Journal of the Royal Statistical Society Series A: Statistics in Society, 2023, **00**, 1–25
https://doi.org/10.1093/jrsssa/qnad031

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series A Statistics in Society

Original Article

A practical revealed preference model for separating preferences and availability effects in marriage formation

Shuchi Goyal[1], Mark S. Handcock[1], Heide M. Jackson[2], Michael S. Rendall[3] and Fiona C. Yeung[1]

[1]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA
[2]Maryland Population Research Center, University of Maryland, College Park, MD, USA
[3]Department of Sociology and Maryland Population Center, University of Maryland, College Park, MD, USA

Address for correspondence: Shuchi Goyal, Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA. Email: sgoyal25@ucla.edu

Abstract

Many demographic problems require models for partnership formation. We consider a model for matchings within a bipartite population where individuals have utility for people based on observed and unobserved characteristics. It represents both the availability of potential partners of different types and the preferences of individuals for such people. We develop an estimator for the preference parameters based on sample survey data on partnerships and population composition. We conduct simulation studies based on the Survey of Income and Program Participation showing that the estimator recovers preference parameters that are invariant under different population availabilities and has the correct confidence coverage.

Keywords: discrete choice, marriage markets, matching, survey on income and program participation, two-sided partnership

1 Introduction to the two-sided matching market

Many social processes of pair formation can be viewed as two-sided matching problems. These scenarios are prevalent in demography, economics, sociology, political science, and education, among other fields. For example, heterosexual marriages, job searching, and residency assignments for medical school graduates all require members of two disjoint groups to mutually consent to form a relationship or match. Yet, the underlying mechanisms which dictate such processes are often opaque. We consider not only how an actor chooses from a set of actors from the opposite side, but also the interactions between pairs of actors in a choice situation and the stability of the matching result. Actors from opposing sides have to choose each other voluntarily in order for a 'match' to occur. Of particular interest to many researchers is the role individual and societal preferences play in the match-making process.

These preferences are difficult to discern for multiple reasons. First, it is challenging to collect data which records complete information about characteristics of observed pairings and the pool of options from which each individual made a selection. Second, the final observed matchings are as much a result of the availability of different types of individuals as they are of individual preferences. For example, in the heterosexual marriage market, women may prefer men who are highly educated. However, a limit in the supply of men with this characteristic means that some women must either choose a partner with lower education levels or remain single. It is important to distinguish the effects of preferences from those of availability in the final matchings realized. This problem has long been recognized in demography and, as we will review in the

Received: November 12, 2020. **Revised:** December 8, 2022. **Accepted:** February 12, 2023

next section, has motivated an impressive body of literature without having been satisfactorily resolved (Choo & Siow, 2006; Dagsvik, 2000; Logan et al., 2008).

Menzel (2015) proves a series of new mathematical results related to the asymptotic distribution of matching outcomes in a two-sided market. In this paper, we develop Menzel's (2015) technical findings for applications in demographic studies of two-sided matching processes. We propose a *revealed preferences model* which, given an observed set of stable matchings in a large population, uses a re-parameterized version of Menzel's (2015) equations to recover latent preference parameters in the population. These preference parameters are used to estimate the total utility of a given partnership, given the characteristics of the individuals in that partnership. To measure uncertainty of parameter estimates, we also propose both an analytical and an empirical approach to compute confidence intervals. We conduct simulation studies to show that for realistic populations, the revealed preferences model reconstructs preference parameters that are invariant under different population availabilities. We also show that the proposed confidence intervals achieve appropriate coverage.

The revealed preferences model can be generalized for applications where an individual is permitted to have multiple relationships, as in the case of an employer and its employees (Yeung, 2019). However, for the purposes of this paper, we focus only on the simpler case in which individuals have at most one partner, also known as one-to-one matchings.

The paper is organized as follows: in Section 2, we provide background information on the general two-sided matching problem and review existing literature which addresses the challenges of identifying individual preferences in such settings. In Section 3, we detail the proposed revealed preferences model and introduce relevant mathematical notation. We also address how we overcome challenges in the identifiability of certain preference parameters. In Section 4, we discuss parameter inference using a surrogate likelihood approach which depends on the sampling process through which the data were obtained. We also describe methods of computing standard errors for parameter estimates and constructing confidence intervals. In Section 5, we demonstrate applications of the revealed preferences model. We provide details on three simulation studies in which we attempt to recover known preferences using our proposed method. We present the results of these simulation studies in Section 6 which demonstrate the model's accurate estimation of parameters. We conclude in Section 7 with a discussion regarding the implications of the results and examples of ways the revealed preferences model might be useful in other fields.

2 Background

In most social settings, relationships are constantly shifting over time. For example, marriages form and dissolve, employees join and leave firms, and students enroll in and drop out of schools. These complex movements are difficult to capture in any data set due to their continuous nature. To circumvent this problem in the context of marriages, we focus on newly formed partnerships in a given sample at a discrete point and assume that this organization of one-to-one matches is *stable*.

The concept of *stable matchings* has been previously explored in depth by economists and statisticians. Stability is achieved when no two individuals who are not currently partnered with each other exist such that both individuals would prefer each other over their current partner. Furthermore, no person in a partnership would prefer to be single over their current partner. Gale and Shapley (see Roth & Sotomayor, 1990) showed that in large populations, there are various stable matchings that can be realized. By assuming matching stability, we are able to assume that the observed data accurately reflect individual and societal preferences at that time point.

One approach to study two-sided matching scenarios is through the use of *two-sided discrete choice models*, so called because individuals in the population have a set of discrete options with which they can match. The goal of two-sided matching models is to obtain the frequencies for the different types of partnerships that can occur, where the partnership type is defined by the combination of observable characteristics of the individuals in the partnership (Dagsvik et al., 2001).

In general, discrete choice models statistically relate the choice decision to the decision maker's attributes and the attributes of the alternatives available. Game theorists and statisticians initially proposed discrete choice models to understand agent preferences in one-sided settings. In these scenarios, each individual has a set of discrete possible choices. Essentially, there is a 'chooser' and a 'chosen'. The agent in the role of chooser is the sole decision maker of their outcome,

although his decision may be affected by the decisions of other choosers around them. The one-sided discrete choice model estimates the utility the chooser would derive from every possible choice in his option set and assumes that agents make the utility-maximizing choice. The parameters of interest are the chooser's preferences.

However, the traditional one-sided discrete choice model is unsuitable for use in two-sided scenarios. First, as mentioned earlier, the option set of each agent is rarely observed completely. Second, the observed matchings in two-sided processes are no longer reflective of the preferences of a single individual, as both actors involved in the partnership must consent to the partnership. That is, rather than dividing the population into groups of 'choosers' and 'chosens', both individuals in the partnership are choosers of each other. Each member of the partnership aims to maximize his or her own utility, and preferences may not necessarily be reciprocal. For example, highly educated women may have a preference for highly educated men, but highly educated men may not have a preference for highly educated women.

Among others, Schoen (1981), Pollak (1986), and Pollard (1997) approached the two-sided matching problem to obtain the frequency distribution of match types. However, the methodologies they propose are limited in that they say little about the behavior of agents in the two-sided market. Thus, there is no apparent mechanism for detecting the underlying preferences which motivate the matchings.

In contrast, Logan et al. (2008), Dagsvik (2000), and Menzel (2015) all theorize two-sided versions of the discrete choice model which consider the role of both preference parameters and availability of partners in matching markets and propose methodology which can implicitly be used to estimate said preferences. Logan et al. (2008) propose a model for bipartite populations where each side has a distinct utility function for partnerships with agents on the opposing side. In the case of heterosexual marriages, all men have an identically defined deterministic component to their utility which depends on the man's own observed characteristics x and the characteristics of his partner z; similarly, all women have an identically defined deterministic component to their utility which depends on the woman's own observed characteristics z and the characteristics of her partner x. Here, $x \in \mathcal{X}$ and $z \in \mathcal{Z}$, where the sample spaces $\mathcal{X}$ and $\mathcal{Z}$ represent the set of possible types of men and women, respectively, and may be continuous or discrete. Unobserved characteristics are accounted for in the utility by including an individual fixed effect term for each actor. Logan et al. (2008) assume that an individual's unobserved option set within the local marriage market can be approximated by the observed sample distribution of characteristics.

Logan et al. (2008) show that their proposed method for small populations could theoretically be used to compute maximum-likelihood estimates (MLEs) of preference parameters. Rather than basing their inference on the true likelihood of the observed match being realized, they propose inferences based on the likelihood that the observed match is stable. For computation of these estimates, they propose Bayesian inference based on Markov chain Monte Carlo (MCMC).

The approach suggested by Logan et al. (2008) is limited in that the Bayesian inference works best for small populations. For example, the authors apply their method to make inferences about gender-based marital preferences using data from the National Survey of Families and Households (NSFH). With a sample containing 314 men and 360 women, they are able to compute parameter estimates for the two-sided model.

However, the method cannot be used with large sample data sets such as the Survey of Income and Program Participation (SIPP), where the number of people of each gender exceeds 16,000 or the American Community Survey (ACS), where the number of people of each gender exceeds 100,000. In such cases, the calculations required to update parameter estimates in each step of the MCMC process are extremely complex and often intractable. Additionally, when large populations with multiple stable matching solutions are studied, the posterior distribution of the parameters may have multiple maxima, thereby also rendering the parameters unidentifiable. Logan et al. (2008) also note limitations in parameter identifiability when certain parallel terms are included in the utility functions.

Dagsvik (2000) focuses the identification and estimation of preference parameters in a closely related two-sided matching market model. He proposes constructing aggregate supply and demand functions based on preferences on both sides of the matching market. When the asymptotic supply and demand functions are equal, they derive equilibrium equations for the number of partnerships achieved between individuals of specific types. These equations imply that availability of

partners and personal preferences are asymptotically separable in their relationship to the distribution of matching outcomes in a large population. This is a significant finding because, intuitively, the ability of people to achieve their preferred partnership outcome is constrained by the existence of partners. Dagsvik (2000) then shows that these equations can be manipulated to obtain point estimates of preference parameters. However, methodology for analytically computing standard errors for these estimates is not presented. In addition, the results only apply to discrete agent types.

Nevertheless, the insights by Dagsvik (2000) lay important groundwork for the work done by Menzel (2015). Specifically, Menzel (2015) proves that the relationships suggested by Dagsvik (2000) hold true for large populations. Menzel (2015) derives equations which establish a relationship between the preference parameters and availabilities of men and women of each type in the population and the limiting distribution of types of matches across the possible outcomes. These calculations prove that in a large population, the interdependency between availability and preferences can be accurately modeled, and therefore that preferences can be recovered independently of the population availability context. Menzel (2015) then proposes that the relationship he develops can be used to construct a likelihood function for observing a particular matching. His results also apply to continuous agent characteristics.

We develop the results of Menzel (2015) to derive reparametrized equations which allow asymptotically stable estimates of the proportions of single and partnered persons of each type in the population. We propose a subclass of two-sided discrete choice models which we refer to as *revealed preference models*. In this subclass of models we, like Logan et al. (2008), Dagsvik (2000), and Menzel (2015), focus on bipartite networks. Actors in the network are divided into two distinct groups. Edges, which represent partnerships, form only between members of opposing groups. Whereas Logan et al. (2008) assume that the full opportunity set of each actor is observed, we allow agents of different observed types to have different opportunity sets (Yeung, 2019). The goal of our study is to extend Menzel's (2015) findings to estimate a set of latent structural parameters that describes the decision-making behavior of a given population which led to the observed matching outcome. The difficulty of this problem is that the set of alternatives for each actor is not generally observed and determined endogenously in the market. Our proposed model utilizes key findings from Menzel (2015) about the limiting distribution of matches in a large population and applies them to estimate preference parameters based on an observed distribution of matches. We extend Menzel (2015) by developing a modification of his estimator that corrects for bias in small populations across a range of sample sizes and sample fractions.

Our study extends from the non-transferable utility assumption following Dagsvik (2000), Logan et al. (2008), and Menzel (2015). Variants of this model have been used to represent decision-making in a matching market that assumes transferable utility (TU) within partnerships, with two recent studies by Dupuy and Galichon (2014) and Chiappori et al. (2017) building on a TU framework developed by Choo and Siow (2006). We note here only the basic commonalities and differences between the TU model of Choo and Siow (2006) and the NTU model of Menzel (2015). The TU model is grounded in the economic theory of Becker's (1973, 1974) model of marriage. It requires the key assumption that the members of a couple engage in within-couple exchanges of utility-providing goods and services. Choo and Siow (2006) interpret these exchanges as determining '…each spouse's share of responsibilities within a marriage'. The major statistical modeling implication is that in a TU model, the choosing individual only considers the prospective match's observable characteristics (Chiappori, 2020). In contrast, within the NTU framework, there is no similar exchange of utility-providing goods and services, and the individual is influenced by the prospective match's observable (to the researcher) characteristics and the characteristics that are to the researcher unobservable. In the NTU case, increased availability leads to increased propensity to find a match.

3 Revealed preferences model

To facilitate our discussion of the revealed preferences model, we will discuss the problem within the context of heterosexual marriages within a two-sex population unless otherwise noted. In this setup, we consider a population with two distinct groups, and individuals are either male or female. At any given point in time, individuals have at most one partner of the opposite sex, and

they also have the outside option to remain single (unpartnered). Both the male and the female must agree to the partnership for that partnership, or 'marriage', to be observed.

Individuals evaluate their marital options using a utility function, which contains a deterministic and random component. Actors of the same gender are assumed to have deterministic components to their utility functions that depend on their own observed characteristics x and those of their potential partners, z. The random component of the utility function accounts for the fact that agents' characteristics are only partially observed. Agents choose the partner from available options who will maximize their own total utility. The latent parameters in the deterministic component of the utility function which govern this pair formation are commonly known as 'preference' parameters in the sense that they represent how actors would choose among different alternatives if given a choice (Logan, 1996a; Logan et al., 2008).

We consider a population with N_w women and N_m men, so that the total population size is $N = N_w + N_m$. N_h represents the number of households in the population, where a household is an entity consisting of either a single (unpartnered) man or woman or a partnered couple, so that $N_h \leq N$, and $N_h = N$ only when all individuals choose to remain single. Using the same notation introduced in Section 2, we observe a p-vector of covariates $x \in \mathcal{X}$ on the women and a q-vector of covariates $z \in \mathcal{Z}$ on the men. Let x_i and z_j denote the observed attributes of woman $i = 1, \ldots, N_w$ and man $j = 1, \ldots, N_m$, respectively. The equations in this section are written generally so that the elements of x and z may be continuous, discrete, or a combination of the two. For ease of presentation, however, in the simulation studies in Section 6 where we apply the revealed preferences model, we assume that x and z are discrete.

Actors may perceive potential partners differently based on their own characteristics. Thus, the perceived utility gained by partnering with a particular opposite-sex individual may differ from one decision maker to the next. However, all actors are assumed to choose the partner within their respective choice sets that maximizes utility. Given the utility-maximizing behavior of the decision makers, we define the utility gained by woman i with observed attributes x_i from partnering with man j with observed attributes z_j as

$$U_{ij} = \underbrace{U(x_i, z_j \mid \underset{\sim}{\theta}^W)}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{\eta_{ij}}_{\substack{\text{unobserved random} \\ \text{component}}}, \tag{1}$$

where $\underset{\sim}{\theta}^W$ is the set of parameters denoting the woman's preferences. The deterministic part of the utility functions depends on variables representing the respective types of women and men. Similarly, we define the utility gained by man j with observed attributes z_j from a partnership with woman i with observed attributes x_i as

$$V_{ji} = \underbrace{V(z_j, x_i \mid \underset{\sim}{\theta}^M)}_{\substack{\text{deterministic} \\ \text{component}}} + \underbrace{\zeta_{ji}}_{\substack{\text{unobserved random} \\ \text{component}}}, \tag{2}$$

where $\underset{\sim}{\theta}^M$ is the set of parameters representing men's preferences. From this point forth in the paper, we will use tilde below a Greek letter to refer to a vector.

Following Menzel (2015), we assume that unobserved random components of the utility functions as defined in Equations (1) and (2) are independently and identically distributed draws from a distribution in the domain of attraction of the extreme-value type-I (Gumbel) distribution. This domain includes Exponential, Gamma, Gaussian, Lognormal, and Weibull distributions. Here, we will focus on the Gumbel itself, but note our model and methods are generalizable.

3.1 Model specifications

Having introduced the general setup of a two-sided discrete choice model, we now go into detail about model forms for the deterministic and random utility components. We focus on the special case where the deterministic components of the utilities in (1) and (2) are additive linear functions;

however, other choices of utility functions can also be used (see Dagsvik, 1994 for inference of latent preferences under other choices of utility functions).

For additive linear utility functions, let

$$U(x_i, z_j \mid \underset{\sim}{\theta}^{W}) = \theta_{w0} + \sum_{k=1}^{K_w} \theta_{wk} X^k(x_i, z_j),$$
$$V(z_j, x_i \mid \underset{\sim}{\theta}^{M}) = \theta_{m0} + \sum_{k=1}^{K_m} \theta_{mk} Z^k(x_i, z_j), \tag{3}$$

where x_i and z_j are vectors measuring observed characteristics of woman i and man j, respectively. The woman's deterministic utility consists of an intercept term θ_{w0} and K_w functions $X^k(x_i, z_j)$ which represent utility that woman i derives from the partnership based on her perception of her own characteristics and the characteristics of man j. For example, $X^k(x_i, z_j)$ might be an indicator function that represents whether certain observed attributes are identical for the pair (i.e., the partnership is homogamous). The corresponding K_m functions for the man's side are denoted as $Z^k(x_i, z_j)$. Here, $\underset{\sim}{\theta}^{W} = [\theta_{w0}, \theta_{w1}, \ldots \theta_{wK_w}]^T$ and $\underset{\sim}{\theta}^{M} = [\theta_{m0}, \theta_{m1}, \ldots \theta_{mK_m}]^T$ are the preference parameters.

The random component of the utility model accounts for unobserved information about individuals in the data which may impact partnership choices. The random terms are assumed to be identically distributed draws from an extreme-value type-I (Gumbel) distribution.

We additionally define the random utility for the choice of remaining single as Menzel (2015) did, so that

$$U_{i0} = 0 + \max_{k=1,\ldots,N_m^{\delta}} \{\eta_{i0,k}\},$$
$$V_{j0} = 0 + \max_{k=1,\ldots,N_w^{\delta}} \{\zeta_{j0,k}\}, \tag{4}$$

for females and males, respectively.

The single household utility specification in Equation (4) implies that the deterministic component of the utility for an individual choosing to be unpartnered is 0. The non-deterministic component of the single utility function of females is defined as the maximum of N_m^{δ} independent draws of $\eta_{i,k}$, the Gumbel-domain-of-attraction distributed random term of the male partnered utility function presented in Equation (1). Similarly, the non-deterministic component of the single utility function for males is the maximum of N_w^{δ} independent draws of $\zeta_{j,k}$ from Equation (2). A interpretation for this formulation is that in a market of N_m men, woman i also considers N_m^{δ} outside latent non-market alternatives (and vice versa for men).

We focus on the case where $\eta_{i,k}$ and $\zeta_{i,k}$ are i.i.d. Gumbel. Since the maximum of N_m^{δ} i.i.d. Gumbel random variables is also Gumbel distributed with the location parameter increased by $\delta \log N_m$, the hyperparameter δ effectively sets the expected utility for an individual choosing to be unpartnered. We choose δ based on prior expectations of how the proportion individuals in the population who are single will change for different market sizes. For this model, we set $\delta = 1/2$. This specification ensures that the share of singles in the market is stable for different market sizes (Menzel, 2015, Assumption 2.2). Intuitively, increasing the value of δ will make the choice of remaining single more attractive in large populations, while decreasing the value of δ makes the single option less attractive.

3.2 Large-population approximation

Let $w(x)$ be the number of women in the population with characteristics x and $m(z)$ be the number of men in the population with characteristics z. For notational convenience, let $\bar{w}(x) = w(x)/N$ and $\bar{m}(x) = m(x)/N$.

Consider a population with utilities drawn from models (1), (2), (3), and (4). Then, the stable matching induces a probability distribution over the observed characteristics. Consider sampling

a random person from the population and their classification of matched or single. Let $f(x, *)$ and $f(*, z)$ be the probability that the person is an unmatched woman of type x and an unmatched man of type z, respectively. Let $f(x, z)$ be the probability the person is in a match between a woman of type x and a man of type z. Finally, let $\bar{f} = \{f(x, z), f(x, *), f(*, z)\}$, $x \in \mathcal{X}$, $z \in \mathcal{Z}$. Together, $\bar{f}$ defines a distribution satisfying the overall normalization constraint

$$\int f(x, z)\, \mathrm{d}x\, \mathrm{d}z + \int f(x, *)\, \mathrm{d}x + \int f(*, z)\, \mathrm{d}z = 1. \tag{5}$$

More specifically,

$$\begin{aligned} \bar{w}(x) &= f(x, *) + f(x, \diamond), \\ \bar{m}(z) &= f(*, z) + f(\diamond, z), \end{aligned} \tag{6}$$

where $f(x, \diamond)$ is the probability the person is a matched woman of type x,

$$\begin{aligned} f(x, \diamond) &= \int f(x, z)\, \mathrm{d}z, \\ f(\diamond, z) &= \int f(x, z)\, \mathrm{d}x. \end{aligned}$$

A major result of Menzel (2015) is that, under mild regularity conditions, if the population size is large and the matching is stable, the frequencies approximately satisfy the relations

$$f(x, z) = 2\, \mathrm{e}^{W(x,z \mid \underset{\sim}{\beta})} f(x, *) f(*, z) \quad \forall x, z, \tag{7}$$

where the factor of 2 counts individuals rather than partnerships and

$$W(x, z \mid \underset{\sim}{\beta}) = U(x, z \mid \underset{\sim}{\theta}^{W}(\underset{\sim}{\beta})) + V(z, x \mid \underset{\sim}{\theta}^{M}(\underset{\sim}{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

is the sum of the deterministic components of the utilities and $\underset{\sim}{\theta}^{W}(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^{M}(\underset{\sim}{\beta})$ are functions such that $\underset{\sim}{\beta}$ parameterizes $W(x, z \mid \cdot)$. The solution must satisfy the population equilibrium conditions on the parameter values, $\underset{\sim}{\beta}$,

$$\begin{aligned} \frac{f(x, \diamond)}{f(x, *)} &= \int 2\, \mathrm{e}^{W(x,s \mid \underset{\sim}{\beta})} f(*, s)\, \mathrm{d}s \quad \forall x, \\ \frac{f(\diamond, z)}{f(*, z)} &= \int 2\, \mathrm{e}^{W(s,z \mid \underset{\sim}{\beta})} f(s, *)\, \mathrm{d}s \quad \forall z. \end{aligned} \tag{8}$$

The typical number of stable matchings possible increases exponentially with the population size. However, all of these stable matchings have the same limiting probability distribution ($\bar{f}$) over the observed characteristics.

Together, (6) and (7) make it possible to obtain estimates $\hat{\underset{\sim}{\beta}}$ of the preference parameters.

3.3 Parametrization and identifiability

We say that a parametrization of the model, $\beta \in B$, is large population identifiable if for each $\beta_1, \beta_2 \in B$ with $\beta_1 \neq \beta_2$ there exists a state of the covariates x and z such that

$$P(\bar{c} \mid \beta_1) \neq P(\bar{c} \mid \beta_2).$$

Based on Equations (7) and (8), and the expression

$$W(x, z \mid \underset{\sim}{\beta}) = U(x, z \mid \underset{\sim}{\theta}^{W}(\underset{\sim}{\beta})) + V(z, x \mid \underset{\sim}{\theta}^{M}(\underset{\sim}{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}$$

only the sum of the partnered individuals' utilities is identifiable. $U(x, z \mid \underset{\sim}{\theta}^W)$ and $V(z, x \mid \underset{\sim}{\theta}^M)$ may not be separably identifiable when they are additive linear functions as in Equation (3) and include parallel terms. In general, let $\underset{\sim}{\theta}^W(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^M(\underset{\sim}{\beta})$ be functions such that

$$W(x, z \mid \underset{\sim}{\beta}) = U(x, z \mid \underset{\sim}{\theta}^W(\underset{\sim}{\beta})) + V(z, x \mid \underset{\sim}{\theta}^M(\underset{\sim}{\beta})), \quad \forall x \in \mathcal{X}, z \in \mathcal{Z}.$$

In this case, $W(x, z)$ can be parameterized in terms of $\underset{\sim}{\beta}$. We will consider parametrizations, where $\underset{\sim}{\beta}$ is identifiable. To emphasize the relationship between $\underset{\sim}{\beta}$, $\underset{\sim}{\theta}^W$, and $\underset{\sim}{\theta}^M$, we refer to the gender-specific preference parameters as $\underset{\sim}{\theta}^W(\underset{\sim}{\beta})$ and $\underset{\sim}{\theta}^M(\underset{\sim}{\beta})$ for the rest of this paper.

3.4 Reparametrization of the model

We can reparametrize these expressions to improve interpretability and ease computation. Define parameters $g(x, *)$ and $g(*, z)$ via the equations

$$\begin{aligned} f(x, *) &= \frac{\bar{w}(x)e^{g(x,*)}}{(1 + e^{g(x,*)})}, \\ f(*, z) &= \frac{\bar{m}(z)e^{g(*,z)}}{(1 + e^{g(*,z)})}, \end{aligned} \tag{9}$$

so that $g(x, *)$ and $g(*, z)$ both have range the real line.

We can interpret $g(x, *)$ as the log-odds that a women with characteristics x is single. Similarly, we can interpret $g(*, z)$ as the log-odds that a men with characteristics z is single. Hence, this reparametrization is essentially from probabilities to logits. We will use $g(x, *)$ and $g(*, z)$ in place of $f(x, *)$ and $f(*, z)$ to ease computation and interpretability. Note that

$$\begin{aligned} f(x, \diamond) &= \frac{\bar{w}(x)}{(1 + e^{g(x,*)})}, \\ f(\diamond, z) &= \frac{\bar{m}(z)}{(1 + e^{g(*,z)})}, \end{aligned}$$

so that (6) is automatically satisfied and (7) becomes

$$f(x, z) = \text{pref}(x, z)\bar{w}(x)\bar{m}(z) \quad \forall x, z, \tag{10}$$

where

$$\text{pref}(x, z) = 2\frac{e^{W(x,z)+g(x,*)+g(*,z)}}{[1 + e^{g(*,z)}][1 + e^{g(x,*)}]} \quad \forall x, z$$

Equation (10) explicitly separates the availability component of the model ($\bar{w}(x)\bar{m}(z)$) from the preferences-related component ($\text{pref}(x, z)$). In this parametrization, (8) becomes

$$\begin{aligned} e^{-g(x,*)} &= \int 2\frac{e^{W(x,s)+g(*,s)}\bar{m}(s)}{1 + e^{g(*,s)}}\, ds \quad \forall x, \\ e^{-g(*,z)} &= \int 2\frac{e^{W(s,z)+g(s,*)}\bar{w}(s)}{1 + e^{g(s,*)}}\, ds \quad \forall z. \end{aligned} \tag{11}$$

4 Data and inference

4.1 Data

The analysis depends on the sampling design that produces the data. Let $c(x, *)$ and $c(*, z)$ be the sample counts of unmatched women of type x and unmatched men of type z, respectively. Let $c(x, z)$ be the sample counts of matches between women of observed characteristics x and men of type z in the population. Finally, let $\bar{c} = \{c(x, z), c(x, *), c(*, z)\}$, $x \in \mathcal{X}$, $z \in \mathcal{Z}$. Together, $\bar{c}$ defines the empirical version of the distribution $\bar{f}$.

We define a household to be a unit which is either 'single' if it contains of a single, unpartnered person or 'partnered' if it contains two individuals in an exclusive partnership. This definition of household is different from the one often utilized in demography work, where households can consist of a combination of unpartnered and partnered individuals, as well as their offspring. Single households are further differentiated by the gender and type of the individual living in it. Each partnered household is further differentiated by the combination of the type of female and the type of male who live in the household. Each household holds either exactly one single person of any gender or one married couple, and a household is characterized by the type(s) of the individual(s) in it.

Our method can be applied with a broad range of complex survey sampling designs, with the requirement that they produce estimates of $\bar{f}$. Here, we focus on the situation where the data are a probability sample of the individuals in a population where the weights are w_i^w for the i^{th} woman and w_j^m for the j^{th} man. It is presumed that the weights are normalized via post-stratification to sum to population quantities over the covariates in the model. It is also presumed that the characteristics of the partner, if any, of sampled individuals are available. We take a super population framework, where the population is sampled from a super population process. Specifically, the N members of the population are independent and identical draws from a super population stochastic process. The sample of women is denoted $\{x_i, z_i, w_i^w\}_{i=1}^{n_w}$, where z_i are the characteristics of the women's partner, if any. If the sampled women is single formally set z_i to *. Similarly, the sample of men is $\{z_j, x_j, w_j^m\}_{j=1}^{n_m}$.

Estimates of $w(x)$ and $m(z)$ may be available from auxiliary surveys. Otherwise, we can use the data alone and standard design-based estimates of $w(x)$ and $m(z)$, written as $\tilde{w}(x)$ and $\tilde{m}(z)$, respectively. Note that these represent *availabilities* and do not depend on the preference parameters. The parameters are then $\underset{\sim}{\psi} = (\underset{\sim}{\beta}, \{g(x, *)\}_{x\in\mathcal{X}}, \{g(*, z)\}_{z\in\mathcal{Z}})$.

4.2 Large-population likelihood approach

Had we observed the entire population, the likelihood for ψ would involve the complex dependencies between the individual choices and matchings in the population. Each of the matchings is interdependent. Our approach is to use as a surrogate for the likelihood for ψ, one based on the likelihood of the observed frequencies of pairings by covariates, $\bar{c}$, and model (7) and (8). Specifically, we approximate the exact likelihood for $\underset{\sim}{\psi}$ by

$$\begin{aligned} &\text{lp} - \text{log-lik}(\underset{\sim}{\beta}, g(x, *), g(*, z) \mid \{x_i, z_i, w_i^w\}_{i=1}^{n_w}, \{z_j, x_i, w_j^m\}_{j=1}^{n_m}) \\ &\quad = \sum_{x\in\mathcal{X}}\sum_{z\in\mathcal{Z}} c(x, z) \log f(x, z) + \sum_{x\in\mathcal{X}} c(x, *) \log f(x, *) + \sum_{z\in\mathcal{Z}} c(*, z) \log f(*, z) \end{aligned} \tag{12}$$

The log-likelihood (12) can be written in terms of $g(x, *)$ and $g(*, z)$ using (10). The values $\tilde{w}(x)$ and $\tilde{m}(z)$ replace $w(x)$ and $m(z)$ in these expressions.

To obtain estimates, (12) can be maximized subject to the constraints expressed in (11) to produce the maximum large-population likelihood estimator (MLPLE), $\hat{\psi}$. This was achieved via a sequential quadratic programming (SQP) algorithm for non-linearly constrained gradient-based optimization (Johnson, 2020; Kraft, 1994). The algorithm optimizes successive second-order (quadratic/least-squares) approximations of the objective function (via BFGS updates), with first-order (affine) approximations of the constraints. We note that there are many possible survey sampling schemes in use, and the sampling could be at the individual level or at the household level. These alternative survey designs are straightforward to incorporate into the above equations and we do not explicate it here.

4.3 Correcting the estimator for bias and confidence coverage

It is likely that the estimator of Section 4.2 will be biased primarily as it is based on a large-population approximation to the generating process followed by a number of statistical approximations. As noted in Section 4.1, we take a super population framework, with N specifying the size of the draw from the super population to the population and $n \leq N$ the size of the subsequent draw of the sample from the population. There is added uncertainty associated with both steps (specifically, the large-population approximation at the first step and the sampling error at the second step).

The large-population approximation does not take into account information in the matching that is not captured by the counts of matches and singles by type. In addition, the super population sampling distribution of these counts is not multinomial. While the utilities in Equations (1) and (2) are independent, the matches are interdependent and hence so are the counts. However, the counts are asymptotically (with N) sufficient for the parameters (Menzel, 2015) and the bias should be smaller for large population sizes.

To address this, we propose using bootstrap procedures to estimate the sampling distribution of the estimator and correct for bias and confidence coverage. We propose two versions of this bootstrap: a parametric version that is preferred where computationally feasible and a classical version to be used for large population sizes.

4.3.1 Parametric bootstrap

If the population size is small (e.g., less than 20,000), we can generate the (stochastic) relational utilities for all population members using Equations (1), (2), and (4) at the MLPLE parameter values. We can then use the Gale–Shapley algorithm to achieve a stable matching for that population. This matching is from the population generating process of the data. We follow it with a sampling of size n using the sampling design of the data including survey weights (e.g., stock-stock, stock-flow, census). We repeat this process b times, so that we have b bootstrapped samples. We fit the revealed preferences model to each of the b samples and obtain the bootstrapped parameter estimates for a single parameter $\underset{\sim}{\psi}$, which we denote as $\underset{\sim}{\psi}^* = [\underset{\sim(1)}{\psi^*}, \underset{\sim(2)}{\psi^*}, \ldots, \underset{\sim(b)}{\psi^*}]$. Doing so requires us to re-solve a constrained maximization problem for each bootstrap sample. This can be computationally expensive but is simply parallelizable (as we have done in the software associated with this paper Handcock et al., 2022).

The empirically estimated bias of $\hat{\psi}$, denoted as $\widehat{\text{bias}}_{\hat{\psi}}$, is equal to the mean of the bootstrapped parameter estimates $\underset{\sim}{\psi}^*$ minus $\underset{\sim}{\hat{\psi}}$. We then propose as our bias-corrected point estimator $\underset{\sim BC}{\hat{\psi}} = 2\underset{\sim}{\hat{\psi}} - \frac{1}{b}\sum_{i=1}^{b} \underset{\sim(i)}{\psi^*}$.

As we are drawing directly from the super-population generating and sampling processes, we believe this will provide a firm basis for bias-reduction and coverage correction for the census case.

4.3.2 Large-population bootstrap

The computational burden of the Gale–Shapley algorithm is large for large populations (e.g., $N > 20,000$). In this case, we consider a classical bootstrap for survey data, simple random resampling b data sets from the original data with replacement so that we have b sets of bootstrapped samples (Shao & Tu, 1995). As before, we fit the revealed preferences model to each of the b samples and obtain the bootstrapped parameter estimates for a single parameter $\underset{\sim}{\psi}$, which we denote as $\underset{\sim}{\psi}^* = [\underset{\sim(1)}{\psi^*}, \underset{\sim(2)}{\psi^*}, \ldots, \underset{\sim(b)}{\psi^*}]$ and propose a bias-corrected point estimator appropriate for survey data $\underset{\sim BC}{\hat{\psi}} = \underset{\sim}{\hat{\psi}} - \frac{N-n}{n-1}\left(\frac{1}{b}\sum_{i=1}^{b} \underset{\sim(i)}{\psi^*} - \underset{\sim}{\hat{\psi}}\right)$ (McCarthy & Snowden, 1985).

In this scenario, $N \gg n$ so that sampling uncertainty dominates errors from the large-population approximation. We then appeal to survey sampling bootstrap asymptotics as justification (Shao & Tu, 1995, Theorem 6.5).

This procedure and its parametric complement appear to work well, as is borne out in the simulation studies of Sections 5 and 6.

4.4 Measuring uncertainty of the estimates

Once we obtain the parameter estimates $\hat{\underset{\sim}{\psi}}$, a natural next step is to measure their uncertainty.

The covariance matrix of the estimates can be approximated by a standard Central Limit Theorem argument. The approximate log-likelihood function, augmented by the constraints, is

$$\text{log-lik}_A(\underset{\sim}{\psi} \mid \{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) \tag{13}$$

$$= \text{lp} - \text{log-lik}(\underset{\sim}{\psi} \mid \{x_i, z_i, w_{wi}\}_{i=1}^{n_w}, \{z_j, x_i, w_{mj}\}_{j=1}^{n_m}) + \sum_{k=1}^{|\mathcal{X}|+|\mathcal{Z}|} \lambda_k h_k(\underset{\sim}{\psi}), \tag{14}$$

where $\{h_k(\underset{\sim}{\psi})\}_{k=1}^{|\mathcal{X}|+|\mathcal{Z}|}$ are constraints (10). Its Hessian is

$$\mathbb{E}\left(\frac{\partial^2 \text{log-lik}_A}{\partial \underset{\sim}{\psi} \partial \underset{\sim}{\psi}'}\right) = \begin{pmatrix} H & J \\ J^T & 0 \end{pmatrix}, \tag{15}$$

where H is the Hessian of (12) with ij^{th} element $\mathbb{E}\left(\frac{\partial^2 \text{lp}-\text{log}-\text{lik}}{\partial \underset{\sim}{\psi} \partial \underset{\sim}{\psi}'}\right)$ and J is the Jacobian matrix of the constraints with kj^{th} element $\frac{\partial h_k(\underset{\sim}{\psi})}{\partial \underset{\sim}{\psi}}$. The estimate of the (asymptotic) covariance matrix of the MLPLE of $\underset{\sim}{\psi}$ is the (1,1) block of the Moore–Penrose inverse of this matrix (Hartmann & Hartwig, 1996).

The accuracy of the estimate of the covariance matrix depends on the application-specific accuracy of the various approximations. Thus, the analytically estimated standard errors may not accurately reflect the standard errors of parameter estimates that are observed over repeated samples from the same population. However, they are easy and fast to compute. It is natural to consider robust (sandwich formula) variance estimators for this situation. However, these performed poorly as they did not adequately take into account the constraints.

As an alternative, we propose estimating standard errors empirically using the bootstrap procedures of Section 4.3. Most directly, the empirically estimated standard error of $\hat{\underset{\sim}{\psi}}$, denoted as $\widehat{\text{se}}_{\hat{\underset{\sim}{\psi}}}$, is equal to standard error of the bootstrapped parameter estimates $\underset{\sim}{\psi}^*$.

We also consider various methods employing bootstrap procedures to compute confidence intervals for each parameter. The *percentile bootstrap*, is the most straightforward of these methods. We denote $\underset{\sim(\alpha)}{\psi^*}$ as the α percentile of the bootstrap parameter estimates $\underset{\sim}{\psi}^*$. The $(1-\alpha)\%$ percentile bootstrap confidence interval for parameter $\underset{\sim}{\psi}$,

$$(\underset{\sim(\alpha/2)}{\psi^*}, \underset{\sim(1-\alpha/2)}{\psi^*}).$$

The second method we employ is the basic bootstrap confidence interval. For the parameter $\underset{\sim}{\psi}$ with estimate $\hat{\underset{\sim}{\psi}}$, we use the basic bootstrap procedure to obtain a $(1-\alpha)$ confidence interval,

$$(2\hat{\underset{\sim}{\psi}} - \underset{\sim(1-\alpha/2)}{\psi^*}, 2\hat{\underset{\sim}{\psi}} + \underset{\sim(\alpha/2)}{\psi^*}).$$

We also consider a modified version of the studentized t bootstrap confidence interval. Here we obtain a $(1-\alpha)\%$ confidence interval as:

$$(\hat{\underset{\sim}{\psi}} - t^*_{(1-\alpha/2)}\widehat{\text{se}}_{\hat{\underset{\sim}{\psi}}}, \hat{\underset{\sim}{\psi}} + t^*_{(\alpha/2)}\widehat{\text{se}}_{\hat{\underset{\sim}{\psi}}}).$$

We test the performances of the analytical confidence intervals as well as those of all three proposed bootstrap confidence interval methods in Section 6.5 as part of our simulation studies.

5 Simulation studies of model and inferential accuracy

In this section, we illustrate the statistical properties of the revealed preferences model by conducting three simulation studies which we refer to hereafter as studies I, II, and III. In simulation study I, we show that the revealed preferences model accurately estimates underlying preference parameters which partially motivate matching outcomes in a population under different availability scenarios. In simulation study II, we investigate the relationship between the population size N and bias of preference parameter estimates produced by the revealed preferences model when census data is available. In simulation study III, we investigate the relationship between the relative sample proportion n_h/N and bias of preference parameter estimates when data are available for a sample of a population. In all three studies, we show the bias-corrected maximum large-population likelihood estimates (MLPLEs) for the preference parameters, adjusted using the methodology proposed in Section 4.3. In addition, in studies II and III, we also show the MLPLEs prior to bias correction and compare them to the bias-corrected MLPLEs, demonstrating that the bias-corrected MLPLEs consistently improve estimate accuracy with little cost to precision.

Together, the simulation studies shown in this paper make a significant contribution to existing literature as they clearly demonstrate the novel ability of our proposed revealed preferences methodology to separate effects of preference and availability on matching outcomes. Previously, Menzel (2015) presented a simulation study with maximum-likelihood estimation of preference parameters. However, his results were extremely limited in that he considers populations that are restricted to size $N \leq 2,000$ and are generated under a single availability scenario. In contrast, we will show that the revealed preferences model recovers preferences for given sample or census data for a wide range of population (sample) sizes and under different availability scenarios. We also demonstrate the use of bias-correction procedures to improve the accuracy of our estimates. For researchers in other fields who will apply our model, we also consider several different specifications for the systematic component of the utility function to demonstrate the flexibility of our proposed approach.

The remainder of this section is structured as follows: we first describe a general procedure for the three simulation studies. We then describe the two availability scenarios considered for generating individuals of different genders and education in each simulated population in Section 5.1. In Section 5.2, we discuss the choice of $\underset{\sim}{\beta}^0$ and the different utility model specifications considered for the function $W(x_i, z_j \mid \underset{\sim}{\beta})$. Once we have defined the availability scenarios and utility model specification, we then provide further detail about the different specifications of each study in Section 5.3.

The basic procedure for the different simulation studies is the same. We begin by assuming a heterosexual marriage market in which males and females base partnership decisions on their own education level and the education of prospective spouses, as well as some other unobserved characteristics. We assume that the marginal distributions of gender and education within the population are known and represented as availability scenario $\mathcal{A} = \{\bar{w}(x), \bar{m}(z)\}$. We also assume that the form of the partnership utility function $W(x_i, z_j \mid \underset{\sim}{\beta})$ and the preference parameters $\underset{\sim}{\beta}$ for individuals in the market are both known.

We suppose a population of size N which reflects the gender and education distributions of availability scenario $\mathcal{A}$ and the partnership preferences $\underset{\sim}{\beta}$. In simulation studies I.i and II, we assume the data consist of information on the full simulated population, while for simulation studies I.ii and III, we suppose that the data are a sample of n_h households from the simulated population. We then obtain the distribution of partnerships $\bar{c}$, either empirically or via large-population approximation described in equation 7. We fit the revealed preferences model to the data to produce estimates $\underset{\sim}{\hat{\beta}}$ of the original preference parameters.

5.1 Choice of availability scenarios

We consider two marginal distributions for gender and education as our availability scenarios, referred to hereafter as $\mathcal{A}_1$ as $\mathcal{A}_2$. Both availability scenarios were chosen based on data from the 2008

Panel of the Survey of Income and Program Participation (SIPP), which has been made publicly available by the United States Census Bureau (U.S. Bureau of the Census, 2020). The 2008 SIPP is a nationally representative panel study that followed individuals in sampled households from 2008 through 2012. Individuals responded to a set of core questionnaires administered every 4 months and in 2009, individuals over the age of 15 answered a series of supplemental survey questions on their marital history, and, if currently married, the date their most recent marriage began.

We limit the analytic sample to individuals 18–59 years old who at wave 2 had married in the past year or were not currently married and were living in households that responded to Waves 1 and 2 of the 2008 SIPP Panel as well as the marital history topical module administered at the Wave 2 interview. We focus on marriages that initiated no more than a year prior to the survey data to ensure we capture preferences at the time the marriage was initiated and to avoid bias due to marital dissolution, remarriage, or educational upgrading (Kalmijn, 1994; Schwartz & Mare, 2005). With these limitations, our analytic sample consists of 21,597 individuals, 1,040 of whom had married in the last year, and 20,527 who remained single in the last year. The 1,040 newly married individuals were by survey design married to another sample member, and, therefore, were in 520 couples in our sample. Within a given year, entering into a marriage is therefore relatively rare, with only 5% of individuals in our analytic sample having entered a new marriage. Thus preferences for marriage, meaning for getting married in a given year, are negative when we run the revealed preferences model in Section 6. This 2008 SIPP sample design corresponds to Menzel's (p. 913) sample of households that are assumed to be drawn from a population resulting from the stable matching. In our case, we have 21,077 households that include 520 couples.

The maximum education level attained by each individual is a categorical variable coded as 1 for less than a high school education, 2 for a high school degree, 3 for some college, and 4 for a bachelors degree or beyond. The education level of female i is stored as x_i and the education level of male j is stored as z_j.

The first availability scenario $\mathcal{A}_1$ is factual (a population like the 2008 SIPP). In other words, it utilizes the gender and education distributions of the overall population based on the 2008 SIPP sample, and the partnership preferences of individuals are equal to preferences estimated in the 2008 SIPP sample. In this availability scenario, about 49.1% of individuals are women and 51.9% are men.

Availability scenario $\mathcal{A}_2$ has the same marginal distribution of education and availability as the non-Hispanic Black population in the 2008 SIPP data. However, the preferences of individuals in availability scenario are kept the same as those of individuals in scenario $\mathcal{A}_1$. Under availability scenario $\mathcal{A}_2$ about 58.0% of individuals are females and 42.0% are males, which reflects a significant gender skew not seen in scenario $\mathcal{A}_1$. In both $\mathcal{A}_1$ and $\mathcal{A}_2$, women are less likely to have less than a high school degree (education category 1) and are more likely to have completed any college (education category 3 or higher).

In simulation studies I.i and I.ii, we simulate populations from both scenarios $\mathcal{A}_1$ and $\mathcal{A}_2$. Given the utility model specification, we assume that in both scenarios all individuals are characterized the same true preference parameters $\underset{\sim}{\beta}^0$. By fitting the revealed preferences model on data from populations based on both availability scenarios, we show that preference parameter estimates are unbiased even as the availability of potential partners changes. Thereafter, in simulation studies II and III, we only simulate populations based on availability scenario $\mathcal{A}_1$ (Tables 1 and 2).

Table 1. Availability scenarios

Availability scenario	Source of availability distribution	Type
$\mathcal{A}_1$	2008 SIPP full sample	Total U.S. population in 2008
$\mathcal{A}_2$	2008 SIPP non-Hispanic Black sample	A realistic sub-population availability

Table 2. Gender and education distributions under the two availability scenarios

	Males		Females	
Education level	**% Population**	**% of Males**	**% Population**	**% of Females**
		Availability scenario $\mathcal{A}_1$		
1 (< high school)	7.4	14.5	5.3	10.9
2 (high school)	14.5	28.5	11.2	22.8
3 (some college)	19.5	38.4	21.0	42.9
4 (≥ bachelors)	9.5	18.6	11.5	23.4
Total	**50.9**	**100.0**	**49.1**	100.0
		Availability scenario $\mathcal{A}_2$		
1 (< high school)	7.2	17.1	7.1	12.3
2 (high school)	13.8	33.0	15.3	26.4
3 (some college)	15.9	37.8	25.4	43.7
4 (≥ bachelors)	5.1	12.1	10.2	17.6
Total	**42.0**	**100.0**	**58.0**	100.0

5.2 Utility model specification

We now discuss three different partnership utility specifications under which we test the performance of the revealed preferences model. We first consider a very simple model specification in which a female experiences a shift in utility, relative to her utility had she remained unpartnered, only when she partners with a man whose education level is the same as her own. The tendency for partnered individuals to share similar characteristics is reflected by *homogamous* pairings, and preference for such partnerships is referred to as *homophily*. We designate this specific model as the *uniform homophily model* because the shift in the deterministic component of the utility is uniform for all types (education levels) of individuals. The set of parameters for this model is denoted as $\underset{\sim}{\beta}^{\mathrm{UH}}$. The sum of woman i and man j's utilities if they partnered with each other is

$$W_{ij}(x_i, z_j \mid \underset{\sim}{\beta}^{\mathrm{UH}}) = \beta_0 + \beta_1 \mathbb{I}\{x_i = z_j\}. \tag{16}$$

The uniform homophily model can be extended if we assume that the utility a woman derives from a partnership is based not only on whether she and her partner have equal education levels, but also on the education level itself. Once again, there is a corresponding utility function for males. We refer to this as a *differential homophily model*, where the change in utility depends not only on partners share a particular trait, but also on the value of trait considered. The set of parameters for this model is denoted as $\underset{\sim}{\beta}^{\mathrm{DH}}$,

$$W_{ij}(x_i, z_j \mid \underset{\sim}{\beta}^{\mathrm{DH}}) = \beta_0 + \sum_{k=1}^{4} \beta_k \mathbb{I}\{x_i = z_j = k\}. \tag{17}$$

The third model we consider is a modified version of the *saturated mix model*, which includes every possible first-order term. In the saturated mix model, women and men both derive a different utility from each possible combination of education levels in the marriage. The full set of parameters is denoted by the vector β^{SM}.

We are able to remove the intercept term β_0 from the utility model because it is a constant value added to the matching utility of every pair. Thus, the sum of the utilities of two individuals in a marriage is given by

$$W(x_i, z_j \mid \underset{\sim}{\beta}^{\mathrm{SM}}) = \sum_{p,q} \beta_{p,q} \mathbb{I}\{x_i = p, z_j = q\}. \tag{18}$$

The term $\beta_{p,q}$ is the coefficient to an indicator which equals 1 if the couple consists of a woman of type p and a man of type q, and 0 otherwise. The saturated mix model consists of $P \times Q$ first-order parameters, where there are P possible types for women and Q possible types for men.

Out of the 21,077 households in the SIPP analytic sample, there is 1 couple which consists of a woman with education level 1 and a man with education level 4, and 1 couple which contains a woman with education level 4 and a man with education level 1. The low counts make estimation of the $\theta_{1,4}$ and $\theta_{4,1}$ parameters difficult, as the joint utility of such couple is perceived as effectively negatively infinite. To facilitate estimation in these cases, we consider pairings between a woman with education level 1 and a man of education level 4 to have equal utility to a pairing between a woman with education level 2 and a man of education level 4. This 'reduces' the $\beta_{1,4}$ and $\beta_{2,4}$ parameters to a $\beta_{1\text{ or }2,4}$ parameter. Likewise, we can equate pairings between a woman with education 4 and man with education 1 to pairings between a woman with education 4 and a man with education 2, so that $\beta_{4,1}$ and $\beta_{4,2}$ are replaced by $\beta_{4,1\text{ or }2}$. Thus, rather than using the fully saturated model with 16 parameters to estimate, we consider a *reduced mix model* with only 14 parameters, represented in vector form as $\underset{\sim}{\beta}^{\text{RM}}$. The situation here is very similar to the 'collapsing cells' situation in contingency table modeling (Agresti, 2012, Section 10.1).

We note that mix models are of particular interest to demographers who have access to large samples from populations. When the size of the available data is small as is the case for simulation studies II and III, however, model saturation can result in biased and highly variable parameter estimates and the less parametrized uniform homophily or differential homophily model may be preferable.

The testing procedure for each model specification is the same, and we outline the basic procedure which is used in simulation study I. We first choose a set of preference parameters $\underset{\sim}{\beta}^0$ given the specific model that we assume is the underlying truth. This is done by using RPM to fit that model on the analytic 2008 SIPP data and calculating parameter estimates $\underset{\sim}{\beta}$. We assume that these estimates are equivalent to the true preference parameters of individuals under all availability scenarios, so that $\underset{\sim}{\beta}^0 = \underset{\sim}{\beta}$. In each simulated population, the known preferences $\underset{\sim}{\beta}^0$ are applied to calculate total household utility for every potential partnership and form a stable matching. We fit the revealed preferences model on the observed stable matching outcome from the simulated population, constraining the MLPLEs to lower and upper bounds of -10 and 10, respectively, and utilize the methodology proposed in Section 4.3 to obtain bias-corrected MLPLEs. We compare these estimates to the true underlying true preferences $\underset{\sim}{\beta}^0$. We make minor modifications to this process for simulation studies II and III which are described below.

5.3 Details for simulation studies I, II, and III

Having established the availability scenarios and utility models, we will consider in this paper, we now provide further detail on each of the simulation studies.

To demonstrate that the revealed preferences model produces unbiased estimates of β given either an observed distribution of partnerships $\bar{c}$ or a large-population approximation of $\bar{c}$, we conduct simulation study I in two parts. In study I.i, we simulate populations of size $N = 6,000$. The generating distribution for the populations may be either availability scenario $\mathcal{A}_1$ or $\mathcal{A}_2$, and a population consists of individuals whose partnership utilities are either all determined by the differential homophily utility model (Equation 17) or the reduced mix utility model (Equation 18). Thus, we consider four possible combinations of availabilities and utility model specifications, and we simulate 1,000 populations of each combination. For every simulated population, based on the utility function and $\underset{\sim}{\beta}^0$ we obtain a stable matching using the Gale–Shapley algorithm. (Gale & Shapley, 1962) We then compute the empirical distribution of partnerships $\bar{c}$ observed in this stable matching. Treating the simulated data as a census, we fit the revealed preferences model to obtain preference parameter estimates.

Ideally, to obtain the distribution of partnerships within a population, we would always use the Gale–Shapley algorithm to first achieve a stable matching for that population. However, a large

amount of memory and computational power is required to create stable partnerships for large population sizes (e.g., greater than 20,000), since the household utility matrices $\{W_{ij}\}_{N_w \times N_m}$ and $\{M_{ij}\}_{N_m \times N_w}$ must be calculated for all potential pairings. In such cases, rather than implementing the Gale–Shapley algorithm to achieve a stable matching, we can approximate the empirical distribution of household types in the outcome and estimate preference parameters based on the large-population approximation (Equation (7)). In general, we suggest using the large-population approximation rather than replicating the actual matching process when working with simulated populations with more than 6,000 individuals.

In study I.ii, we show that a large-population approximation of $\bar{c}$ is suitable for unbiased estimation of preference parameter estimates. We begin once again assuming that a population can be characterized by the same four combinations of availabilities and utility model specifications considered in study I.i. In this case, however, we suppose that $N = 300$ million within a single population. Rather than simulating the population directly, we approximate the distribution of partnerships that would occur in a stable matching within such a population. We then sample about 20 thousand households from this approximated distribution, fit the revealed preferences model to the sample data, and obtain preference parameter estimates. For each combination of availability and utility model, we take 1,000 samples.

We note here that populations generated using availability scenario $\mathcal{A}_1$ can be considered 'factual' in that they resemble the 2008 SIPP sample. In other words, both the underlying marginal distributions of gender and education $\mathcal{A}_1$ and the preferences β used to generate matchings in the simulated population are based on the 2008 SIPP. In contrãst, populations generated using availability scenario $\mathcal{A}_2$ are 'counter-factual' as the population composition changes while preferences of the 2008 SIPP are maintained.

In simulation study II, we simulate 1,000 populations each of size $N = 60$, 600, and 6, 000 with the assumption that the education and gender for individuals in all populations are generated based on availability $\mathcal{A}_1$ and all individuals have a uniform homophily utility model (Equation 16) for partnership. We choose the uniform homophily model for this part of the study to avoid negatively infinite estimates at $N = 60$. We also make a small modification here to the model testing procedure described previously; we do not set the true underlying preferences $\underset{\sim}{\beta}^0$ equal to $\underset{\sim}{\tilde{\beta}}^{\mathrm{UH}}$, the preference estimates obtained by fitting the uniform homophily model on the SIPP data. Instead, we increase the intercept term in $\underset{\sim}{\tilde{\beta}}^{\mathrm{UH}}$ by a magnitude of 4 to increase the number of partnerships and facilitate stable estimation of preference parameters. For each simulated population, we use the Gale–Shapley algorithm to obtain a stable matching and fit the revealed preferences model to the observed $\bar{c}$ for the entire population. We then compare the bias of the median parameter MLPLEs and bias-corrected MLPLEs at each N as N increases. We also evaluate the effectiveness of using a bootstrap approach for bias correction of $\underset{\sim}{\hat{\beta}}$ at different N.

For simulation study III, we simulate populations of size $N = 6, 000$ with the assumption that the education and gender for individuals in all populations are generated based on availability $\mathcal{A}_1$ and all individuals have a differential homophily utility model (Equation 17) for partnership. For each stable population, after using the Gale–Shapley algorithm to reach a stable matching, we sample $n_b = 600$, 1, 200, or 3, 000 households. Similar to simulation study II, rather than set $\underset{\sim}{\beta}^0 = \underset{\sim}{\tilde{\beta}}^{\mathrm{DH}}$, we increase the intercept term in $\underset{\sim}{\tilde{\beta}}^{\mathrm{DH}}$ by 4 units to increase partnership rates. We fit the revealed preferences model to the sample data and compare the performance the mean MLPLEs and bias-corrected MLPLEs $\underset{\sim}{\hat{\beta}}$ as n_b increases.

6 Results

6.1 Simulation study I.i: population data

For simulation study I.i, we simulate populations of size $N = 6, 000$ from 'factual' availability $\mathcal{A}_1$ and 'counterfactual' availability $\mathcal{A}_2$ and utilized the Gale–Shapley algorithm to perform stable matching on the individuals in each simulated population. The utility derived from each potential partnership was calculated based on $\underset{\sim}{\beta}^0$ for a specified deterministic utility function and an

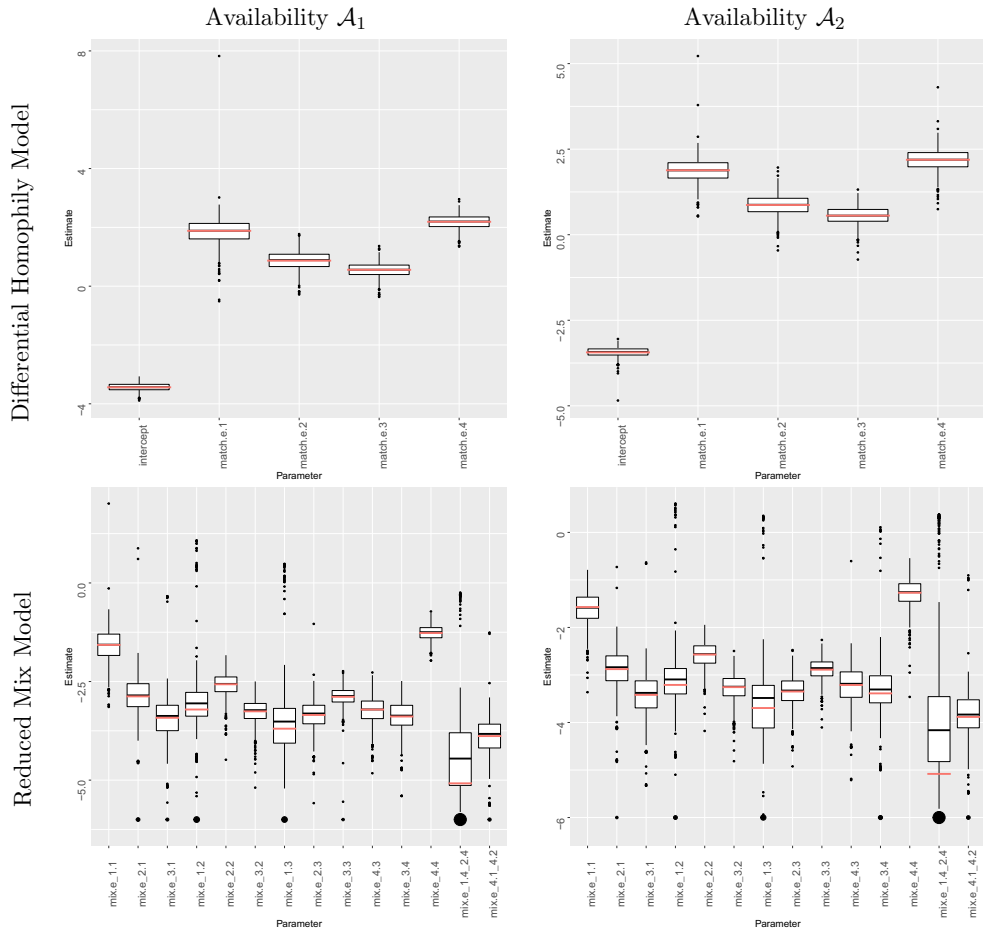

Figure 1. Distribution of bias-corrected MLPLEs in simulation study I.i: Population data with $N = 6,000$ (1,000 simulations).

extreme-value Type-I distributed random error term. The utility a woman achieves by staying single is equal to maximum value of $\sqrt{N_w}$ random draws from an extreme-value Type-I distribution.

The plots in Figure 1 show the distribution of the 1,000 bias-corrected MLPLEs for each combination of availability scenario $\mathcal{A} \in \{\mathcal{A}_1, \mathcal{A}_2\}$ and two utility model specifications (differential homophily and reduced mix). The red lines in the plots represent the true β_0 preference values which induced the Gale–Shapley matchings. Negatively infinite estimates are recognized via a point mass at value -6 with an area proportional to the number of such estimates.

The medians and standard deviations, of parameter estimates for the match and reduced mix models are presented in Online Supplementary Material, Tables 3 and 4. For this and all following simulation studies, we compute standard deviation as a standardized version of the interquartile range. Tables with numerical results are in Online Supplementary Material, Appendix A.

Although availability of individuals differs between $\mathcal{A}_1$ and $\mathcal{A}_2$, under both model specifications the revealed preferences model produces estimates of the true preference parameters which are about equal in accuracy and precision. Based on the plots for study I.i in Figure 1, the mean estimates of all reduced mix model parameters except $\beta_{1\text{ or }2,4}$ appear to align with the true values fairly well in all availability scenarios. Furthermore, the estimates for all parameters, with the exception of $\beta_{1\text{ or }2,4}$, resemble a normal distribution.

We note that when using the reduced mix model, for both availability scenarios the distribution of $\hat{\beta}_{1\text{ or }2,4}$ displays a right skew. When the population has very few or no pairings of a certain type, the model estimates the total utility of such a pairing as very negative, if not infinitely so. In our

implementation of this model, we impose an upper bound of 10 and a lower bound of −10 on all parameters. The high frequency of extremely negative values (≤ -6) in the parameter estimates of $\beta_{1\text{ or }2,4}$ indicates that in that specific population, there were very few or no households which contained a matching between a woman with education level 1 or 2 and a man with education level 4.

We ran simulation study I.i with both the differential homophily and reduced mix model specifications on a third availability scenario (results not shown), in which men outnumber women 3:1 and educational attainment was highly asymmetric across genders. We found that in this artificially extreme case, the occurrence of highly negative estimates of $\beta_{1\text{ or }2,4}$ increased. Furthermore, the estimates of $\beta_{1,3}$ and $\beta_{2,3}$ also showed a strong right skew. In general, the standard deviation of the parameter estimates tends to increase as the population becomes more skewed.

6.2 Simulation study I.ii: sampling from a large population

In this simulation study, we simulate samples from large populations using availabilities $\mathcal{A}_1$ and $\mathcal{A}_2$, each with a nominal size of $N = 300$ million and a household sample size of $n_h = 21,077$ (equivalent to the size of the analytic SIPP sample). We find that the resulting estimates are very robust to the population size as long as it is modestly large (e.g., $N > 6,000$). We choose to study large populations as they are typical in demography. Brien (1997), for example, compares model performance for three levels of population aggregation of the marriage market: in descending order, state, metropolitan area, and county group. He finds that the highest, state level of aggregation best explains marriage differentials between population subgroups.

We employ a large-population approximation of stable matching outcomes in the simulated population that would be observed if individuals had true preferences $\underset{\sim}{\beta}^0$, either based on a differential homophily or a reduced mix utility model. The plots in Figure 2 show the distribution of the 1,000 parameter estimates $\underset{\sim}{\hat{\beta}}$ for each combination of simulating availability scenario and revealed preferences model specification. The red lines in the plots represent the true values $\underset{\sim}{\beta}^0$ which we are attempting to recover.

The first row of Figure 2 shows the distributions of the parameter estimates under the differential homophily model given large simulated population. The medians and standard errors of the differential homophily model parameters are presented in Online Supplementary Material, Table 5.

In both availability scenarios, we observe that the mean estimate for each parameter in the differential homophily model is very close to the true value. We also note that when simulating from availability scenarios $\mathcal{A}_1$ and $\mathcal{A}_2$, the standard errors of the parameter estimates stay about the same. However, we also ran this simulation study under the artificially extreme availability scenario described in the results for study I.i (results not shown) and found that in that case the standard error nearly tripled for all parameters.

The second row of Figure 2 shows the distributions of the parameter estimates under the reduced mix model when the simulated population size is large. Due to space constraints, we relegate Online Supplementary Material, Table 6, which shows the medians and standard errors of the parameter estimates, to Online Supplementary Material, Appendix A. The revealed preferences model recovers the true preference parameters $\underset{\sim}{\beta}^{\text{RM},0}$ for all availability scenarios. Furthermore, the standard deviations of all parameter estimates stay similar across the availability scenarios.

6.3 Simulation study II: small population sizes

Simulation study II is carried out for two primary purposes. The first purpose is to illustrate how the revealed preferences model can be used with population data that includes small to very-small population sizes. The second is to show the relationship between population size N and estimate bias and the relationship between population size N and the effectiveness of our proposed bias correction methodology.

We simulate 1,000 populations each of sizes $N = 60$, 600, and 6,000 from availability scenario $\mathcal{A}_1$. We then use the Gale–Shapley algorithm to obtain a stable matching in the population, with true preference parameters $\underset{\sim}{\beta}^{\text{UH},0}$ based on the uniform homophily model and the inflated

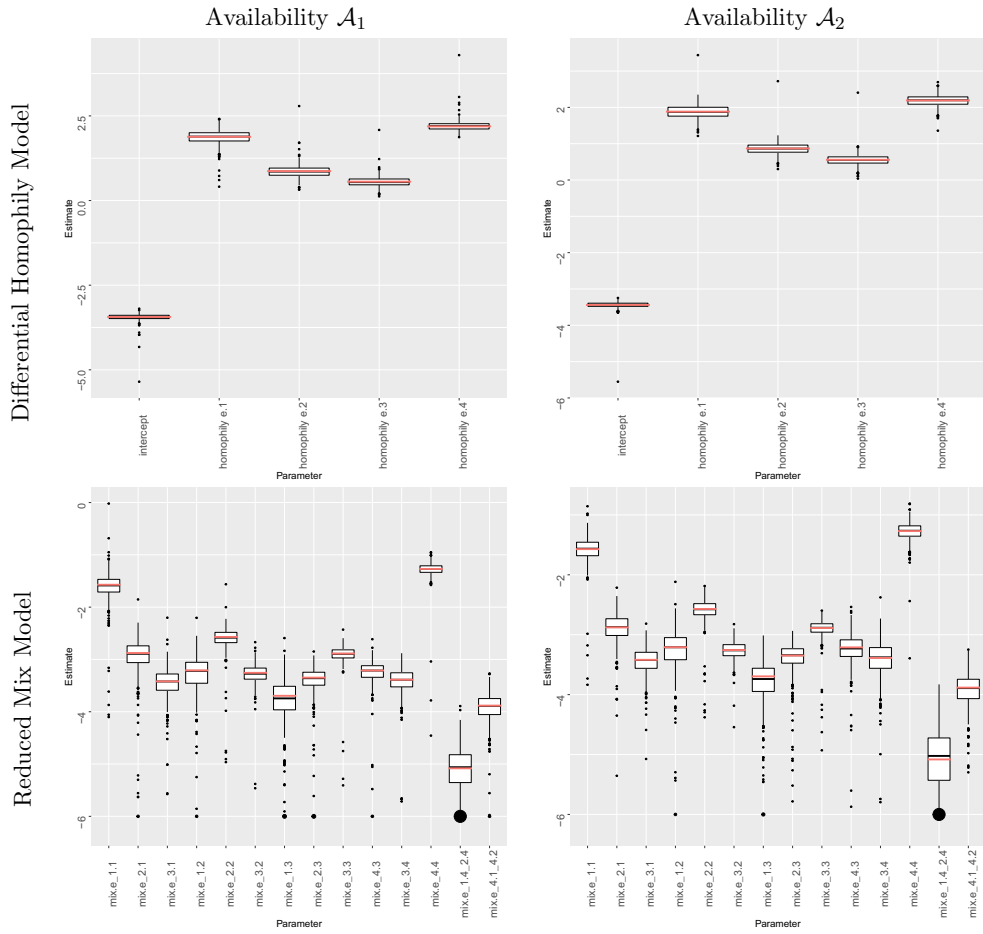

Figure 2. Distribution of bias-corrected MLPLEs in simulation study I.ii: Sample data with $n_h = 21,077$ from a population of $N = 300$ million (1,000 simulations)

intercept. The distributions of the maximum large-population likelihood estimates (MLPLEs) and the bias-corrected MLPLEs for each N are shown in Figure 3. The median estimates and standard deviations of the MLPLEs and bias-corrected MLPLEs given in Online Supplementary Material, Table 7, respectively.

The panels in the first column of Figure 3 show model estimates for each parameter when $N = 60$. Each panel corresponds to a single parameter and shows two distributions; the left box plot shows the distribution of the MLPLEs and the right box plot shows the distribution of the bias-corrected MLPLEs. The second and third columns of Figure 3 show the same information for $N = 600$ and $N = 6,000$.

At each population size, the MLPLEs for both the intercept term and the uniform homophily preference term underestimates the true value $\tilde{\beta}^{\mathrm{UH},0}$, though the bias of the latter term is of a much smaller magnitude than of the former. For both parameters, bias decreases as the population N increases. The standard deviation of the MLPLE estimate decreases substantially as N increases; we see in Online Supplementary Material, Table 7 that when N increases by a factor of 10, the standard deviation decreases by a factor of approximately 1/3 for the intercept parameter and 1/4 for the homophily parameter.

When bias correction methodology is applied, the bias in the estimates of both the intercept and the homophily parameter decreases for all population levels. The improvement of the estimates due to bias correction is especially clear for the intercept term. We also notice that for both

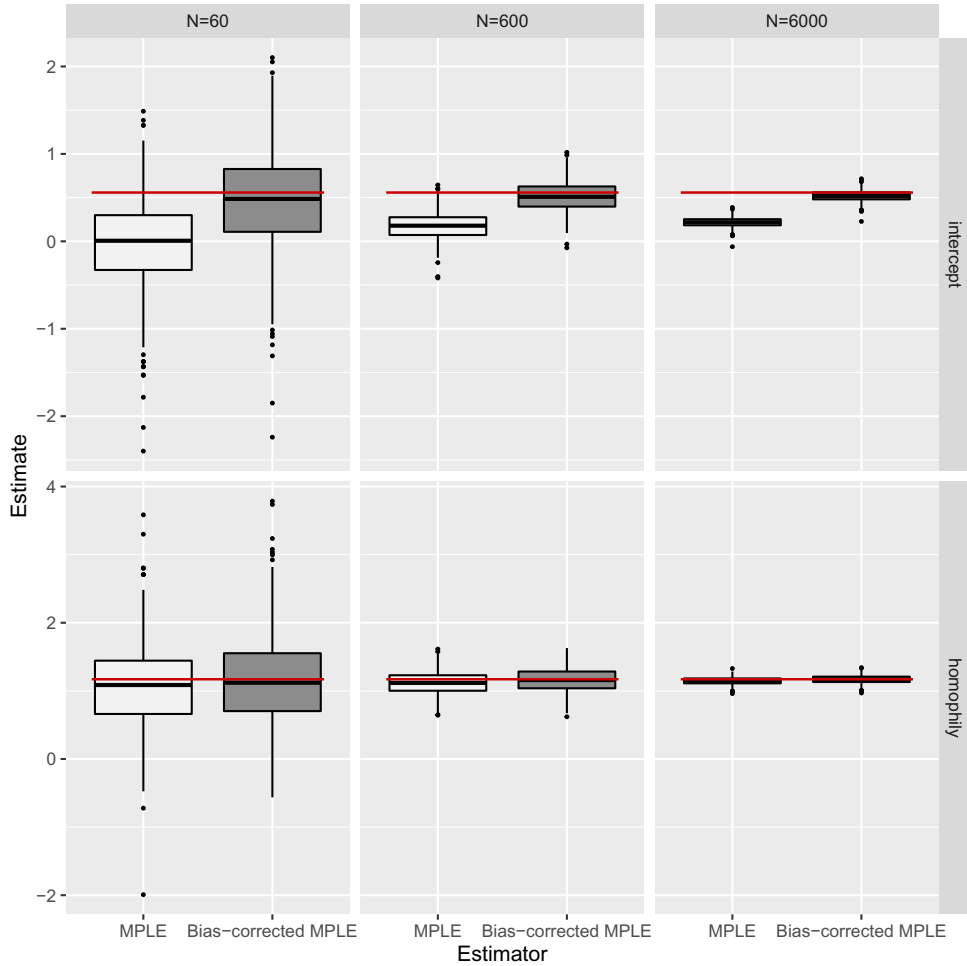

Figure 3. Simulation study II: Distribution of uniform homophily MLPLEs and bias-corrected MLPLEs for different population sizes N; 1,000 simulations.

parameters the difference in the mean MLPLE and mean bias-corrected MLPLE is greatest at $N = 60$. The bias-corrected MLPLEs have a slightly higher standard deviation than the non-bias-corrected MLPLEs, though the magnitude of this different decreases with population size N.

6.4 Simulation study III: increasing relative sample size

In simulation study III, we investigate the relationship between the sample size n_h and the bias of MLPLEs when fitting the revealed preferences model, as well as the impact of bias correction methodology on the estimates as sample size increases.

Figure 4 shows the distribution of parameter MLPLEs and bias-corrected MLPLEs at each value of n_h, while the medians and standard deviations of the estimates are given in Online Supplementary Material, Table 8. At all three values of n_h, the mean MLPLE estimate underestimates the true value. We see much less bias, though still a small amount (<0.05 units), in the estimates for the matching preferences at each education level. The variance of the MLPLEs decreases as the sample size increases.

After bias-corrected methods are used, the difference between the truth $\beta^{\mathrm{DH},0}$ and the bias-corrected MLPLEs becomes very small. As with simulation study II, we see that a coñsequence of bias correction is a slight increase in variance for $n_h = 600$ and 1200. At $n_h = 3,000$, however, the impact of bias

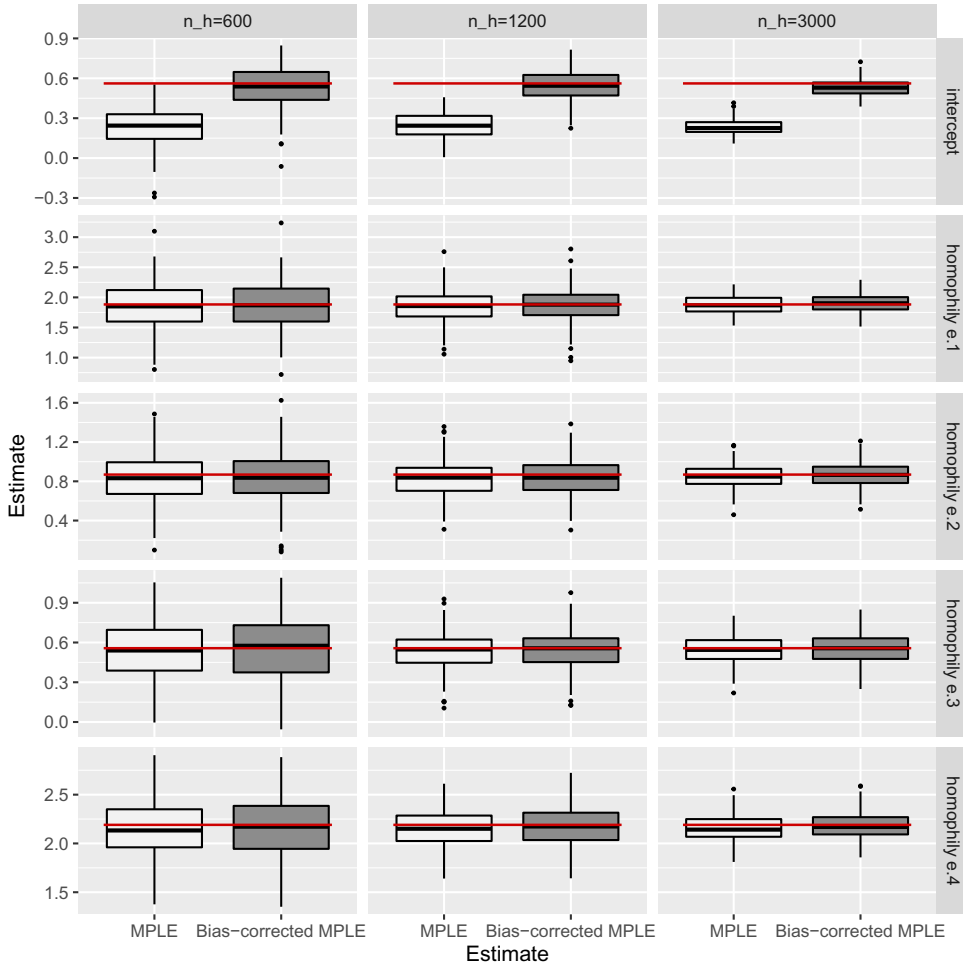

Figure 4. Simulation study III: Distribution of differential homophily MLPLEs and bias-corrected MLPLEs for different n_h, where N=6,000; 200 simulations.

correction on the variance of estimates is ambiguous. While the variance increases with bias correction for the intercept parameters and the parameters indicating homogamy on the education levels 2 (high school education) and 3 (some college), the variance of the parameters indicating homogamy at education levels 1 (less than high school) and 4 (college degree or higher) actually decreases.

We repeated this exercise at $N = 1,000$ and $n_h = 100, 200,$ and 500 (results not shown) and obtained results that were consistent with the earlier findings. Specifically, the MLPLEs showed some bias at all n_h, with bias in the intercept MLPLEs being much higher than in other parameters. The bias-corrected MLPLEs were closer estimates of the true $\beta^{\mathrm{DH},0}$. As n_h increased, the variance of both the MLPLEs and the bias-corrected MLPLEs decreased. In general we find that as long as the sample size is large enough to ensure non-zero entries in $\bar{c}$ are rare, the bias-corrected MLPLEs have high accuracy and improve in precision as n_h increases.

6.5 Confidence intervals and coverage probabilities

To supplement the findings in simulation study I.ii, we calculate 95% confidence intervals for bias-corrected MLPLEs based on samples from simulated populations of size $N = 300$ million, and we compare the empirical coverage rates of the true parameter values to the 95% threshold.

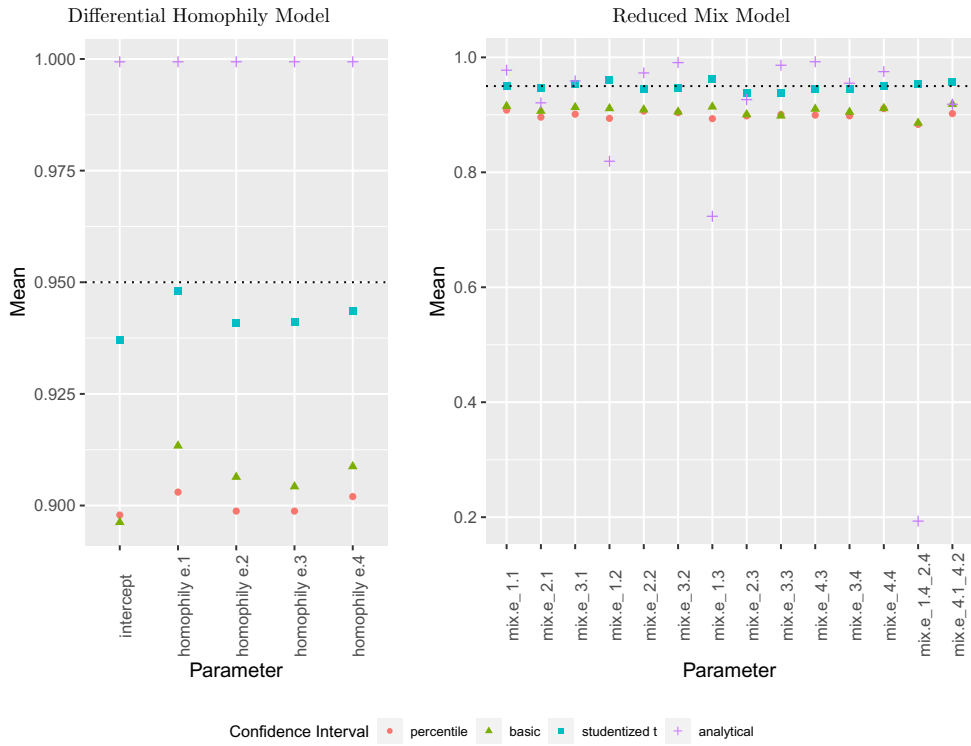

Figure 5. Mean empirical coverage probability by bootstrap confidence intervals for model parameters (40 sets of 200 simulations from Availability scenario $\mathcal{A}_1$).

To calculate empirical coverage rates, we simulate $S = 200$ samples from large large populations from availability $\mathcal{A}_1$. For each sample, we fit the reduced mix model and produce analytical 95% confidence intervals based on the approximated Hessian matrix, as detailed in Section 4.4. We additionally implement the basic, percentile, and modified studentized t bootstrap methods also discussed in Section 4.4 to construct empirical 95% confidence intervals. An illustration of the coverage results from a single set of 200 simulations are presented for selected parameters in Online Supplementary Material, Figures 6 and 7 in Appendix B.

The process of simulating 200 populations and constructing confidence intervals for each simulation was repeated 40 times, so that we observed an empirical coverage rate across 200 simulations 40 times. We show the mean coverage rates of the reduced mix model parameters using the various confidence intervals in the right-hand panel of Figure 5. The dotted black line at 0.95 denotes the 95% threshold we aim to achieve. The analytical confidence intervals appears to be the most volatile; across the 14 parameters estimated in the reduced mix model, the mean coverage rate of the analytical confidence intervals ranged from 19.3 to 99.2%. The three bootstrap confidence intervals have a more consistent performance; within each interval type, the range of the mean coverage rates across the parameters is about 2 percentage points. The basic and percentile bootstraps both display undercoverage, with mean coverage rates around 90% across parameters. The studentized t interval achieves mean coverage rates closest to the 95% target.

We pay special attention to the coverage rates for the $\beta_{1\text{ or }2,4}$ parameter. This parameter corresponds to a preference for couples with a female of education level 1 or 2 and a male of education level 4. As noted earlier, the number of couples of this type in the SIPP data and in the simulated samples was very small. The mean coverage rates of the percentile, basic, and analytical confidence intervals are all lowest for this parameter, likely because of the low count of such couples in the data. We note, however, that the performance of the studentized t interval does not appear to be affected by the low couple count. In fact, the mean coverage rate of the studentized t interval for $\beta_{1\text{ or }2,4}$ is 95.3%.

The coverage rates shown Figure 5 were produced based on populations simulated from the 'factual' availability scenario $\mathcal{A}_1$. We repeated the procedure to evaluate confidence interval coverages using populations simulations from the 'counterfactual' availability scenario $\mathcal{A}_2$ (results not shown. We found no evidence that the change in population availabilities impacted the coverage rates of the bootstrap confidence intervals.

We also repeated this process to evaluate the performance of confidence intervals for differential homophily model parameters. In this case, we found that the analytical confidence intervals were two to three times wider than the student t intervals and captured the true value 100% of the time for all parameters, indicating overcoverage. We again observed that the studentized t confidence intervals consistently achieved the highest coverage rate of the bootstrap procedures. The basic and percentile bootstrap 95% confidence intervals show slight undercoverage, falling between 89.6% and 91.3% coverage. A plot of mean coverage rates by analytical and bootstrap confidence intervals for the differential homophily model is provided in the left panel of Online Supplementary Material, Figure 5 under Appendix B. We show coverage results from a single set of 200 simulations for selected parameters in the differential homophily model in Online Supplementary Material, Figures 8 and 9 in Appendix B

7 Discussion

The ability to extract preferences separably from availabilities is a key feature of the revealed preferences model and methodology, which we propose in this paper. In simulation study I.i, we simulate a small population ($N = 6,000$) and run the Gale–Shapley algorithm to obtain a stable matching. Given statistics of the types of matchings, we are able to compute parameter estimates which are very close to the true values. We note that Logan (1996b) was able to show a similar result for his initial special case of the model.

In simulation study I.ii, we simulate a large population and obtain an approximate distribution of household types in a stable matching. We sample couples and individuals from this matching and then maximize (12) over the sample data to obtain parameter estimates, showing that the method accurately recovers true preference parameter values even under various different availabilities of prospective partners. In both simulation studies I.i and I.ii, the distribution of the parameter estimates appears Gaussian in most cases. The standard errors decrease when the population size is larger, as in simulation study I.ii.

When there are very few or none of a certain type of couple in the data, the total utility of such a pairing is estimated be negative infinity. As an example, we refer to the estimates of $\beta_{1\text{ or }2,4}$ in simulation study I.ii, shown in the first column of Figure 2. If we observed no couples in which a woman has education level 1 or 2 and the man has education level 4, then the parameter estimate for the utility model term indicating such a match is negative infinity. This artifact is a form of separation also seen for generalized linear models (Heinze & Schemper, 2002). The high concentration of parameter estimates for $\beta_{1\text{ or }2,4}$ under -6 correctly captures this and reflects the lower utility corresponding to such pairings.

For both availability scenarios $\mathcal{A}_1$ ('factual') and $\mathcal{A}_2$ ('counterfactual') under the differential homophily model, the standard errors of the estimates in simulation study I.ii are smaller than the corresponding values in Simulation study I.i (small population scenario). As in simulation study I.i, the distributions of the parameter estimates appear to follow a Gaussian distribution.

In simulation studies II and III, we investigated the performance of the revealed preference model under different population and sample sizes. In simulation study II we assumed access to population data.

We found that for different population sizes N, the bias-corrected MLPLEs provided accurate estimates of preference parameters with the variance of estimates decreasing inversely with N. This is a significant finding as the previous formulation of the model proposed by Menzel (2015) required n/N to be small. We show that even when $n/N = 1$, the bias-corrected MLPLEs obtained using bootstrap methods recover true preference for pairings. The bias-corrected MLPLEs are similarly effective in reduced estimate bias in simulation study III, in which we obtain samples of different sizes from populations of N=6,000 individuals. We note again that bias in the MLPLEs is mitigated through the bootstrap bias correction. Together, the findings of simulation studies II and III provide strong support for the use of bias-corrected MLPLEs to estimate

preferences in revealed preferences models and show that accurate estimates can be achieved for a wide range of n/N.

We also evaluate different methods of accounting for uncertainty in our estimates. Based on results in Section 6.5, we believe that the approximation of the Hessian matrix leads to volatile analytical confidence intervals which deviate from the threshold coverage rate of 95%. These confidence intervals are often too wide or narrow to be useful. We also find that among the three bootstrap based methods for producing confidence intervals, the mean coverage probabilities of the studentized t interval were the closest to 95%, while the percentile and basic method-based confidence intervals demonstrate slight undercoverage.

The revealed preferences model can be used to make inferences which are particularly useful in demographic studies. For example, the preference parameter estimates when we fit the reduced mix specification of the revealed preferences model to the 2008 SIPP data are given in column 3 of Online Supplementary Material, Table 4. The estimated utility of pairings in which both individuals have the same education level is substantially higher than it is for pairings where individuals have different education levels. Homophilous behavior is expected by researchers who study matching problems. It is also consistent with the findings of Logan et al. (2008), who presented results which implied a preference for homophily in race and religion in heterosexual marriages.

An important issue not addressed by this paper is the identification of the effective population that constitutes the market. A useful additional concept is that of awareness, that is, the set of people a person effectively chooses among. It is possible to model the probability that a person is aware of another as a function of observed characteristics of the individuals (e.g., geographic distance, age difference) (Menzel, 2015). Incorporating them requires a significant expansion of the model (for example, geographic distance is a continuous variable, requiring the integral version of the model). For a treatment of this, see Zhang (2022).

Conflict of interest: None declared.

Funding

We are grateful for support from the National Science Foundation BIGDATA: Applications program, grant NSF IIS-1546259, and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, population research infrastructure grants P2C-HD041041 and P2C-HD041022 and training grant T32-HD007545.

Data availability

An open-source R package implementing the methods developed in this paper, rpm, (Handcock et al., 2022), was used to do the simulation studies and analyze the case-studies. The package is available on CRAN (R Core Team, 2020).

The SIPP data used as the basis for the simulation studies described in Section 5 are available within the rpm package. Instructions on reproducing analysis are given at https://github.com/handcock/rpm.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series A* online.

References

Agresti A. (2012). *Categorical data analysis* (3rd ed.). Wiley.

Becker G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, *81*(4), 813–846. https://doi.org/10.1086/260084

Becker G. S. (1974). A theory of marriage: Part II. *Journal of political Economy*, *82*(2, Part 2), S11–S26. https://doi.org/10.1086/260287

Brien M. J. (1997). Racial differences in marriage and the role of marriage markets. *Journal of Human Resources*, *32*(4), 741–778. https://EconPapers.repec.org/RePEc:uwp:jhriss:v:32:y:1997:i:4:p:741-778. https://doi.org/10.2307/146427

Chiappori P.-A. (2020). The theory and empirics of the marriage market. *Annual Review of Economics*, *12*(1), 547–578. https://doi.org/10.1146/annurev-economics-012320-121610

Chiappori P.-A., Salanié B., & Weiss Y. (2017). Partner choice, investment in children, and the marital college premium. *American Economic Review*, *107*(8), 2109–2167. https://doi.org/10.1257/aer.20150154

Choo E., & Siow A. (2006). Who marries whom and why. *Journal of Political Economy*, *114*(1), 175–201. https://doi.org/10.1086/498585

Dagsvik J. K. (1994). Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica: Journal of the Econometric Society*, *62*(5), 1179–1205. https://doi.org/10.2307/2951512

Dagsvik J. K. (2000). Aggregation in matching markets. *International Economic Review*, *41*(1), 27–58. https://doi.org/10.1111/1468-2354.00054

Dagsvik J. K., Brunborg H., & Flaatten A. S. (2001). A behavioral two-sex marriage model. *Mathematical Population Studies*, *9*(2), 97–121. https://doi.org/10.1080/08898480109525498

Dupuy A., & Galichon A. (2014). Personality traits and the marriage market. *Journal of Political Economy*, *122*(6), 1271–1319. https://doi.org/10.1086/677191

Gale D., & Shapley L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, *69*(1), 9–15. http://www.jstor.org/stable/2312726. https://doi.org/10.1080/00029890.1962.11989827

Handcock M. S., Admiraal R. M., Yeung F. C., Jackson H. M., Rendall M. S., & Goyal S. (2022). ***rpm:*** Modeling of Revealed Preferences Matchings. University of California, Los Angeles, Los Angeles, CA. http://CRAN.R-project.org/package=rpm. R package version 0.70.

Hartmann W. M., & Hartwig R. E. (1996). Computing the Moore-Penrose inverse for the covariance matrix in constrained nonlinear estimation. *SIAM Journal on Optimization*, *6*(3), 727–747. https://doi.org/10.1137/S1052623494260794

Heinze G., & Schemper M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409–2419. https://doi.org/10.1002/sim.1047

Johnson S. G. (2020). *The NLopt nonlinear-optimization package*. http://github.com/stevengj/nlopt.

Kalmijn M. (1994). Assortative mating by cultural and economic occupational status. *American Journal of Sociology*, *100*(2), 422–452. https://doi.org/10.1086/230542

Kraft D. (1994). Algorithm 733: Tomp-fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software*, *20*(3), 262–281. https://doi.org/10.1145/192115.192124

Logan J. A. (1996a). Opportunity and choice in socially structured labor markets. *American Journal of Sociology*, *102*(1), 114–160. http://www.jstor.org/stable/2782189. https://doi.org/10.1086/230910

Logan J. A. (1996b). Opportunity and choice in socially structured labor markets. *American Journal of Sociology*, *102*(1), 114–160. http://www.jstor.org/stable/2782189. https://doi.org/10.1086/230910

Logan J. A., Hoff P. D., & Newton M. A. (2008). Two-sided estimation of mate preferences for similarities in age, education, and religion. *Journal of the American Statistical Association*, *103*(482), 559–569. https://doi.org/10.1198/016214507000000996

McCarthy P. J., & Snowden C. B. (1985). *The bootstrap and finite population sampling* (Vol. *95*). Vital and health statistics. Series 2, Data evaluation and methods research. https://stacks.cdc.gov/view/cdc/12908.

Menzel K. (2015). Large matching markets as two-sided demand systems. *Econometrica*, *83*(3), 897–941. https://doi.org/10.3982/ECTA12299

Pollak R. A. (1986). A reformulation of the two-sex problem. *Demography*, *23*(2), 247–259. https://doi.org/10.2307/2061619

Pollard J. H. (1997). Modelling the interaction between the sexes. *Mathematical and Computer Modelling*, *26*(6), 11–24. https://doi.org/10.1016/S0895-7177(97)00166-0

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Roth A. E., & Sotomayor M. A. O. (1990). *Two-sided matching: A study in game-theoretic modeling and analysis*. Econometric Society Monographs. Cambridge University Press.

Schoen R. (1981). The harmonic mean as the basis of a realistic two-sex marriage model. *Demography*, *18*(2), 201–216. https://doi.org/10.2307/2061093

Schwartz C. R., & Mare R. D. (2005). Trends in educational assortative marriage from 1940 to 2003. *Demography*, *42*(4), 621–646. https://doi.org/10.1353/dem.2005.0036

Shao J., & Tu D. (1995). *The jackknife and bootstrap*. Springer Series in Statistics. Springer New York. https://books.google.com/books?id=VO3SBwAAQBAJ.

U.S. Bureau of the Census (2020). 2008 Survey of Income and Program Participation (SIPP). https://www.census.gov/programs-surveys/sipp/data/datasets.2008.html.

Yeung F. C. (2019). *Statistical revealed preference models for bipartite networks [Ph.D. thesis]*. University of California at Los Angeles. https://escholarship.org/uc/item/5ct600k6.

Zhang X. (2022). *An awareness model for a two-sided matching market [Master's thesis]*. University of California at Los Angeles. https://escholarship.org/uc/item/7nm7d0ft.

A. Supplementary Tables

Table 3: Medians and standard deviations (SDs) of differential homophily model bias corrected MLPLEs in simulation study I.i (1,000 simulations, $N = 6,000$)

Parameter	Truth	Availability			
		$\mathcal{A}_1$		$\mathcal{A}_2$	
	$\underset{\sim}{\beta}^{\mathrm{DH},0}$	Median	SD	Median	SD
intercept	-3.439	-3.435	0.136	-3.425	0.133
homophily e.1	1.883	1.887	0.391	1.879	0.332
homophily e.2	0.868	0.886	0.310	0.875	0.290
homophily e.3	0.557	0.561	0.238	0.558	0.256
homophily e.4	2.191	2.198	0.243	2.194	0.308

The homophily t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

Table 4: Medians and standard deviations (SDs) of reduced mix model bias corrected MLPLEs in simulation study I.i (1,000 simulations, $N = 6,000$)

Education Parameter		Truth	Availability			
			$\mathcal{A}_1$		$\mathcal{A}_2$	
Female	Male	$\underset{\sim}{\beta}^{\mathrm{RM},0}$	Median	SD	Median	SD
1	1	-1.572	-1.565	0.401	-1.585	0.330
2	1	-2.877	-2.854	0.433	-2.837	0.389
3	1	-3.419	-3.374	0.477	-3.377	0.422
1	2	-3.209	-3.070	0.496	-3.097	0.406
2	2	-2.570	-2.561	0.277	-2.561	0.270
3	2	-3.256	-3.226	0.283	-3.247	0.269
1	3	-3.695	-3.619	0.627	-3.524	0.672
2	3	-3.348	-3.311	0.348	-3.328	0.306
3	3	-2.888	-2.867	0.216	-2.855	0.217
4	3	-3.211	-3.207	0.330	-3.186	0.397
3	4	-3.387	-3.365	0.372	-3.311	0.425
4	4	-1.270	-1.249	0.190	-1.257	0.274
1 or 2	4	-5.082	-5.139	4.128	-4.838	4.349
4	1 or 2	-3.883	-3.829	0.450	-3.839	0.441

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors

Table 5: Medians and standard deviations (SDs) of differential homophily model bias corrected MLPLEs in simulation study I.ii (1,000 simulations, $n_h = 21,077$)

Parameter	Truth	Availability			
		$\mathcal{A}_1$		$\mathcal{A}_2$	
	$\underset{\sim}{\beta}^{\mathrm{DH},0}$	Median	SD	Median	SD
intercept	-3.439	-3.437	0.072	-3.435	0.064
homophily e.1	1.883	1.889	0.180	1.875	0.181
homophily e.2	0.868	0.854	0.156	0.864	0.145
homophily e.3	0.557	0.544	0.127	0.553	0.127
homophily e.4	2.191	2.200	0.115	2.195	0.149

The homophily e.t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

Table 6: Medians and standard deviations (SDs) of reduced mix model bias corrected MLPLEs in simulation study I.ii (1,000 simulations, $N = 21,077$)

Education Parameter		Truth	Availability			
			$\mathcal{A}_1$		$\mathcal{A}_2$	
Female	Male	$\underset{\sim}{\beta}^{\mathrm{RM},0}$	Median	SD	Median	SD
1	1	-1.572	-1.585	0.180	-1.563	0.167
2	1	-2.877	-2.892	0.238	-2.873	0.207
3	1	-3.419	-3.421	0.230	-3.422	0.198
1	2	-3.209	-3.219	0.297	-3.210	0.273
2	2	-2.570	-2.584	0.146	-2.576	0.137
3	2	-3.256	-3.273	0.157	-3.261	0.137
1	3	-3.695	-3.745	0.335	-3.740	0.287
2	3	-3.348	-3.360	0.185	-3.343	0.178
3	3	-2.888	-2.893	0.115	-2.884	0.104
4	3	-3.211	-3.212	0.164	-3.230	0.206
3	4	-3.387	-3.388	0.200	-3.378	0.256
4	4	-1.270	-1.271	0.092	-1.264	0.128
1 or 2	4	-5.082	-5.069	0.410	-5.047	0.529
4	1 or 2	-3.883	-3.889	0.230	-3.891	0.240

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors

Table 7: Simulation study II: MLPLEs and bias corrected MLPLEs for different N with Availability $\mathcal{A}_1$ and uniform homophily preferences (1,000 simulations)

Parameter	Truth $\underset{\sim}{\beta}^{\mathrm{UH},0}$	Bias Correction	$N = 60$		$N = 600$		$N = 6,000$	
			Median	SD	Median	SD	Median	SD
intercept	0.558	No	0.007	0.465	0.179	0.151	0.218	0.052
		Yes	0.485	0.533	0.509	0.171	0.520	0.059
homophily	1.170	No	1.086	0.580	1.117	0.168	1.147	0.053
		Yes	1.120	0.630	1.159	0.182	1.166	0.058

Table 8: Simulation study III, MLPLEs for different n_h with Availability $\mathcal{A}_1$ and differential homophily preferences at $N = 6,000$ (200 simulations)

Parameter	Truth $\underset{\sim}{\beta}^{\mathrm{DH},0}$	Bias Correction	$n_h = 600$		$n_h = 1,200$		$n_h = 3,000$	
			Median	SD	Median	SD	Median	SD
intercept	0.561	No	0.223	0.142	0.227	0.092	0.218	0.057
		Yes	0.531	0.156	0.537	0.107	0.519	0.064
homophily e.1	1.883	No	1.859	0.363	1.848	0.262	1.844	0.153
		Yes	1.891	0.388	1.884	0.294	1.886	0.170
homophily e.2	0.868	No	0.872	0.262	0.861	0.186	0.870	0.121
		Yes	0.866	0.295	0.846	0.199	0.872	0.129
homophily e.3	0.557	No	0.567	0.216	0.569	0.141	0.581	0.087
		Yes	0.551	0.242	0.541	0.159	0.564	0.097
homophily e.4	2.191	No	2.122	0.269	2.106	0.178	2.110	0.119
		Yes	2.206	0.281	2.173	0.197	2.193	0.126

Education level codes: 1 =<high school, 2 =high school, 3 =some college, 4 =≥bachelors
The homophily e.t parameter is the coefficient of an indicator which equals 1 if both partners have education level t.

B. Confidence intervals from 200 simulations

Figures 6 and 7 show the analytical confidence intervals and the empirical bootstrap confidence intervals produced over 200 simulations for the $\beta_{4,4}$ and $\beta_{1 \text{ or } 2,4}$ parameters in the reduced mix model. These figures coincide with the simulation results related to uncertainty estimates described in Section 6.5. The horizontal axis gives the simulation index, and the vertical axis shows the range of the interval. The solid point at the center of each interval indicates the parameter estimate in the bootstrapped sample at that index. The horizontal red line in each plot represents the true parameter value, and intervals in blue are those which failed to include the true value. We provide the empirical coverage rate of the parameter for each method of confidence interval in the top-right corner of the plots.

The first three panels of Figure 6 show the 200 confidence intervals for $\beta_{4,4}$ produced by each of the three bootstrapping methods which were described in Section 4.4. The three methods for constructing the bootstrapped confidence intervals produce very similar results, with the basic bootstrap method achieving 95% coverage and the percentile and modified studentized t methods achieving 96% coverage. Furthermore, the confidence intervals appear to have similar lengths across the three methods.The bottom-right panel

shows the analytical confidence intervals produced for $\beta_{4,4}$ based on the same simulated populations. We note that the analytical 95% confidence intervals only achieve 83% coverage in this set of simulations, indicating undercoverage.

The performances of the three bootstraps methods are more varied more when evaluating the $\beta_{1 \text{ or } 2,4}$ parameter. The modified studentized t and the percentile bootstrap confidence intervals achieve a coverage rate of 88% and 86.5%, respectively, while the basic bootstrap intervals achieve much lower coverage of 78.5%. Furthermore, the percentile and studentized t methods produce intervals which are generally wider than those produced by the basic bootstrap method. The analytical confidence intervals in the bottom-right panel of the figure are so narrow that few of them capture the true value, resulting in a poor coverage rate of 10.5%.

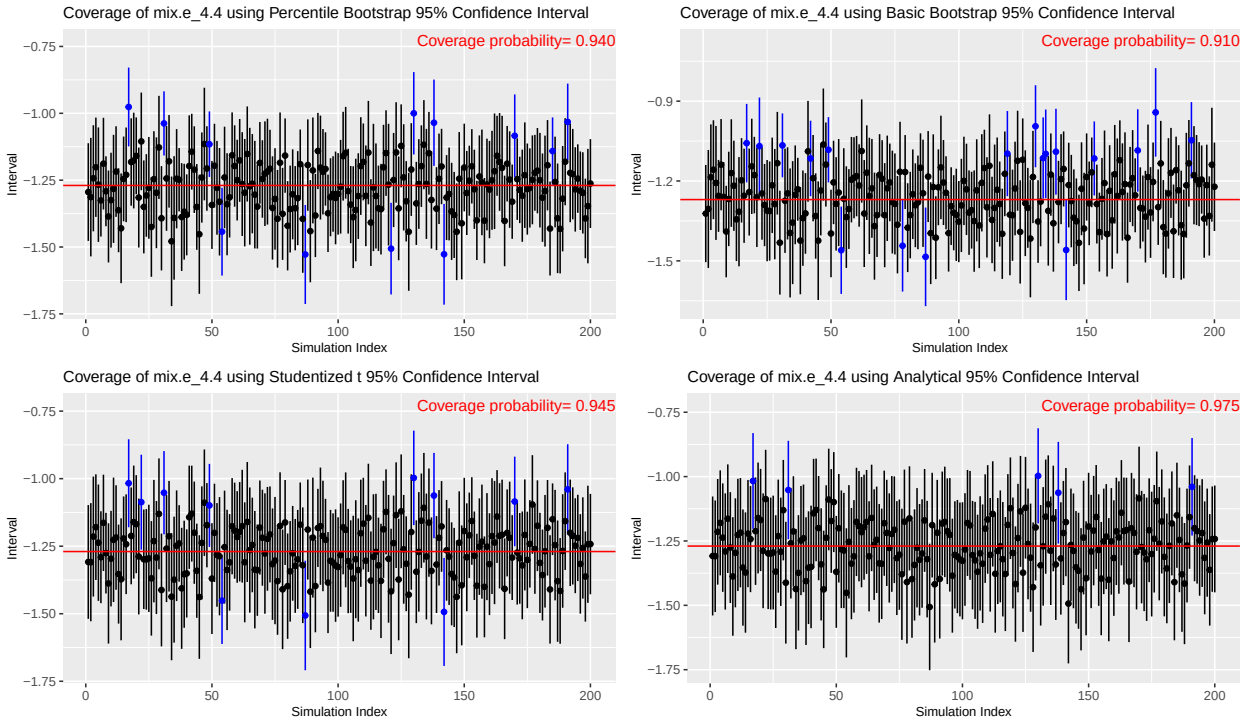

Fig. 6: Coverage of $\beta_{4,4}$ in reduced mix model over 200 simulations

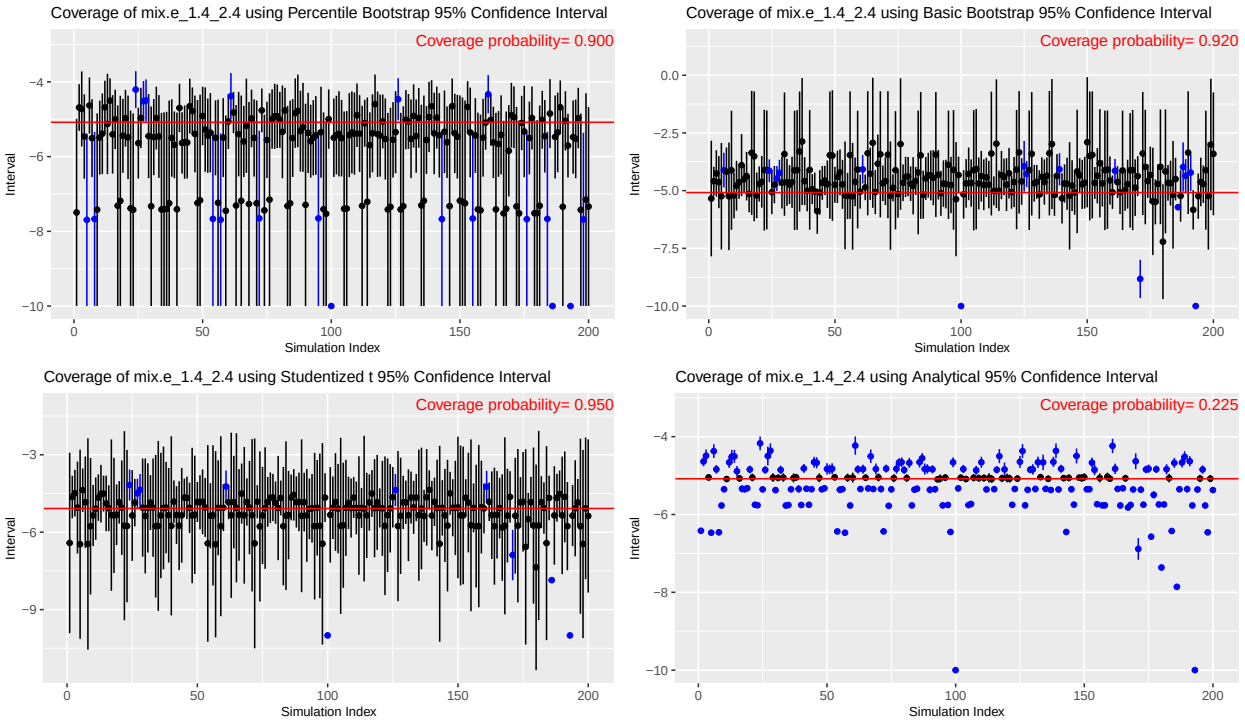

Fig. 7: Coverage of $\beta_{1 \text{ or } 2,4}$ in reduced mix model over 200 simulations

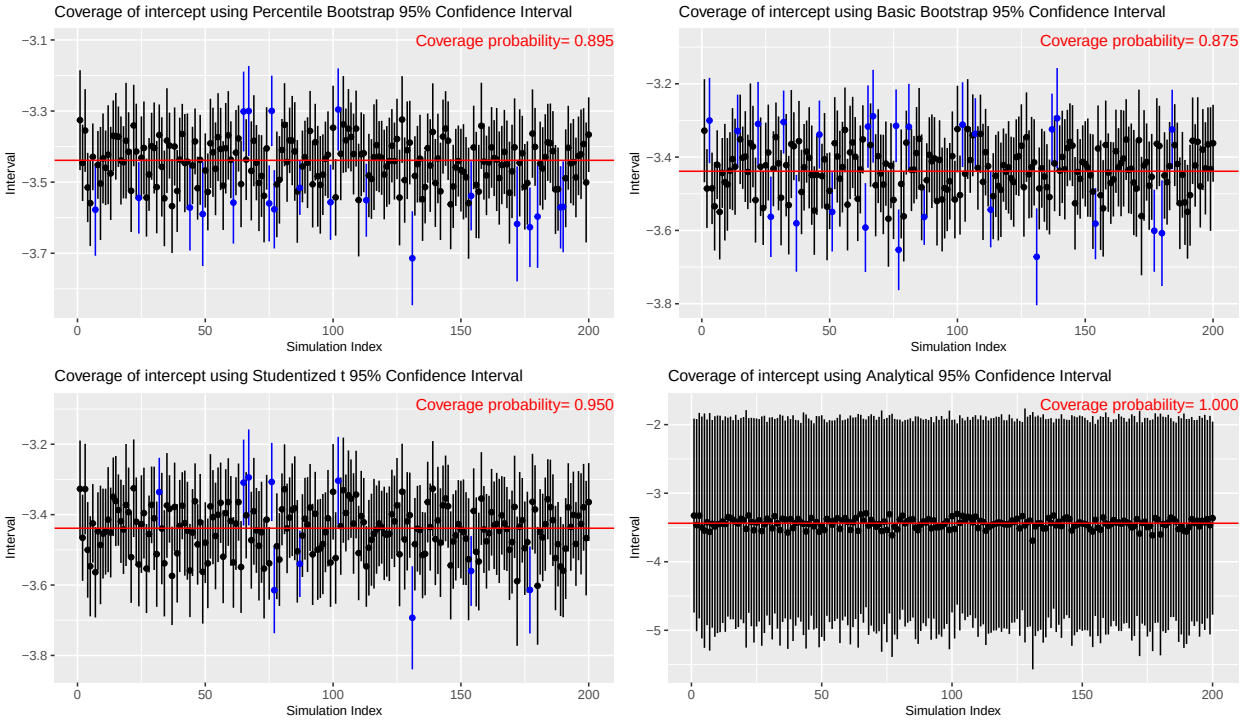

Fig. 8: Coverage of intercept β_0 in differential homophily model over 200 simulations

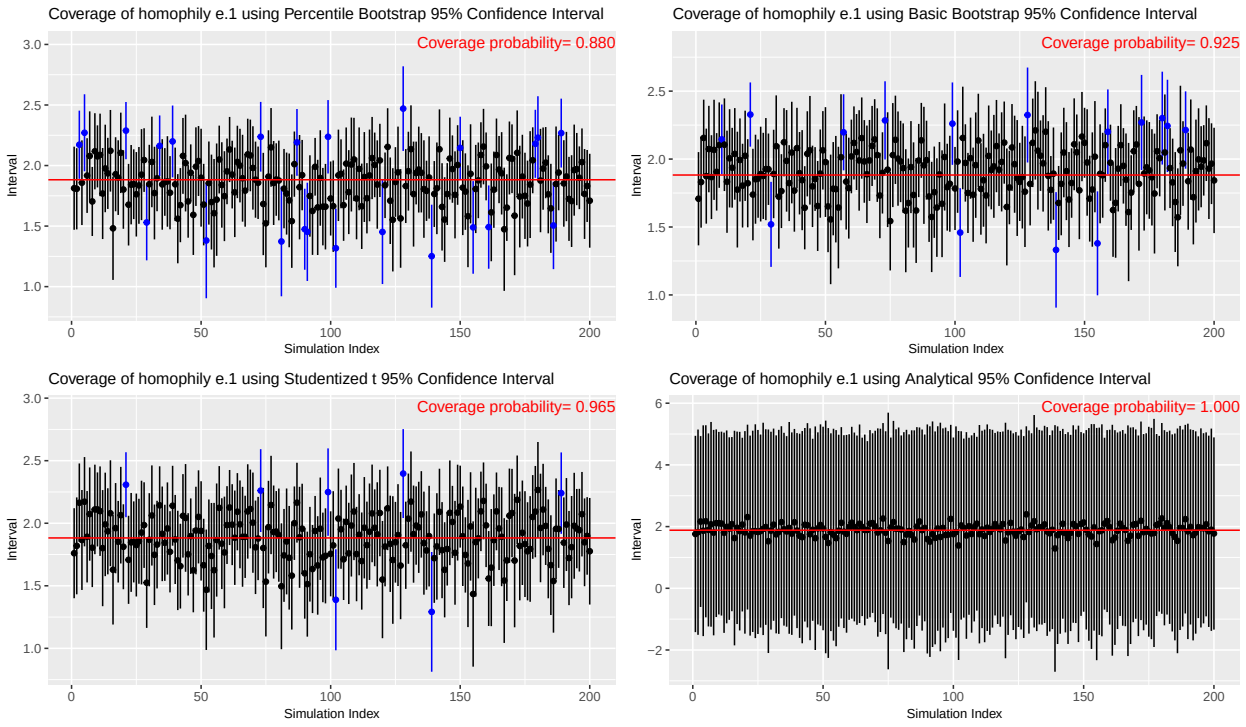

Fig. 9: Coverage of $\beta_{\text{homophily e.1}}$in differential homophily model over 200 simulations