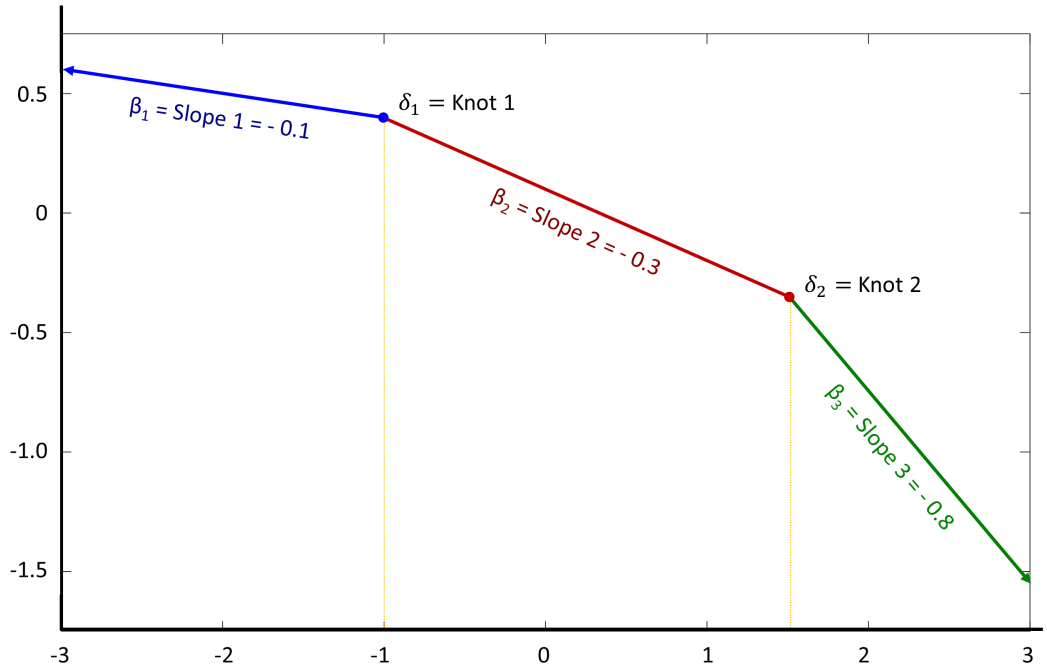


FIGURE S-1b: Trilinear Spline Basis Function

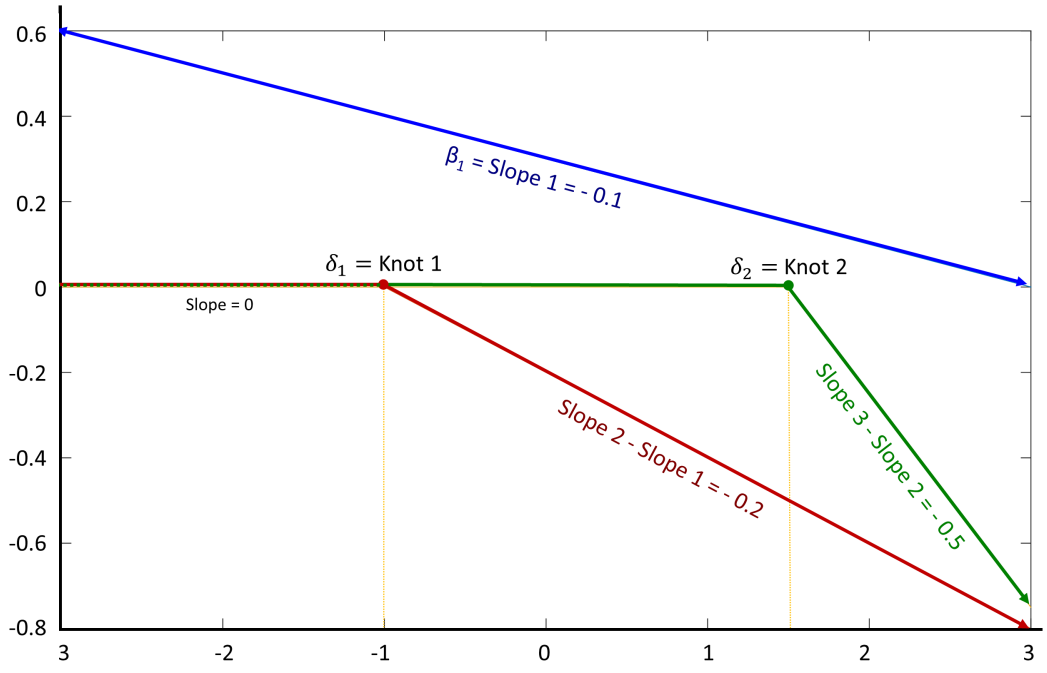
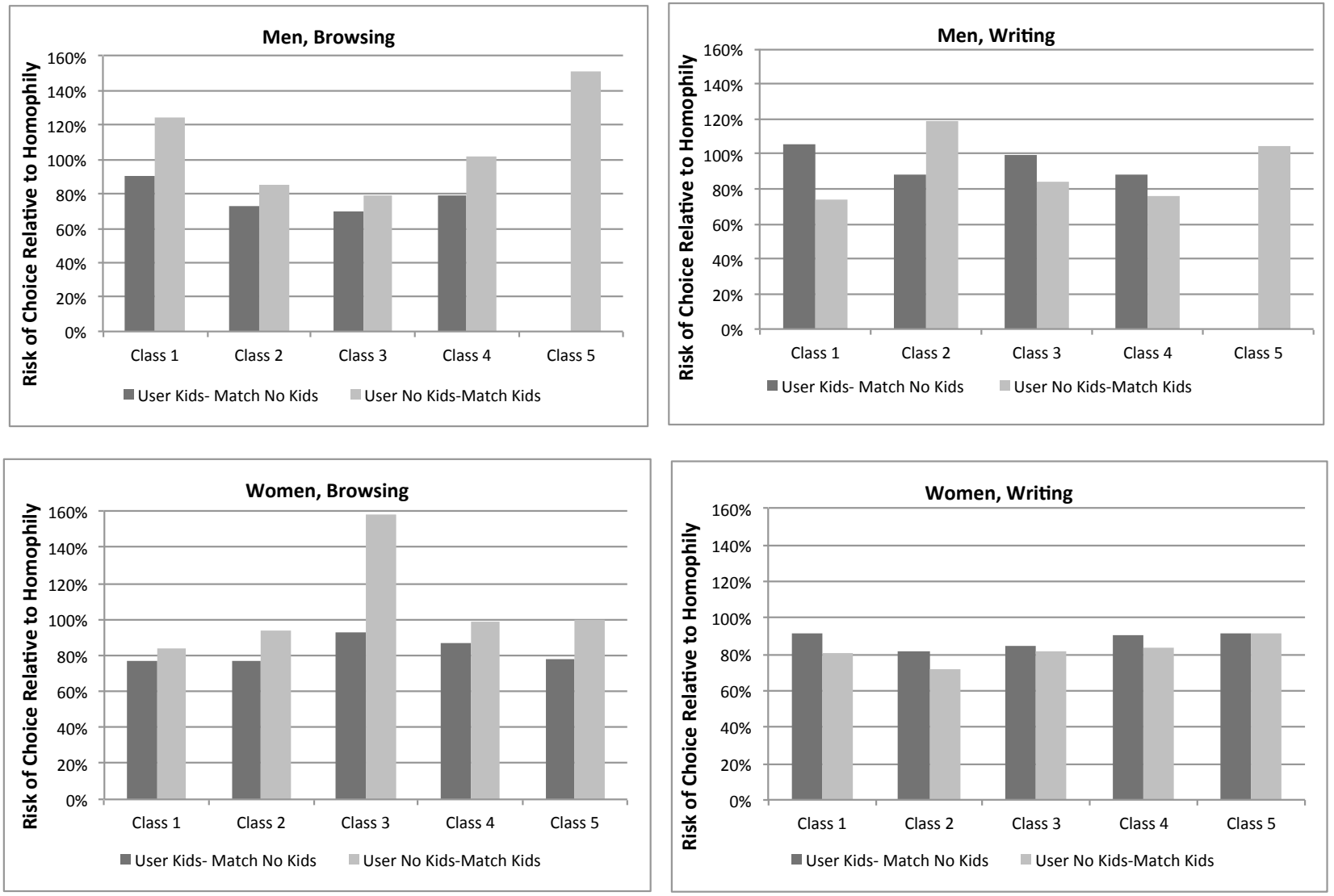
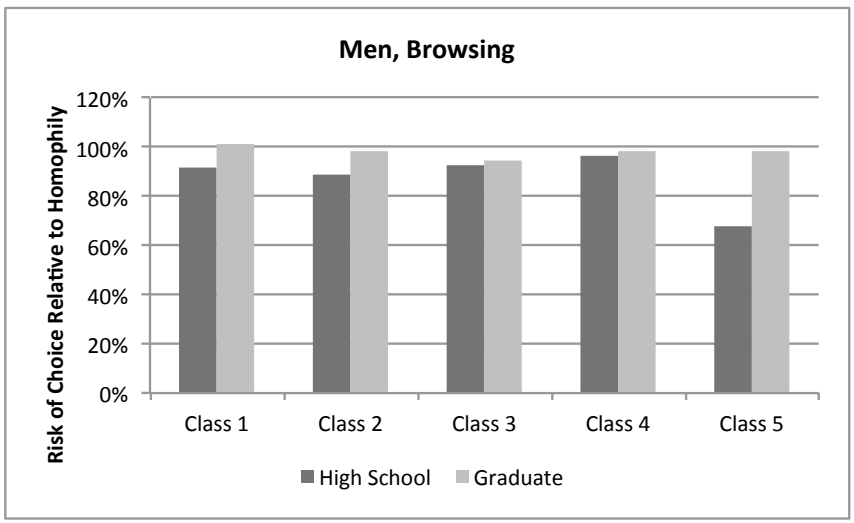


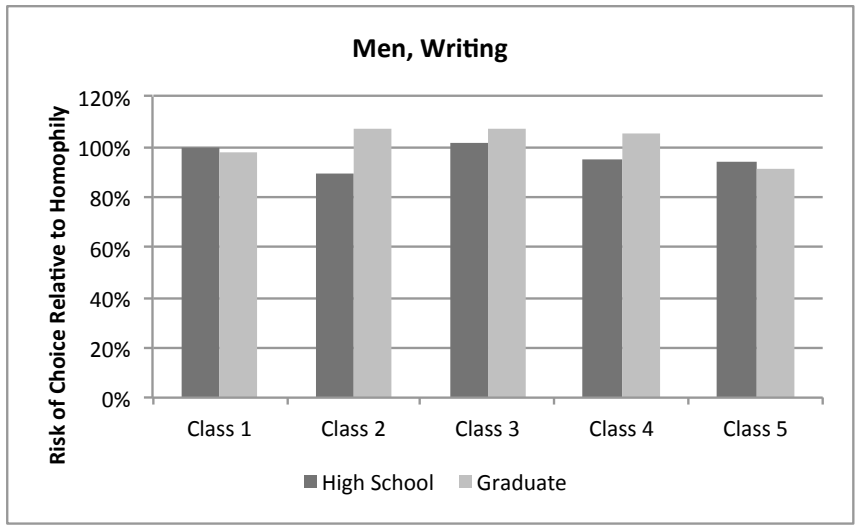
FIGURE S-2. Probability of Browsing and Writing, by Child Status[1]

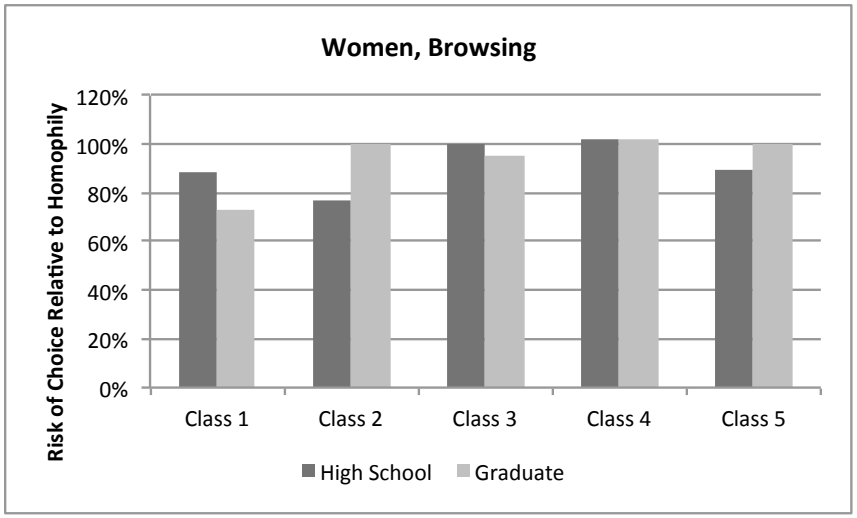


[1] Value for Kids-No Kids on for men in Class 5 omitted, as there are no men with children in this class.

FIGURE S-3. Probability of Browsing and Writing by Education, College Educated Respondents







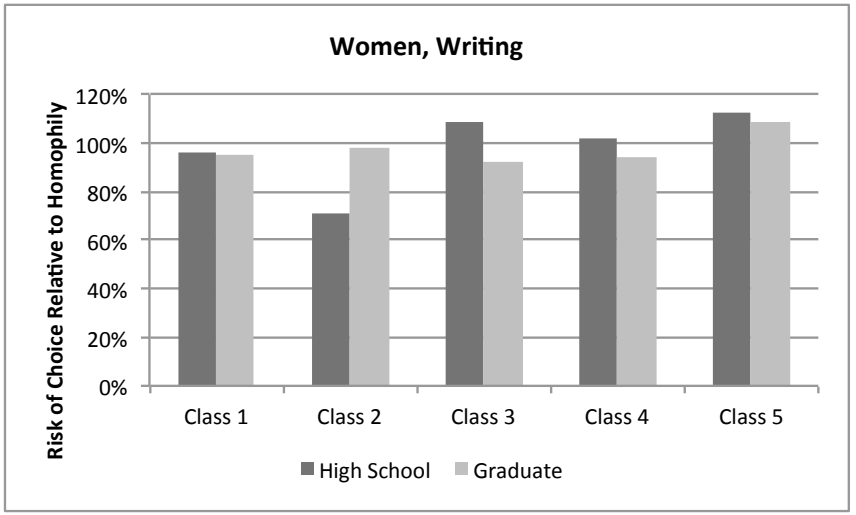


FIGURE S-4. Probability of Browsing and Writing by Education, Post-College Educated Respondents

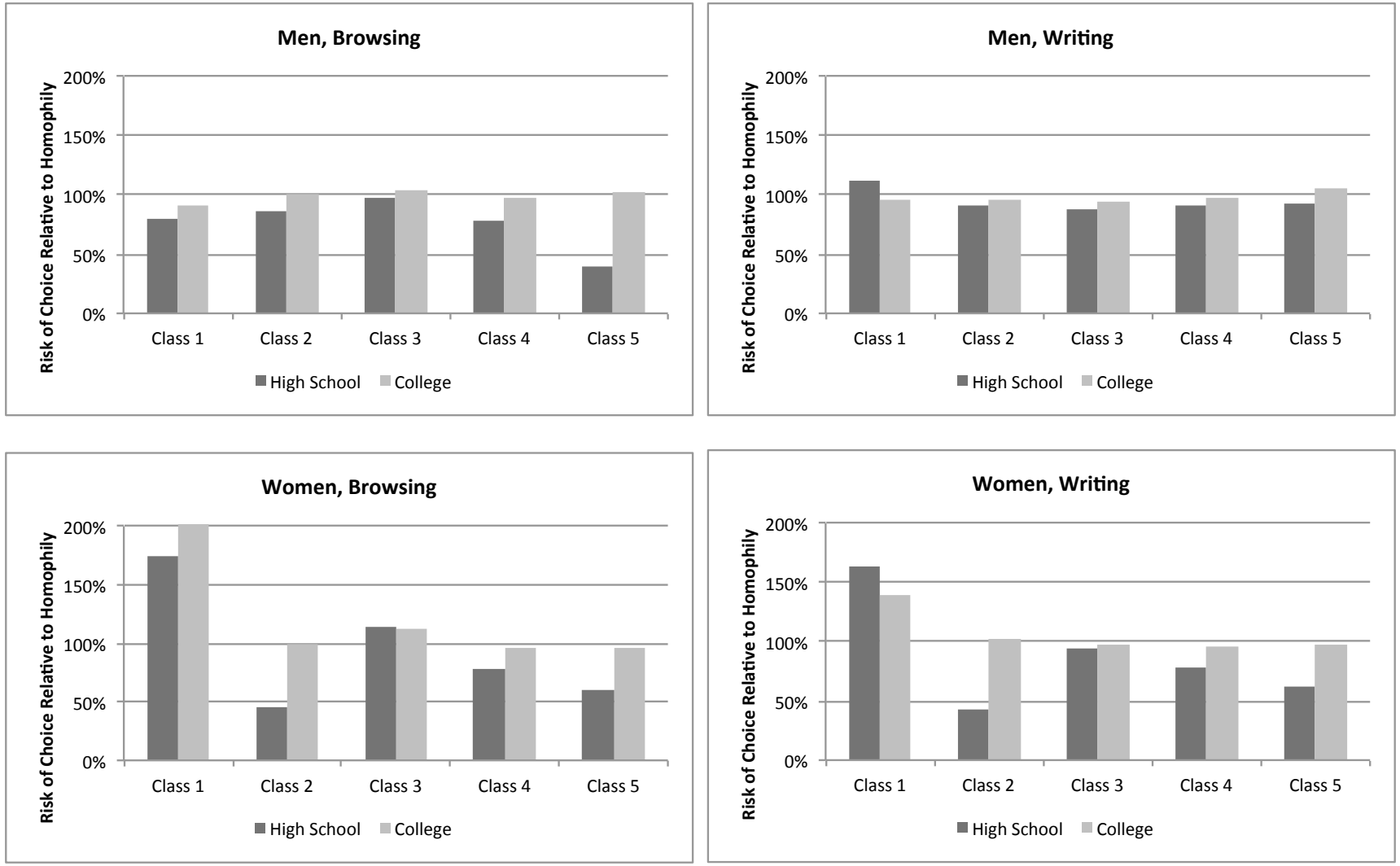
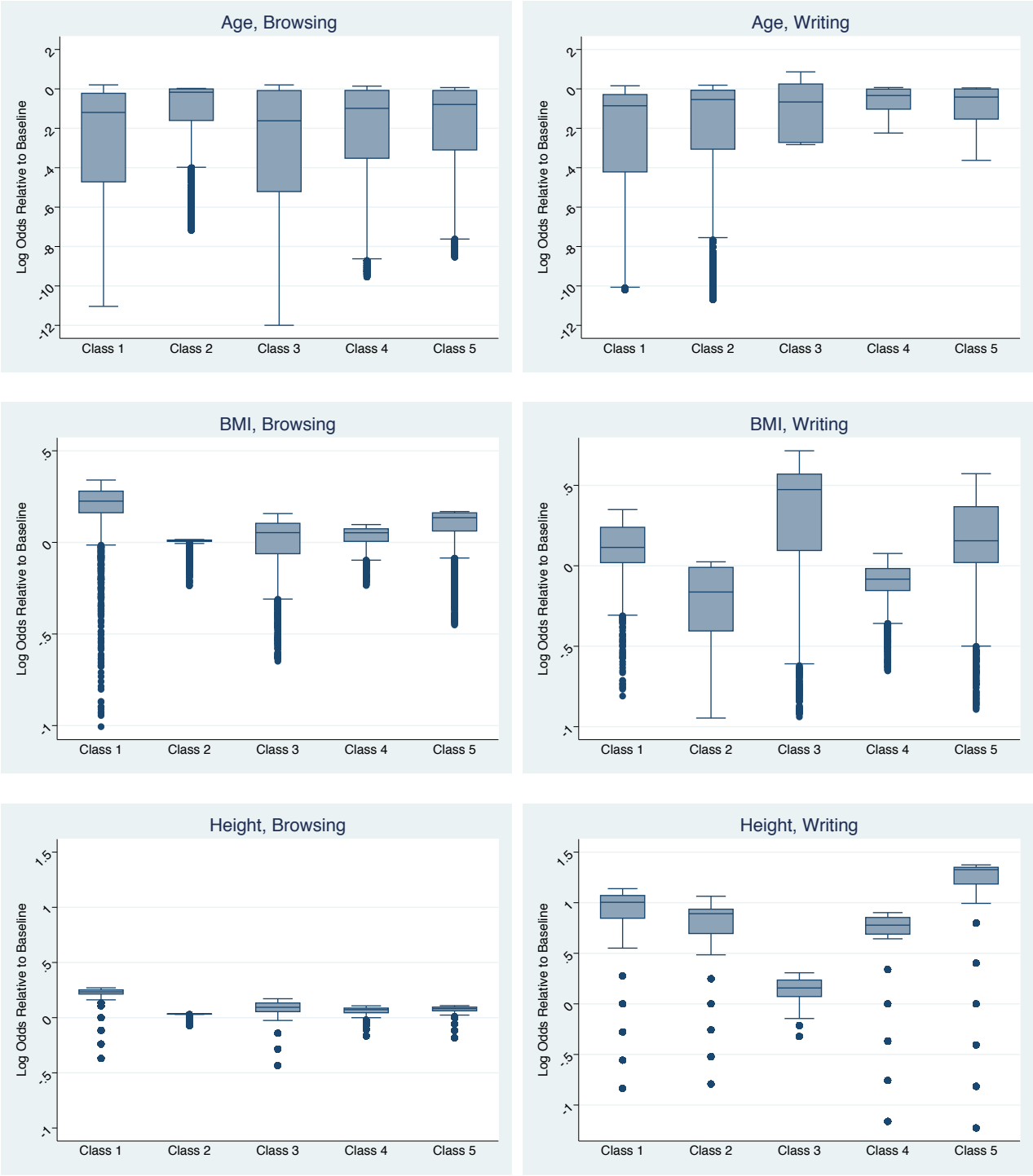
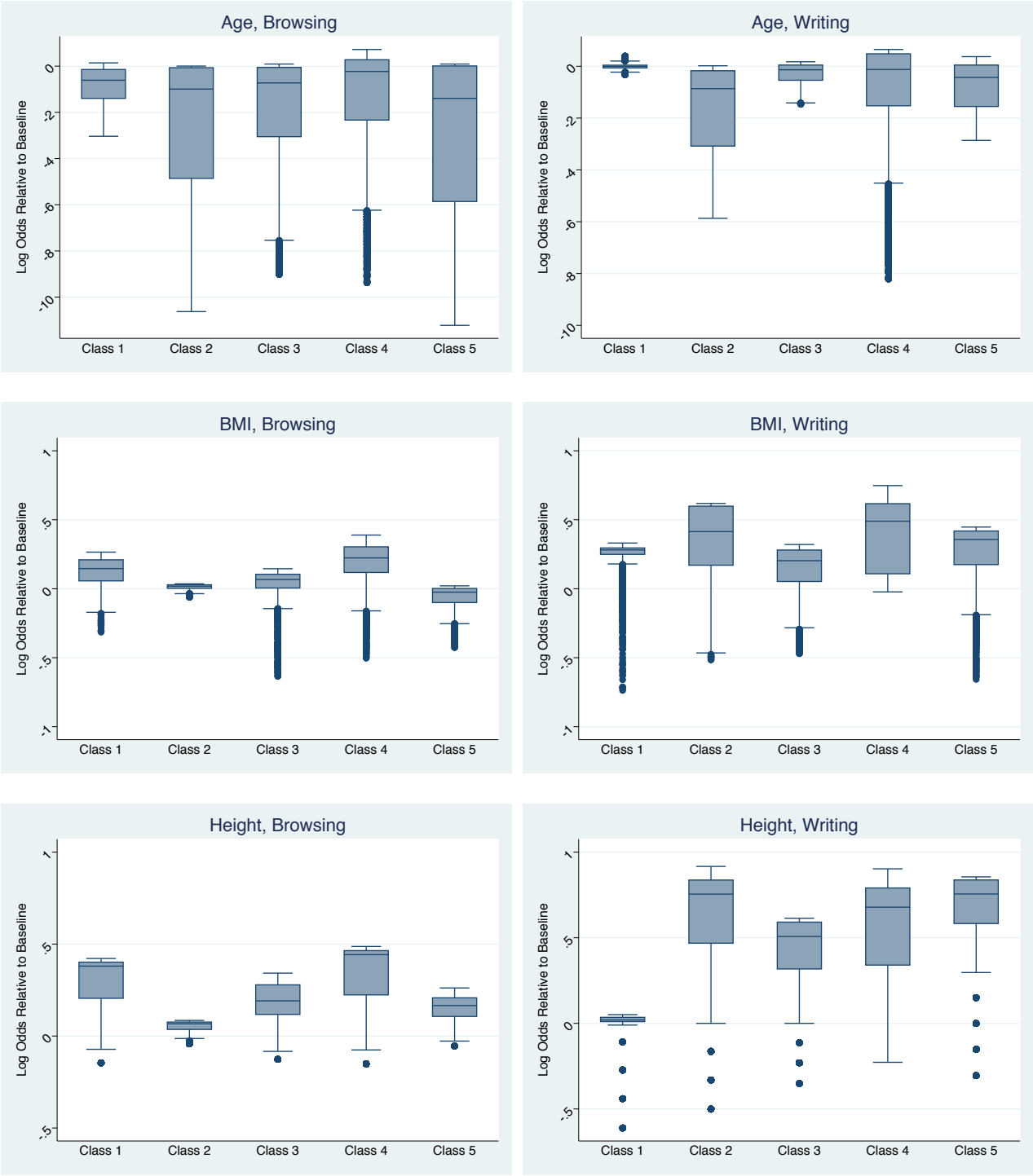


FIGURE S-5: Effect of Continuous Covariates on Log-Odds of Browsing & Writing, Women



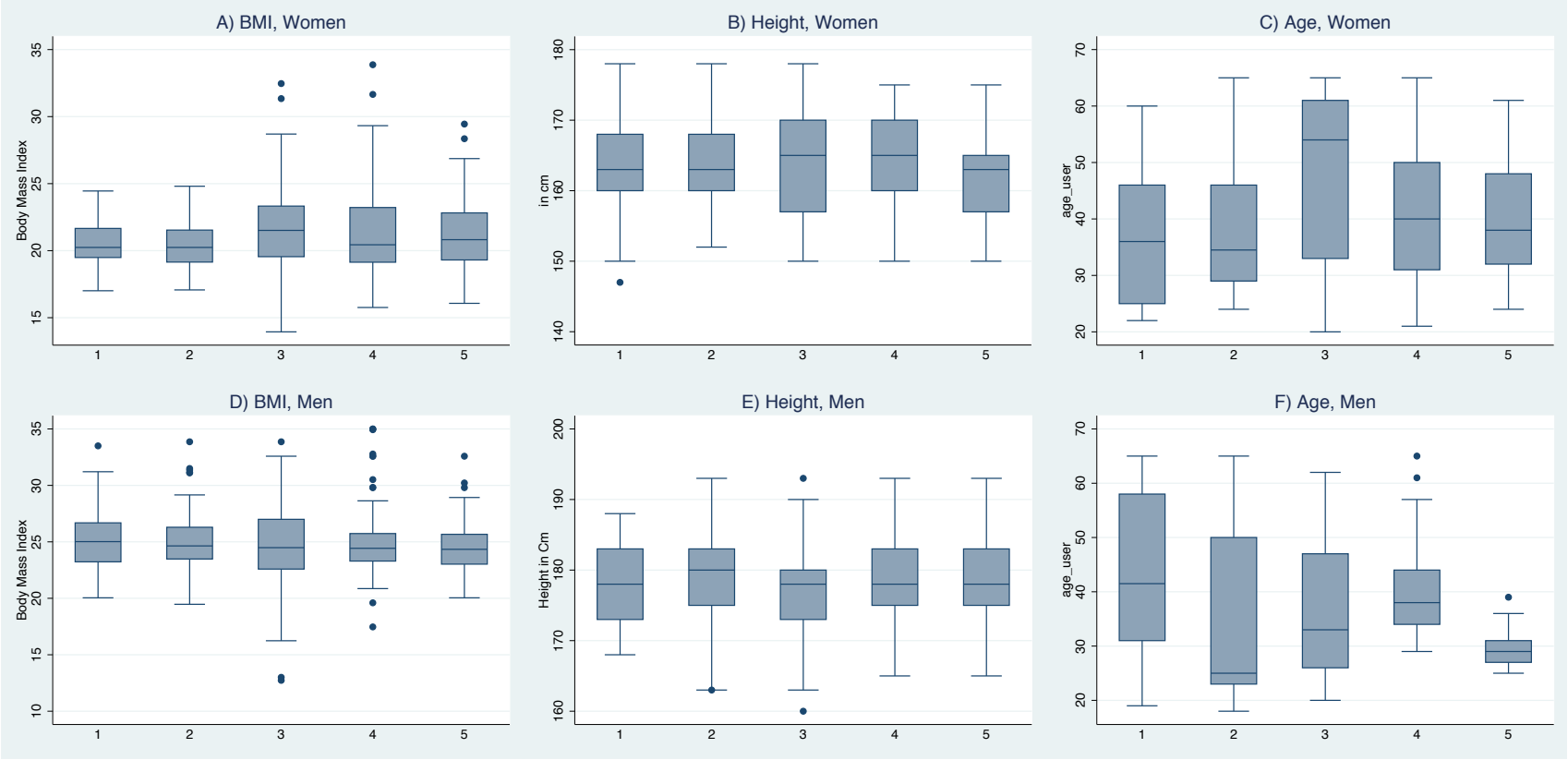
Note: Y-Axis denotes log-odds relative to a "baseline" of homophily; that is, the user and the potential match share the same value for an attribute.

FIGURE S-6: Effect of Continuous Covariates on Log-odds of Browsing & Writing, Men



Note: Y-Axis denotes log-odds relative to a "baseline" of homophily; that is, the user and the potential match share the same value for an attribute.

FIGURE S-7. Range of Continuous Covariates within Latent Classes



Note: Box plots are based on modal class assignment. The box reflects the lower and upper quartiles of each variable, with the median shown in a darkened horizontal line. The so-called "whiskers" are drawn to span 1.5 of the upper and lower interquartile range. Individual outliers beyond these bounds are represented as dots.

FIGURE S-8a. Effect of Age, Height, and BMI on Log Odds of Browsing and Writing, Conventional Models for Women

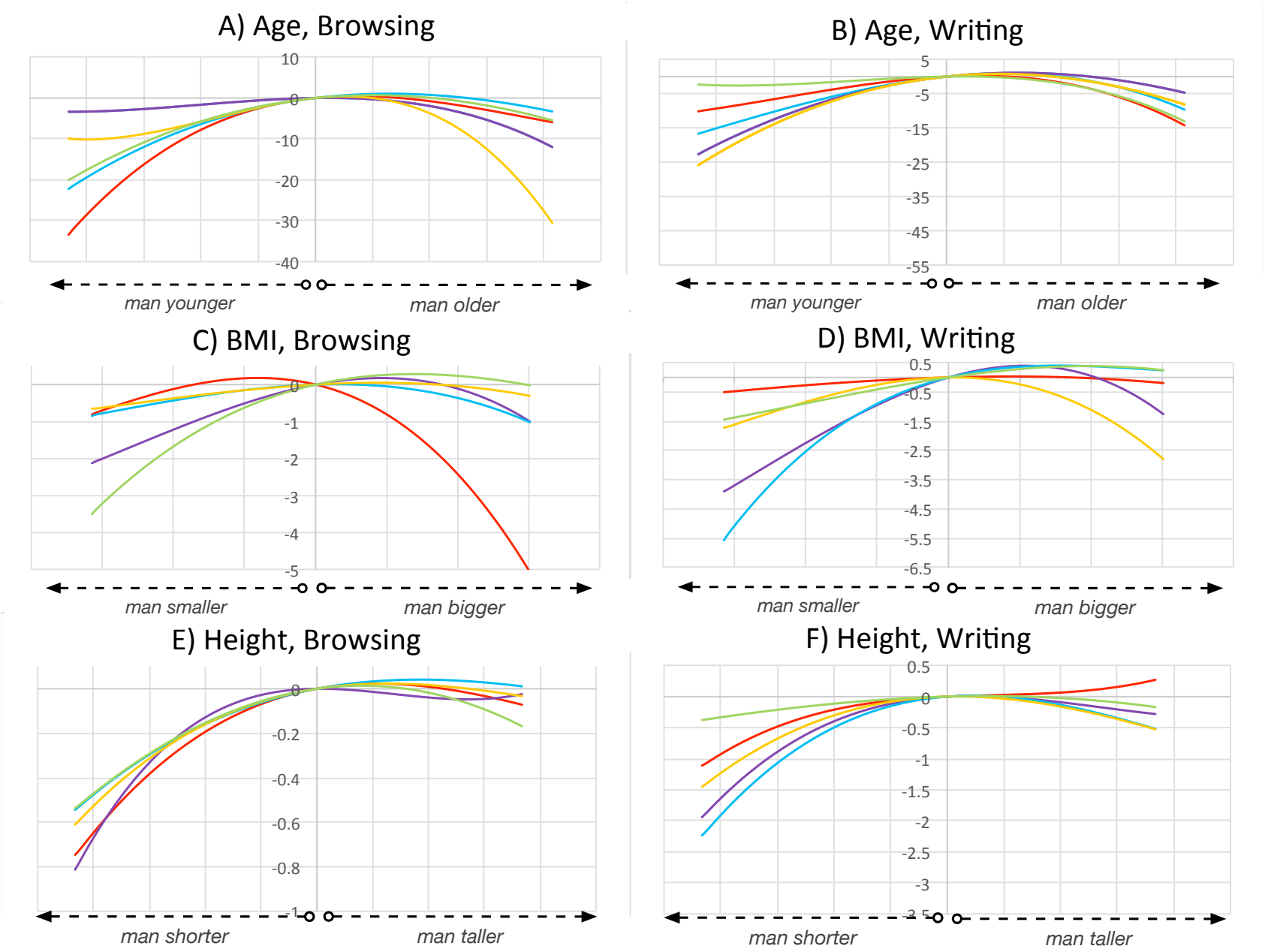


FIGURE S-8b. Effect of Age, Height, and BMI on Log Odds of Browsing and Writing, Multistage Spline Models for Women

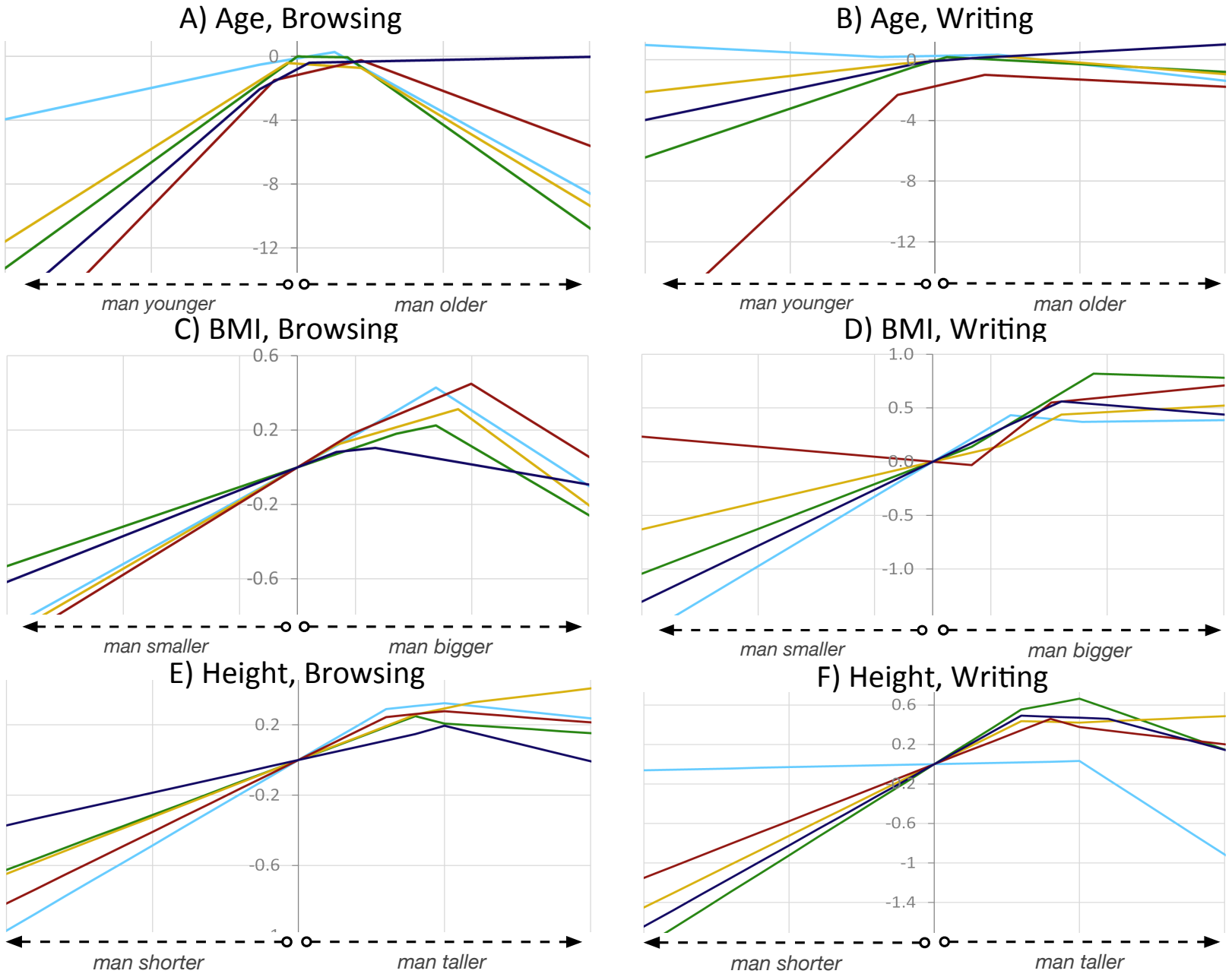


FIGURE S-8c. Effect of Age, Height, and BMI on Log Odds of Browsing and Writing, Conventional Models for Men

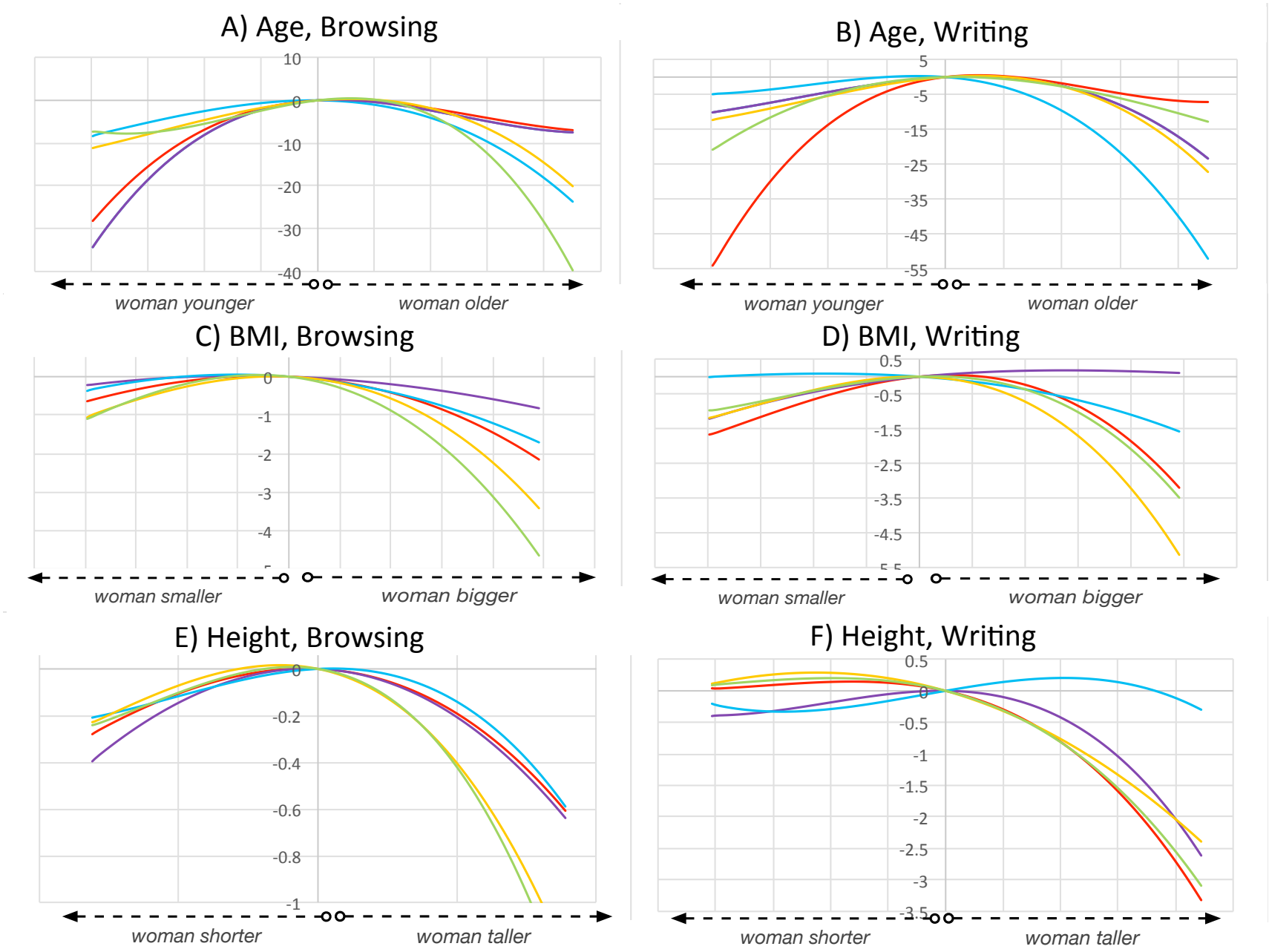
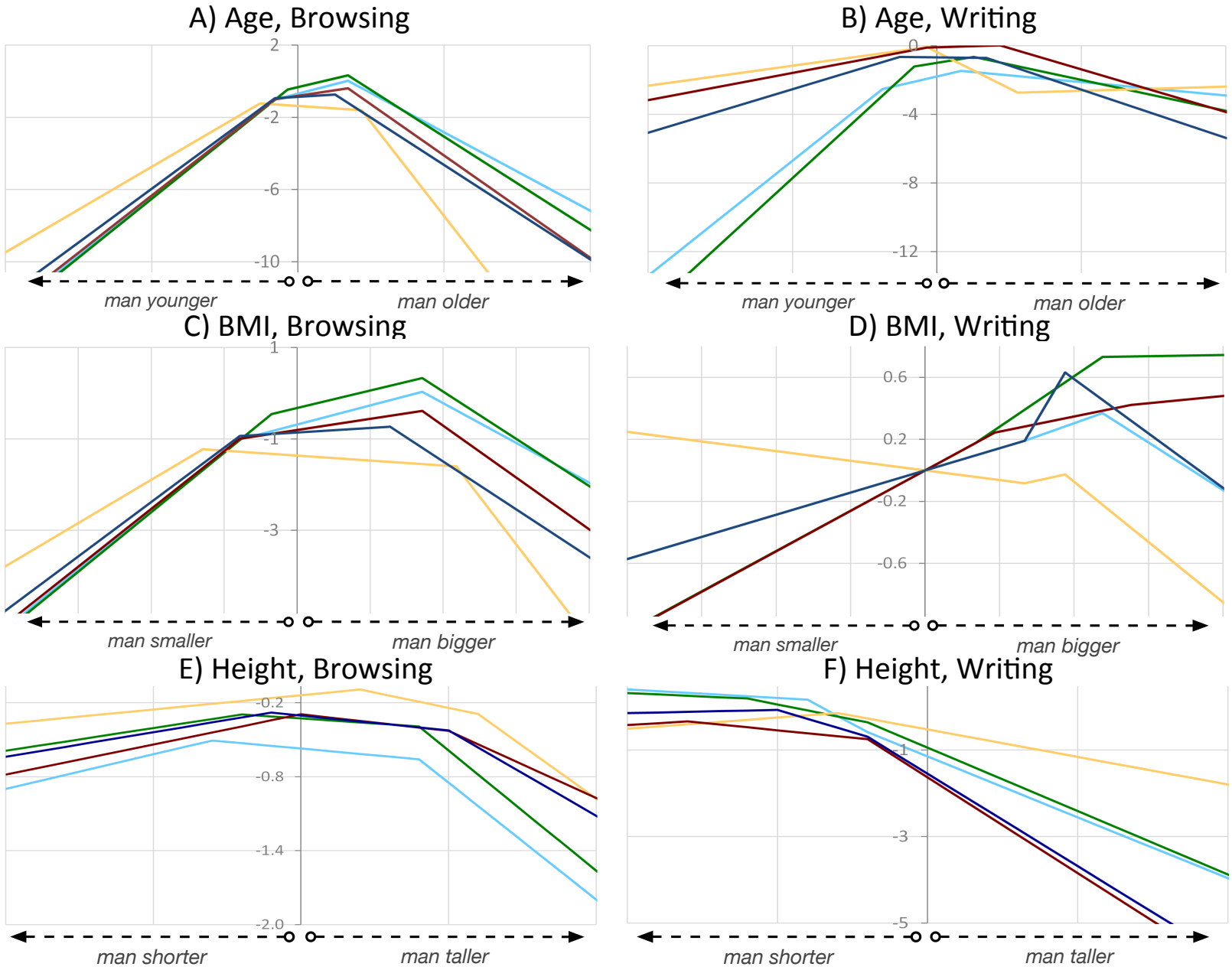


FIGURE S-8d. Effect of Age, Height, and BMI on Log Odds of Browsing and Writing, Multistage Spline Models for Men



List of Tables

Table S-1: Descriptive Statistics for Online Dating Population

	Men		Women	
Demographics	Mean	s.e.	Mean	s.e.
Age	34.93	11.07	34.7	11.81
Height	70.2	2.39	64.45	2.48
BMI	24.82	2.84	21.12	3.39
Photo (% with)	91		91	
Never Married (%)	74		70	
Kids (%)	24		28	
Smoker (%)	10		8	
Education (%)				
High School or Less	15		16	
College	48		48	
Graduate Degree	37		37	
Site Behavior				
Number Browsed	121.54	126.86	134.83	125.29
Number Written	21.4	30.89	11.96	16.86
N =	696		1159	

Table S-2a: Summary Statistics for Women by Modal Class Assignment

	Class				
	1	2	3	4	5
Profiles Browsed	156.12	151	98.19	95.33	145.78
Messages Sent	8.27	12	12.04	20.18	6.83
Age	29.89	33.92	40.1	36.57	36.02
Height (inches)	64.39	64.17	64.53	65.07	64.27
Smoker (%)	12	0	10	18	2
Kids (%)	22	20	38	35	33
Photo (% with)	93	93	91	88	89
BMI	21.39	20.47	21.96	21.36	21.11
Education (%)					
High School or Less	32	1	38	24	4
College	65	48	36	44	40
Graduate Degree	3	51	26	32	56
Never Married (%)	77	78	60	62	62
N	217	371	111	234	226
	23%	39%	12%	25%	24%

Table S-2b: Summary Statistics for Men by Modal Class Assignment

	Class				
	1	2	3	4	5
Profiles Browsed	174.02	109.83	73.36	172.46	111.14
Messages Sent	25.96	10.77	32.11	25.06	17.66
Age	42	34.88	35.4	39.17	28.63
Height (inches)	70.08	70.22	69.82	70.36	70.38
Smoker (%)	5	7	16	13	8
Kids (%)	22	36	34	28	0
Photo (% with)	95	91	86	89	94
BMI	25.19	24.9	24.75	24.98	24.53
Education (%)					
High School or Less	19	17	27	16	2
College	42	53	46	46	49
Graduate Degree	39	30	27	38	49
Never Married (%)	76	59	65	66	100
N	59	162	146	151	178
	8%	23%	21%	22%	26%

Table S-3. Coefficient Estimates From Latent Class Models, Continuous Attributes

	Men					Women				
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 1	Class 2	Class 3	Class 4	Class 5
Browsing Stage										
Height Difference										
	0.097***	0.063***	0.064***	0.082***	0.037***	0.056***	0.037***	0.021***	0.049***	0.040***
Slope1	(0.017)	(0.009)	(0.010)	(0.009)	(0.008)	(0.009)	(0.005)	(0.007)	(0.007)	(0.006)
	-0.080***	-0.104**	-0.030	-0.064***	0.011	-0.078***	-0.053***	-0.073***	-0.076***	-0.064***
Slope2	(0.024)	(0.042)	(0.019)	(0.016)	(0.037)	(0.012)	(0.008)	(0.021)	(0.012)	(0.009)
	-0.034*	0.031	-0.015	-0.030**	-0.088**	-0.168***	-0.179***	-0.122**	-0.082***	-0.124***
Slope3	(0.019)	(0.040)	(0.028)	(0.013)	(0.036)	(0.016)	(0.016)	(0.013)	(0.018)	(0.020)
Cutpoint1	6	8	7	6	8	3	4	8	6	5
Cutpoint2	10	9	11	9	9	10	10	12	11	11
Logged BMI Difference										
	1.735***	1.063***	1.819***	1.927***	1.234***	2.109***	4.029***	2.41**	3.945***	1.632***
Slope1	(0.375)	(0.250)	(0.458)	(0.241)	(0.271)	(0.3833)	(0.4344)	(0.927)	(0.615)	(0.292)
	0.112	-0.406	-0.904	-0.614*	-0.867	-4.314***	-4.562***	-1.391	-4.293***	-1.723**
Slope2	(0.697)	(0.948)	(0.671)	(0.363)	(0.599)	(0.527)	(0.503)	(1.116)	(0.768)	(0.497)
	-3.859***	-2.492**	-3.210***	-3.253***	-2.655***	-3.232***	-5.385***	-3.605***	-0.687*	-3.034***
Slope3	(0.700)	(1.042)	(0.611)	(0.364)	(0.727)	(0.702)	(0.568)	(0.556)	(0.414)	(0.480)
Cutpoint1	6	10	5	6	8	5	4	3	3	6
Cutpoint2	14	14	16	17	15	17	18	15	14	14
Logged Age Difference										
	3.930***	13.257***	11.584***	18.867***	15.845***	12.877***	13.023***	9.473***	12.655***	11.905***
Slope1	(0.156)	(0.180)	(0.224)	(0.475)	(0.382)	(0.183)	(0.171)	(0.205)	(0.240)	(0.186)
	-0.849**	-13.520***	-12.797***	-14.649***	-6.079***	-8.698***	-9.207***	-10.555***	-10.218***	-10.917***
Slope2	(0.296)	(0.356)	(0.356)	(0.524)	(0.516)	(0.262)	(0.242)	(0.351)	(0.331)	(0.285)
	-13.209***	-12.644***	-9.812***	-11.056***	-9.383***	-12.866***	-14.150***	-19.701***	-13.770***	-11.456***
Slope3	(0.351)	(0.465)	(0.522)	(0.195)	(0.332)	(0.246)	(0.209)	(0.715)	(0.291)	(0.286)
Cutpoint1	8	11	10	9	8	9	10	8	9	9
Cutpoint2	14	15	16	16	12	15	15	16	15	14

Table S-3 (continued). Coefficient Estimates From Latent Class Models, Continuous Attributes

Writing Stage										
Height Difference										
	0.006	0.1844***	0.145***	0.115***	0.164***	-0.040**	-0.032***	0.051**	0.043**	0.015
Slope1	(0.028)	(0.056)	(0.022)	(0.017)	(0.030)	(0.012)	(0.009)	(0.024)	(0.018)	(0.015)
	0.004	-0.128**	-0.151***	-0.198**	-0.175***	-0.333***	-0.106***	0.019	-0.111***	-0.222***
Slope2	(0.052)	(0.065)	(0.039)	(0.066)	(0.038)	(0.070)	(0.023)	(0.053)	(0.024)	(0.050)
	-0.200***	-0.159***	0.019	0.047	-0.067*	0.092	-0.155**	-0.180**	-0.374***	-0.220*
Slope3	(0.051)	(0.035)	(0.033)	(0.064)	(0.031)	(0.153)	(0.074)	(0.065)	(0.042)	(0.132)
Cutpoint1	7	5	6	8	6	8	6	6	4	7
Cutpoint2	10	10	9	9	11	10	10	9	10	10
Logged BMI Difference										
	3.241***	2.085*	1.259**	-0.467	1.606***	1.430	2.599	-0.622	2.609**	1.429
Slope1	(0.942)	(1.278)	(0.644)	(0.756)	(0.806)	(1.238)	(1.697)	(0.930)	(0.809)	(0.987)
	3.747*	1.165	1.516	4.718***	-0.169	0.273	0.656	1.670	-1.634	6.680**
Slope2	(1.767)	(1.643)	(1.355)	(1.171)	(1.769)	(2.578)	(2.077)	(3.406)	(1.162)	(3.205)
	0.572	-3.419**	-2.475**	-3.717***	2.874*	-4.769	-3.178**	-4.462	-0.509	11.616***
Slope3	(1.823)	(1.314)	(1.743)	(0.832)	(1.743)	(3.025)	(1.289)	(3.167)	(0.878)	(3.056)
Cutpoint1	8	5	7	5	8	8	5	8	6	8
Cutpoint2	15	16	13	12	13	14	14	11	16	11
Logged Age Difference										
	-0.977**	6.433***	2.131**	17.904**	3.965***	13.400**	15.422***	2.332***	3.171***	5.070***
Slope1	(0.387)	(1.136)	(0.750)	(8.927)	(0.692)	(4.636)	(2.163)	(0.411)	(0.476)	(1.458)
	1.347**	-0.705	-0.655	-13.502	-2.456**	-9.506**	-12.702***	-10.777***	-2.661***	-5.265**
Slope2	(0.569)	(1.651)	(0.855)	(8.997)	(0.970)	(4.774)	(2.293)	(0.875)	(0.631)	(1.623)
	-2.562**	-6.744***	-2.987***	-5.349***	-0.475	-5.459***	-6.319***	8.933***	5.474***	-5.426***
Slope3	(1.122)	(1.066)	(0.859)	(0.507)	(0.793)	(0.776)	(0.619)	(2.539)	(0.812)	(1.431)
Cutpoint1	7	9	8	8	10	7	9	10	10	8
Cutpoint2	16	12	16	15	14	13	14	17	16	15
N			696					1159		
Class Size	0.091	0.234	0.196	0.225	0.254	0.192	0.31	0.098	0.198	0.202

Note: Standard deviations of estimates shown in parentheses. Model generates both estimates of places (knots) where changes in evaluations of utility occur as well as slopes at those changepoints. Specific spline knots where slope changes occur are identified as follows (note that by design, all contain the first segment). See Appendix A text for more discussion of spline bases and interpretation of cutpoints.

Table S-4. Coefficient Estimates From Latent Class Models, Selected Categorical Attributes

	Men					Women				
BROWSING STAGE	Class 1	Class 2	Class 3	Class 4	Class 5	Class 1	Class 2	Class 3	Class 4	Class 5
Marital Status										
Single-Single	0.208***	1.660***	0.290***	-0.003	-0.612***	0.072*	-0.114***	-.135**	0.127**	-0.135***
	(0.037)	(0.054)	(0.048)	(0.024)	(0.067)	(0.041)	(0.023)	(0.057)	(0.046)	(0.027)
Divorced-Divorced	0.233***	0.203***	0.407***	0.415***	-1.392	0.282***	0.083**	0.415***	0.243***	0.189***
	(0.048)	(0.035)	(0.054)	(0.040)	(3.503)	(0.047)	(0.039)	(0.061)	(0.042)	(0.035)
Kids										
Kids-Kids	0.211***	0.792***	0.909***	0.547***	0.205	0.604***	0.600***	0.152**	0.317***	0.604***
	(0.050)	(0.036)	(0.049)	(0.038)	(3.632)	(0.044)	(0.040)	(0.059)	(0.042)	(0.036)
No Kids-No Kids	-0.392***	0.362***	0.529***	-0.027	-0.705***	0.604***	0.136***	-0.766***	0.025	0.004
	(0.037)	(0.048)	(0.050)	(0.025)	(0.068)	(0.044)	(0.031)	(0.066)	(0.047)	(0.028)
Education										
College-High School	-0.117***	-0.142***	-0.064**	0.037**	-0.484***	0.021	-0.337***	0.043	0.011	-0.156***
	(0.030)	(0.027)	(0.029)	(0.016)	(0.030)	(0.020)	(0.020)	(0.028)	(0.022)	(0.023)
College-College	0.045*	0.090***	0.087***	-0.001	0.263***	0.291***	0.161***	0.020	-0.025	0.079***
	(0.025)	(022)	(0.023)	(0.016)	(0.021)	(0.016)	(0.014)	(0.031)	(0.018)	(0.018)
College-PostCollege	0.071**	0.052**	-0.023	-0.300***	0.221***	-0.312***	0.176***	-0.062*	0.014	0.078***
	(0.027)	(0.022)	(0.029)	(0.022)	(0.019)	(0.028)	(0.012)	(0.034)	(0.019)	(0.015)
GradSchool-High School	-0.233***	-0.201***	-0.058*	0.120***	-1.090***	0.155***	-0.961***	0.108***	-0.289***	-0.630***
	(0.031)	(0.029)	(0.032)	(0.017)	(0.034)	(0.023)	(0.022)	(0.031)	(0.025)	(0.027)
GradSchool-College	0.023	0.113***	0.075**	0.180***	0.567***	0.583***	0.465***	0.040	0.099***	0.273***
	(0.026)	(0.024)	(0.026)	(0.017)	(0.024)	(0.020)	(0.014)	(0.035)	(0.022)	(0.020)
GradSchool-PostCollege	0.207***	0.089***	-0.018	0.180***	0.523***	-0.738***	0.496***	-0.148***	0.190***	0.357***
	(0.027)	(0.023)	(0.031)	(0.017)	(0.023)	(0.035)	(0.014)	(0.036)	(0.021)	(0.017)
Smoking										
Smoker-Smoker	0.197	-0.650**	-0.475***	-0.139*	0.208*	-0.218**	2.155***	-1.097***	-0.505***	0.095
	(0.268)	(0.247)	(0.109)	(0.084)	(0.110)	(0.082)	(0.436)	(0.171)	(0.085)	(0.189)
NonSmoker-NonSmoker	0.706***	0.784***	0.229***	-.391***	0.702***	0.608***	1.059***	0.696***	0.466***	0.848***
	(0.050)	(0.042)	(0.042)	(0.023)	(0.032)	(0.023)	(0.025)	(0.043)	(0.025)	(0.031)
No Photos	-2.590***	-2.739***	-2.666***	-2.724***	-2.682***	-2.850***	-2.862***	-2.945***	-2.808***	-2.690***
	(0.064)	(0.058)	(0.041)	(0.041)	(0.055)	(0.041)	(0.032)	(0.073)	(0.047)	(0.038)

Table S-4 (continued). Coefficient Estimates From Latent Class Models, Selected Categorical Attributes

WRITING STAGE										
Marital Status										
Single-Single	-0.357***	1.537***	0.370***	0.051	0.146	-0.535**	0.381***	-1.204***	0.123	0.037
	(0.082)	(0.151)	(0.080)	(0.053)	(0.159)	(0.162)	(0.096)	(0.120)	(0.088)	(0.105)
Divorced-Divorced	0.679***	-0.128	0.362***	0.156*	5.029	0.301**	0.424***	-0.339**	-0.046	0.391***
	(0.095)	(0.091)	(0.079)	(0.093)	(11.340)	(0.129)	(0.125)	(0.129)	(0.070)	(0.115)
Kids										
Kids-Kids	-0.105	0.266**	0.018	0.282***	-2.780	0.170	0.430***	0.368**	-0.210**	0.210*
	(0.104)	(0.092)	(0.071)	(0.088)	(11.451)	(0.130)	(0.130)	(0.129)	(0.070)	(0.113)
No Kids-No Kids	0.748***	-0.321**	0.379***	0.663***	-0.084	0.511**	0.869***	0.451**	0.413***	0.175
	(0.084)	(0.152)	(0.085)	(0.064)	(0.153)	(0.166)	(0.116)	(0.140)	(0.092)	(0.111)
Education										
College-High School	0.010	-0.200***	-0.022	-0.104**	-0.028	-0.012	0.415***	0.163**	0.058	0.086
	(0.067)	(0.060)	(0.044)	(0.046)	(0.064)	(0.066)	(0.087)	(0.060)	(0.038)	(0.094)
College-College	0.014	0.025	-0.060*	0.002	-.105**	0.067	0.220***	0.009	0.033	-0.123*
	(0.053)	(0.056)	(0.035)	(0.036)	(0.043)	(0.052)	(0.052)	(0.067)	(0.32)	(0.072)
College-PostCollege	-0.025	0.175***	0.082*	0.103**	-0.077*	-0.054	0.194***	-0.172**	-0.091**	0.037
	(0.060)	(0.052)	(0.044)	(0.037)	(0.042)	(0.087)	(0.049)	(0.074)	(0.033)	(0.061)
GradSchool-High School	-0.186**	-0.095	-0.136**	-0.113**	-0.134**	0.414***	-1.009***	-0.063	-0.292***	-0.594***
	(0.065)	(0.067)	(0.047)	(0.049)	(0.069)	(0.065)	(0.086)	(0.068)	(0.044)	(0.093)
GradSchool-College	-0.135**	0.003	0.013	0.028	0.125**	0.090	0.523***	-0.014	0.108**	0.263***
	(0.055)	(0.055)	(0.038)	(0.038)	(0.047)	(0.060)	(0.051)	(0.074)	(0.036)	(0.066)
GradSchool-PostCollege	-0.051	0.091	0.122**	0.085**	0.009	-0.504***	0.480***	0.049	0.184***	0.332***
	(0.059)	(0.058)	(0.047)	(0.037)	(0.046)	(0.091)	(0.048)	(0.076)	(0.036)	(0.059)
Smoking										
Smoker-Smoker	1.038**	0.676	0.059	0.406*	0.785***	-0.052	0.970	-10.157	-0.008	0.849
	(0.449)	(0.512)	(0.182)	(0.210)	(0.181)	(0.313)	(1.060)	(99.520)	(0.182)	(0.744)
NonSmoker-NonSmoker	-0.118	0.312**	0.451***	0.503***	-0.024	0.100	0.480***	1.322***	0.561***	0.513***
	(0.120)	(0.108)	(0.055)	(0.062)	(0.069)	(0.091)	(0.102)	(0.139)	(0.056)	(0.163)
No Photos	0.359**	0.310*	0.185	0.235**	-0.123	-0.522**	0.143	-0.451*	-0.133	0.251
	(0.154)	(0.169)	(126)	(0.106)	(0.143)	(0.221)	(0.106)	(0..252)	(0.116)	(0.144)
N			696					1159		
Class Size	0.091	0.234	0.196	0.225	0.254	0.1889	0.3128	0.0993	0.1983	0.2007

Note: Standard errors of estimates in parentheses. In variables, first term refers to attributes of the user and the second term refers to attributes of the match. Large standard error on Smoker-Smoker for Class 3 in writing stage due to no observations at this stage.