Revealing Gender-Specific Costs of STEM in an Extended Roy Model of Major Choice

Marc Henry
Romuald Méango
Ismaël Mourifié

Working Paper

2020-035

05/2020

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

REVEALING GENDER-SPECIFIC COSTS OF STEM IN AN EXTENDED ROY MODEL OF MAJOR CHOICE

MARC HENRY, ROMUALD MÉANGO, AND ISMAËL MOURIFIÉ

Abstract. We derive sharp bounds on the non consumption utility component in an extended Roy model of sector selection. We interpret this non consumption utility component as a compensating wage differential. The bounds are derived under the assumption that potential wages in each sector are (jointly) stochastically monotone with respect to an observed selection shifter. The lower bound can also be interpreted as the minimum cost subsidy necessary to change sector choices and make them observationally indistinguishable from choices made under the classical Roy model of sorting on potential wages only. The research is motivated by the analysis of women's choice of university major and their underrepresentation in mathematics intensive fields. With data from a German graduate survey, and using the proportion of women on the STEM faculty at the time of major choice as our selection shifter, we find high costs of choosing the STEM sector for women from the former West Germany, especially for low realized incomes and low proportion of women on the STEM faculty, interpreted as a scarce presence of role models.

Keywords: Roy model, partial identification, stochastic monotonicity, women in STEM.

JEL subject classification: C31, C34, I21, J24

INTRODUCTION

The underrepresentation of women in mathematics intensive education and occupations is of concern to academics and policy makers, especially since it is perceived as one of the main drivers of the gender wage gap (see Daymont and Andrisani [1984], Zafar [2013] and Sloane et al. [2019]). A survey of the scholarship on the issue can be found in Kahn and Ginther [2017]. We examine the issue from the point of view of college major choice. The literature on major choice, reviewed in Altonji et al. [2016], emphasizes the importance of factors beyond expected earnings in the determination of major choice: see Delavande and Zafar [2019] and Arcidiacono et al. [2019] on the role of non pecuniary factors and outcomes,

The first version is of August 21, 2019. The present version is of May 14, 2020. This research was supported by SSHRC Grant 435-2018-1273 and Leibniz Association Grant SAW-2012-ifo-3. The research was conducted in part, while Ismaël Mourifié was visiting the Becker Friedman Institute at the University of Chicago. He thanks his hosts for their hospitality and support. The authors also acknowledge the excellent research assistance of Ivan Sidorov, as well as helpful comments from Chris Dobronyi, Jonathan Eaton, Jim Heckman, Thomas Russell, Azeem Shaikh and seminar audiences at Cambridge, Chicago, Rice, Texas A&M, Vanderbilt and York. Correspondence address: Department of Economics, Max Gluskin House, University of Toronto, 150 St. George St., Toronto, Ontario M5S 3G7, Canada

Kaplan and Schulhofer-Wohl [2018], Wiswal and Zafar [2015, 2018] on the importance of job amenities, Cunha et al. [2005], Eisenhauer et al. [2015a] on the importance of psychic costs. Mas and Pallais [2017], Wiswal and Zafar [2018] and Mourifié et al. [2020] find gender specific response to non pecuniary factors in the valuation and choice of occupations. This motivates the current theoretical analysis of (partial) identification of gender-specific non pecuniary drivers of major choice within what is generally known in the literature as the extended Roy model (see Heckman and Vytlacil [1999] for details on model genealogy and attribution).

Bayer et al. [2011], d'Haultfœuille and Maurel [2013] and Eisenhauer et al. [2015b] analyze the extended Roy model and provide competing strategies to identify the non pecuniary driver of choice under separability, support assumptions and exclusion restrictions. Saltiel [2018] looks specifically at women's college major choice through the lens of a multi-stage extended Roy model inspired by Heckman et al. [2016, 2018]. We eschew such identifying assumptions and derive sharp bounds on the non pecuniary driver of choice based on minimal assumptions. The model allows for multiple interpretations of the non pecuniary component, including anticipated lack of support for women in mathematics intensive education and occupations, mathematics talent misperception, as well as gender stereotyping of occupations.

The primitives of the model are potential labor market outcomes (Y_0, Y_1), such as wages, in both sectors 0 (non mathematics intensive college major) and 1 (mathematics intensive college major). From now on, Sector 1 will be referred to as the STEM sector, and Sector 0 as the non STEM sector, where STEM stands for Science, Technology, Engineering and Mathematics. The realized outcome Y is equal to the potential outcome in the selected sector, so that $Y = Y_1 D + Y_0(1 - D)$, where D is the observed random sector selection indicator. In the classical Roy model, $Y = \max\{Y_0, Y_1\}$. In the extended Roy model we consider, choices are affected by potential outcomes and a vector of observable variables W. Individuals choose Sector 1 when $Y_1 - C(Y_1, W) > Y_0$, and Sector 0 when $Y_1 - C(Y_1, W) < Y_0$, where C is a relative cost of choosing Sector 1 relative to Sector 0. Since we are looking at reasons women are driven away from STEM majors, we will restrict the relative cost function C to be non negative.

In allowing for the non pecuniary cost function C to depend on potential outcome Y_1, we depart from previous work on the identification of extended Roy models, such as Bayer et al. [2011] and d'Haultfœuille and Maurel [2013]. Our specification can be rationalized by the maximization of utilities that may not be quasi-linear. Hence, we allow the marginal rate of substitution between amenities and income to be a function of the latter. In the context of labor sector choice, we expect the marginal rate of substitution between wage and a female friendly environment to vary according to the skill requirements of the job, hence the wage.

The objective of this analysis is to characterize the collection of cost functions that rationalize observed choices. We will refer to this set as the *identified set*. We interpret the non pecuniary cost function as a wage differential that compensates for the real or perceived disamenities of the STEM sector for women. Hence the identified set characterizes our ability to draw inference on compensating wage differentials in this context. An alternative interpretation of the lower bound of the identified set, which we derive in closed form, is as the minimum subsidy for Sector 1 to make choices observationally indistinguishable from choices that conform to the classical Roy model of sector selection based only on potential outcomes.

Without additional constraints on the model environment, zero cost can always rationalize choices, since potential outcomes $Y_0 = Y_1 := Y$ could have given rise to the observed data. Hence, the identified set is of little use in this case. We therefore assume that a subvector Z of the observable variables $W = (X, Z)$ can only affect potential outcomes and selection relevant costs in one direction. Hence Z is a vector of observable variables that have a monotonic effect on potential outcomes in the following sense: the joint distribution of (Y_0, Y_1) conditional on $X = x, Z = z$ is (weakly) increasing in first order stochastic dominance when z increases (in the componentwise order) for all x. This condition is a generalization of the Manski and Pepper [2000] monotone instrumental variable assumption, and is also related to the stochastic monotonicity constraint in Blundell et al. [2007]. For ease of notation, we will fix $X = x$ for the whole analysis and omit it from notation. However, it is important to remember that the cost function C can be a function of individual and sector characteristics contained in X. In particular, we expect it to depend on gender. In addition, the stochastic monotonicity assumption is expected to hold conditionally on individual and sector characteristics contained in X, particularly gender.

Combining stochastic monotonicity of potential outcomes with monotonicity of the cost function provides a set of sufficient conditions for our assumption. We argue that factors that positively impact the formation of cognitive and non cognitive skills, while reducing perceived costs of the STEM sector, are likely to satisfy this set of sufficient conditions. Maternal educational attainment and the proportion of female role models (faculty, alumni, invited speakers) are prime candidates (see Breda et al. [2018] and Riegle-Crumb and Moore [2014]).

Under the stochastic monotonicity constraint, we derive a characterization of the set of cost functions that rationalize the data using moment inequalities. We also derive the bounds of this set of cost functions in closed form. To derive the lower bound, we use the lower monotone envelope of realized outcomes, a notion related to but distinct from the monotonization of Chernozhukov et al. [2009]. We then derive testable implications of the model. We also provide similar results for an extension of the model where individuals base their decisions on their expectations of potential outcomes at the time of sector choice. Finally, we derive bounds on the distribution of non pecuniary costs in a generalized version of the model, where these costs may depend on unobservable characteristics of the individuals, beyond their potential outcomes.

Mourifié et al. [2020] document rejection of the Roy model of sorting on labor market outcomes for women in the sample of German graduates in the 2005 and 2009 graduating cohorts of the German DZHW Graduate Survey (see Baillet et al. [2017]). We therefore analyze women's choice of major within the framework of our extended Roy model and derive confidence regions for the minimum cost of STEM that rationalize choices. The substantive assumptions we make are that the proportion of women on the STEM faculty in the individual's region at the time of major choice (as a proxy for the presence of role models and better amenities for women) only affects potential labor market outcomes of women graduates positively. We find very significant costs of choosing STEM fields for German women from the former Federal Republic. We observe that among the 2009 graduation cohort, 2 out of 10 have minimum cost of STEM larger than 20% of their income, and 1 out of 10 have minimum cost larger than 40% of their income. The costs are particularly pronounced for women in the lower quartile of the income distribution and for women whose region had low rates of feminization of the STEM faculty at the time of major choice.

Outline. The next section presents the extended Roy model and the main identification results. Section 2 discusses structural underpinning of the model. Section 3 presents extensions to imperfect foresight and the generalized Roy model. Section 4 applies the methodology to women's major choices in Germany. The last section concludes. Proofs of the main results are collected in the appendix.

1. ANALYTICAL FRAMEWORK

1.1. Extended Roy Model of Major Choice.

We adopt the framework of the potential outcomes model $Y = Y_1 D + Y_0(1-D)$. Y, with support $\mathcal{Y} \subseteq [\underline{b}, \infty)$, is an observed scalar outcome (with $\underline{b} \in \mathbb{R}$ and $\underline{b} = 0$ in most cases of interest), D is an observed selection indicator, which takes value 1 if Sector 1 is chosen, and 0 if Sector 0 is chosen, and Y_1, Y_0, are unobserved potential outcomes. In the context of major choice, the outcome of interest will be income in the year following graduation. Sector 1 will consist of all STEM majors and Sector 0, the rest. Decision makers choose their sector of activity based on the realizations of Y_0 and Y_1, and a vector of observed exogenous characteristics Z with support $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$. In the context of women's major choice, the vector Z will be either $Z = (Z_0, Z_1)$

or simply $Z = Z_1$, where Z_d is the proportion of women on the faculty in Sector d in the individual's region or prospective university at the time of choice. Additional observed exogenous covariates will be omitted from the notation. In the context of major choice in Germany, these include gender, visible minority status and a dummy for residence in the former East Germany, as a less affluent region.

Since Y, D and Z are observed, the distribution π of (Y, D, Z) is directly identified from the data. We call Π the set of admissible data generating processes. Unless otherwise specified, Π is the set of probability distributions on $\mathcal{Y} \times \{0, 1\} \times \mathcal{Z}$. We summarize the model with the following assumptions.

Assumption 1 (Potential outcomes)**.** *Observed outcomes are the realizations of a random variable Y satisfying $Y = Y_1D + Y_0(1 - D)$, where (Y_0, Y_1) is a pair of possibly dependent unobserved random variables with support $\mathcal{Y} \subseteq [\underline{b}, \infty)$, and D is an observed indicator variable.*

The original Roy model posits sector selection based only on the comparison of potential outcomes, so that $Y = \max\{Y_0, Y_1\}$. Given our focus on the underrepresentation of women in STEM fields, and rejections of the original Roy model selection rule in Mourifié et al. [2020], we entertain the possibility that other factors affect sector selection in favor of the non STEM sector. In the context of women's major choices, C might be a gender-specific cost of studying in a STEM field or a gender-specific cost of working in the STEM sector. The former may be the result of the lack of support for female students, or a fear of mathematics carried over from schooling (see Xie et al. [2015] for a survey of the sociological literature on the subject). The latter may be related to family-friendliness of employment outside STEM.

Hence, in our model, the decision by the individual is based on the comparison between $Y_1 - C$ and Y_0, where Y_1 is the potential outcome (wage) in the STEM sector and C is an unobserved possibly non pecuniary cost of STEM studies and professional activities.

Assumption 2 (Selection)**.** *The selection indicator satisfies $Y_1 - C(Y_1, Z) > Y_0 \Rightarrow D = 1$, $Y_1 - C(Y_1, Z) < Y_0 \Rightarrow D = 0$, where C is a measurable function on $\mathcal{Y} \times \mathcal{Z}$ and $y \mapsto y - C(y, z)$ is invertible for all $z \in \mathcal{Z}$.*

We focus on the case, where cost C is a function of the potential outcome Y_1 and another vector of observable variables Z (observable in both sectors), but we later also discuss the case, where the cost may depend on other unobservables as well. The dependence of the cost function on potential outcomes distinguishes the assumption below from other extended Roy models cited above. As we show in Section 2 in the discussion of the selection rule's possible structural underpinnings, it allows us to go beyond the comparison of quasi-linear utilities.

We expect the costs incurred by women studying or working in the mathematics intensive sector to be less pronounced for women with higher mathematics ability, hence decreasing in Y_1. In the following, we consider the weaker assumption that the rate of increase of such costs with Y_1 be less than 1.

Assumption 3 (Shape)**.**

(1) *The function $y \mapsto y - C(y,z)$ is increasing for each $z \in \mathcal{Z}$.*
(2) *The function C is non negative.*
(3) *The function $z \mapsto C(y,z)$ is non increasing (in the componentwise ordering) for each $y \in \mathcal{Y}$.*

In the following, Assumption 3(2) will be maintained throughout. Assumption 3(3) will only appear in a set of sufficient conditions for Assumption 4 in Lemma 1 of Section 1.2.

1.2. **Stochastically Monotone Instrumental Variables.** Without further restrictions, the trivial choice $C = 0$ would rationalize any data generating process under Assumptions 1 and 2. Indeed, for any given random vector (Y, D) generating observations, the choices $Y_0 = Y_1 := Y$ and $C = 0$ trivially satisfy Assumptions 1 and 2. To restore testability, we introduce a shape restriction on the joint distribution of potential outcomes (Y_0, Y_1). A traditional approach to restoring testability without parametric restrictions is to allow some observed covariates to affect sector selection only. However, such restrictions are difficult to justify in the context of college major choice. Parental educational attainment is likely to be correlated with unobserved parental cognitive and non cognitive investments in their children (see Card [2001]). Distance to college and other instruments designed to affect educational attainment choices are not suitable for major choice. Other variables that are very relevant to a woman's major choice, such as the proportion of women on the faculty, are expected to affect potential outcomes for women as well as their choices. We resort instead to a weaker instrumental notion, where the instrument Z may affect the joint distribution of potential outcomes, but only in one direction, in terms of first order stochastic ordering. For more details on the multivariate first order stochastic ordering, refer to Shaked and Shanthikumar [2007], Section 6.B.1. When comparing vectors, "$\geq$" denotes the componentwise partial order.

Definition 1 (First Order Stochastic Dominance)**.** A distribution F_1 on $\mathbb{R}^k$ is said to be *first order stochastically dominated* by a distribution F_2 if there exists random vectors X_1 with distribution F_1 and X_2 with distribution F_2 such that $X_2 \geq X_1$. By extension, a random vector with distribution F_2 is also said to stochastically dominate a random vector with distribution F_1.

Replacing traditional exclusion and independence restrictions with a monotonicity restriction allows us to restore testability of the Roy model with variables such as parental educational attainment and the proportion of women on the faculty. These variables are likely to affect potential outcomes as well as choices. However, it can be argued that the effect on potential outcomes can only be positive, as additional unobserved parental investment and additional support and role models for women in STEM are not expected to have a negative effect on their future labor market prospects. The mathematical formalization of this monotonicity notion is stochastic monotonicity.

Definition 2 (Stochastic Monotonicity)**.** A random vector X on $\mathbb{R}^k$ is said to be *stochastically monotone* (non decreasing) with respect to a random vector Z on $\mathbb{R}^l$ if the conditional distribution of X given $Z = z_1$ is first order stochastically dominated by the conditional distribution of X given $Z = z_2$, for any pair (z_1, z_2) of elements of the support of Z such that $z_2 \geq z_1$.

Using Definition 2, we can now formally state our instrumental constraint, we call *stochastically monotone instrumental variable* constraint, hereafter SMIV.

Assumption 4 (SMIV)**.** *The random vector* $(Y_0, Y_1 - C(Y_1, Z))$ *is stochastically monotonically non decreasing with respect to* Z*.*

A simple implication of Assumption 4 is the following.

Assumption $4'$. *The random vector* $(Y_0, Y_1 - C(Y_1, Z))$ *is such that for each* y*,* $\mathbb{P}(Y_0 \leq y, Y_1 - C(Y_1, Z) \leq y | Z = z)$ *is monotonically non increasing in* z*.*

Sufficient conditions include the case, where the vector of potential outcomes (Y_0, Y_1) is stochastically monotone with respect to Z, and the cost function is decreasing in z. In the context of women's major choice, where the costs are associated with lack of support for female students in STEM fields, we expect such costs to decrease with the presence of female faculty and role models in STEM fields and in the educational level of the student's mother, hence the assumption that such costs are decreasing when the values of the chosen sector selection variables Z increase. This is not necessary for Assumption 4, but combined with the assumption that more female faculty or a higher educational attainment of the mother cannot hurt a woman's prospects, it yields the following set of sufficient conditions for Assumption 4.

Lemma 1 (Sufficient conditions)**.** If Assumptions 3(1) and 3(3) hold and (Y_0, Y_1) is stochastically monotone non decreasing with respect to Z, then Assumption 4 (SMIV) holds.

The sufficient conditions of Lemma 1 includes the SMIV assumption in Mourifié et al. [2020], which is inspired by the monotone instrumental variable (MIV) of Manski and Pepper

[2000]. Unlike MIV, it places restrictions on the joint distribution of potential outcomes, as opposed to the marginals only, and drives our characterization of the model's empirical content in Theorem 1. Section 2 discusses structural underpinning for the selection and the shape assumptions.

1.3. **Characterization of the Cost Function.** The object of inference is the hidden cost function C. In particular, we seek a minimal function C that rationalizes the data under the extended Roy model of Assumptions 1-4. From a policy point of view in the context of women's major choice, this minimal cost function can be interpreted as the smallest subsidy necessary to offset the psychological costs from lack of support or fear of mathematics, or offset the lack of support for child-care in the STEM sector, and restore efficiency, i.e. restore choices that are observationally indistinguishable from choices based on the pure Roy model, where the sector is selected to maximize potential outcomes.

1.3.1. *Identified Set.* Before turning to the identification of the minimal cost function, we characterize the set of all cost functions that can rationalize the data under Assumptions 1, 2, 3(2) and 4. This will be the content of Theorem 1. We start with a formal definition of the set of cost functions that rationalize the data under the model assumptions.

Definition 3 (Identified Set)**.** For any $\pi \in \Pi$, we call $\mathcal{C}(\pi)$ the collection of functions $C : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^+$, such that there exists a random vector (Y_0, Y_1, D, Z) where $((1-D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 1, 2, 3(2) and 4 are satisfied.

We are interested in characterizing the identified set $\mathcal{C}(\pi)$ with moment inequalities, and conducting inference on an element $\underline{C}$ in $\mathcal{C}(\pi)$, which will be interpreted as the minimal cost function that can rationalize observations within the extended Roy model (Assumptions 1, 2, 3(2) and 4).

Under Assumptions 1 and 2, when $D = 1$, $Y - DC(Y,Z) = Y_1 - C(Y_1, Z) \geq Y_0$, and when $D = 0$, $Y - DC(Y,Z) = Y_0 \geq Y_1 - C(Y_1, Z)$. Hence we have

$$Y - DC(Y,Z) = \max\{Y_0, Y_1 - C(Y_1, Z)\}. \tag{1.1}$$

Hence, Assumptions 1, 2, 3(2) and 4 imply stochastic monotonicity of $Y - DC(Y,Z)$ with respect to Z, which can be expressed as a collection of moment inequalities involving the infinite dimensional parameter C. Conversely, stochastic monotonicity of $Y - DC(Y,Z)$ with respect to Z is shown to imply that C can rationalize the data, i.e., that we can construct a vector of potential outcomes (Y_0, Y_1) such that Assumptions 1, 2, 3(2) and 4 are satisfied. This discussion is formalized in the next theorem (proved in the appendix).

Theorem 1 (Characterization)**.** *For any $\pi \in \Pi$, the identified set $\mathcal{C}(\pi)$ is equal to the set of non negative measurable functions C on $\mathcal{Y} \times \mathcal{Z}$ such that $\mathbb{P}(Y - DC(Y,Z) \geq \underline{b} | Z = z) = 1$*

for each z, and for any (Y, D, Z) with distribution π, and all pairs $z \geq \tilde{z}$ of elements of $\mathcal{Z}$,

$$\inf_{y \in \mathcal{Y}} \left[\mathbb{P}(Y - DC(Y, Z) > y | Z = z) - \mathbb{P}(Y - DC(Y, Z) > y | Z = \tilde{z})\right] \geq 0. \tag{1.2}$$

A notable aspect of the characterization in Theorem 1 is that it shows observational equivalence of Assumption 4 and the much weaker Assumption 4′. More precisely, for any $\pi \in \Pi$, and any function $C : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^+$, such that there exists a random vector (Y_0, Y_1, D, Z) where $((1-D)Y_0 + DY_1, D, Z)$ has distribution π, if Assumption 1, 2, Assumption 3(2) and the weaker Assumption 4′ are satisfied, then (1.2) holds, so that any inference based on (1.2) is valid under the weaker instrumental restriction in Assumption 4′. Conversely, if π satisfies (1.2) for some function C, then there exists a random vector (Y_0, Y_1, D, Z) where $((1 - D)Y_0 + DY_1, D, Z)$ has distribution π and Assumption 1, 2, 3(2) and the stronger Assumption 4 are satisfied.

1.3.2. *Closed form bounds on the cost function.* To complement the characterization of the identified set in Theorem 1, we derive bounds $\underline{C}$ and $\bar{C}$ on costs functions in the identified set $\mathcal{C}(\pi)$. Under an additional shape restriction, we further show that the bound $\underline{C}$ belongs to the identified set $\mathcal{C}(\pi)$, so that it is an achievable sharp bound.

Under Assumptions 1, 2 and 3(2), $Y \geq Y - DC(Y, Z) = \max\{Y_0, Y_1 - C(Y_1, Z)\} \geq Y_0$. Hence, for each $y \in \mathcal{Y}$, we have

$$\mathbb{P}(Y \leq y|z) \;\leq\; \mathbb{P}(Y - DC(Y, Z) \leq y|z) \;\leq\; \mathbb{P}(Y \leq y, D = 0|z) + \mathbb{P}(D = 1|z)1\{y \geq \underline{b}\},$$

where the last inequality uses standard worst-case bounds for the distribution of potential outcome Y_0. Now, under Assumption 4′ (hence, also under Assumption 4, which is stronger), the middle term is monotone non increasing in z for each y, and is right-continuous as a function of y. Hence, defining

$$\begin{aligned} \underline{F}(y|z) \;&:=\; \lim_{\tilde{y} \downarrow y} \sup \left\{\mathbb{P}(Y \leq \tilde{y}|\tilde{z}) : \tilde{z} \in \mathcal{Z}, \tilde{z} \geq z\right\}, \\ \bar{F}(y|z) \;&:=\; \lim_{\tilde{y} \downarrow y} \inf \left\{\mathbb{P}(Y \leq y, D = 0|\tilde{z}) + \mathbb{P}(D = 1|\tilde{z})1\{y \geq \underline{b}\} : \tilde{z} \in \mathcal{Z}, \tilde{z} \leq z\right\}, \end{aligned} \tag{1.3}$$

we have

$$\underline{F}(y|z) \;\leq\; \mathbb{P}(Y - DC(Y, Z) \leq y|z) \;\leq\; \bar{F}(y|z), \tag{1.4}$$

for all (y, z), whenever Assumptions 1, 2, 3(2) and 4′ hold. This yields the testable implication for our model that bounds $\underline{F}(y|z)$ and $\bar{F}(y|z)$ cannot cross. Since $\mathbb{P}(Y - DC(Y, Z) \leq y|z) = \mathbb{P}(Y - C(Y, Z) \leq y, D = 1|z) + \mathbb{P}(Y \leq y, D = 0|z)$, and since $\mathbb{P}(Y - C(Y, Z) \leq y, D =$

$1|z)$ is non decreasing in y, (1.4) also yields the bounds

$$\begin{aligned} L(y|z) &:= \sup_{\tilde{y}\leq y}\{\underline{F}(\tilde{y}|z) - \mathbb{P}(Y \leq \tilde{y}, D=0|z)\} \\ &\leq \mathbb{P}(Y - C(Y,Z) \leq y, D=1|z) \qquad (1.5) \\ &\leq \inf_{y\leq\tilde{y}}\left\{\bar{F}(\tilde{y}|z) - \mathbb{P}(Y \leq \tilde{y}, D=0|z)\right\} \;:=\; U(y|z). \end{aligned}$$

Finally, we have the following lemma, establishing bounds on the cost functions in the identified set of Definition 3.

Corollary 1 (Bounds on the Cost Function)**.** All cost functions C in the identified set $\mathcal{C}(\pi)$ of Definition 3 satisfy for all $(y,z)\in\mathcal{Y}\times\mathcal{Z}$:

$$\underline{C}(y,z) \;:=\; y - L_-\left(F_1(y|z)|z\right) \;\leq\; C(y,z) \;\leq\; y - U^-\left(F_1(y|z)|z\right) \;:=\; \bar{C}(y,z),$$

where $F_1(y|z) := \mathbb{P}(Y \leq y, D=1|Z=z)$ and U^- and L_-, defined respectively by $U^-(x) := \inf\{y : U(y) \geq x\}$ and $L_-(x) := \sup\{y : L(y) \leq x\}$, are generalized inverses of U and L from (1.5).

To prove that the lower bound of Corollary 1 is attained (hence is sharp), we need to verify that $\underline{F}(\cdot|z)$ is a well defined cdf for all z, and hence is suitable candidate for the distribution function of $Y - DC(Y,Z)$. This is shown in Lemma 3 in the Appendix. We also need the following shape and regularity assumptions.

Assumption 5. *The functions $F_1(\cdot|z)$ defined in Corollary 1 and $\underline{F}(\cdot|z) - \mathbb{P}(Y \leq \cdot, D = 0|z)$, where $\underline{F}$ is defined in (1.3) are continuous and increasing for all $z\in\mathcal{Z}$. The function $\underline{C}$ defined in Corollary 1 is continuous in y for all $z\in\mathcal{Z}$.*

Corollary 2 (Sharpness)**.** Under Assumption 5, the bound $\underline{C}$ of Corollary 1 is an element of the identified set $\mathcal{C}(\pi)$. Hence it is an achievable sharp bound.

The closed form solutions for the bounds on the identified set provides a simpler basis for inference than the characterization of Theorem 1, which relies on a continuum of moment inequalities. Inference on $\underline{C}$ and $\bar{C}$ will be discussed in Section 4.3.

2. Structural underpinnings of the selection equation

Structural underpinnings for the selection rule of Assumption 2 are now discussed. It can be interpreted as a transformation of a decision rule based on relative utilities in both sectors. Suppose individuals choose the sector based on the comparison between utility in Sector 1, $u_1(Y_1,Z)$, and utility in Sector 0, $u_0(Y_0,Z)$, with u_0 and u_1 satisfying the following traditional shape and regularity conditions.

Assumption 6 (Utility model)**.** *For $d=0,1$, Sector d utility $u_d : \mathcal{Y}\times\mathcal{Z}\mapsto\mathbb{R}$ satisfies:*

(1) *Utility u_d is continuously differentiable and $y \mapsto u_d(y,z)$ is increasing with range $\mathbb{R}$ for each $z \in \mathcal{Z}$ and for $d = 0, 1$.*
(2) *For each $(y,z) \in \mathcal{Y} \times \mathcal{Z}$, $u_0(y,z) \geq u_1(y,z)$.*
(3) *For each $(y,z) \in \mathcal{Y} \times \mathcal{Z}$, and $d = 0, 1$, $\nabla_z u_0(u_0^{-1}(u_1(y,z),z),z) \leq \nabla_z u_1(y,z)$.*

Sector specific utility u_d can be rationalized by different amenities. Take $Z = (Z_0, Z_1)$ for instance, where Z_d is the proportion of women on the faculty in Sector d. Women applicants may prefer Sector 0 because the proportion of women in the faculty is larger, hence $u_0(y,z) = u(y,z_0) > u(y,z_1) = u_1(y,z)$. In this latter case, where $u_d(y,z) = u(y,z_d)$, $d = 0, 1$, we say that the instrument is sector specific.

Sector specific utility may also be rationalized with reference dependence (Thaler [1980] and Tversky and Kahneman [1991]) based on gender profiling: social conditioning makes women prefer Sector 0. In the case a vector of instruments Z can potentially affect both sector specific utilities, the mother's educational attainment being one example, Assumption 6(3) holds if Assumption 6(2) does, $\nabla_z u_1(y,z) \geq \nabla_z u_0(y,z)$ and $\nabla_z \left(\partial u_1(y,z)/\partial y\right) \geq 0$.

Assumptions 2 and 3 can be recovered from sector specific utilities in the following way.

Proposition 1 (Utility Model)**.** Under Assumption 6(1), if $u_d(Y_d, Z) > u_{1-d}(Y_{1-d}, Z)$ implies $D = d$, for $d = 0, 1$, then Assumptions 2 and 3(1) hold with $C(y,z) := y - u_0^{-1}(u_1(y,z),z)$, where the inverse of u_0 is taken relative to the first argument. In addition, we have the following.

(1) If in addition, Assumption 6(2) holds, then Assumption 3(2) holds.
(2) If in addition, Assumption 6(3) holds, then Assumption 3(3) holds.

Example 1 (Quasi-linear utilities)**.** In the special case $u_d(y,z) = y + g_d(z)$, for some function g_d, $d = 0, 1$, the cost function becomes $C(y,z) = g_0(z) - g_1(z)$. Assumption 3 holds when $g_0(z) - g_1(z)$ is non negative and non increasing in z. In the context of women's major choice, this is satisfied, for instance, if the proportion of female faculty has no effect on utility from non STEM majors, so that $g_0(z) = 0$, and has an increasing positive effect on utility from STEM majors, so that g_1 is increasing in z.

Note that in Example 1 above, the cost function $C(y,z) = g_0(z) - g_1(z)$ is a function of z only, as in the extended Roy models analyzed in the literature (see Bayer et al. [2011] and d'Haultfœuille and Maurel [2013]). More generally, the cost function $C(y,z) = y - u_0^{-1}(u_1(y,z),z)$ based on the utility model of Assumption 6 depends on potential outcomes unless the constraint $\partial u_1(y,z)/\partial y = \partial u_0(u_0^{-1}(u_1(y,z),z)z)/\partial y$ holds.

When the selection rule of Assumption 2 is derived from utility maximization, the cost function C can be interpreted as a compensating wage differential. Suppose women perceive inferior amenities in the STEM sector. Call $(y,z) \mapsto \tilde{C}(y,z)$ the compensating differential defined by $u_1(y,z) = u_0(y - \tilde{C}, z)$. Then $\tilde{C}(y,z) = y - u_0^{-1}(u_1(y,z),z) = C(y,z)$. Hence, $C(y,z)$ in Assumption 2 is a monetary adjustment that makes women, whose talents entitle them to identical (uncompensated) wages in both sectors, indifferent between the two sectors. As defined, it is a willingness to pay for the better amenities of the non-STEM sector, or equivalently, the equivalent variation to move from non-STEM to STEM. Defining the model symmetrically would allow us to partially identify willingness to accept compensation and compensating variation as well. Note, however, that the identified set would then be different from the one we derive for willingness to pay and equivalent variation.

3. EXTENSIONS: IMPERFECT FORESIGHT AND GENERALIZED ROY MODEL

3.1. **Imperfect Foresight.** In order to analyze the sensitivity of our results to the assumption that individuals make sector choices with perfect knowledge of their future earnings in each sector, we also consider a variation of the model, called imperfect foresight, where choices are made using expectations based on the decision maker's information set $\mathcal{I}$ at the time of decision. We assume throughout this section that the instrument Z is $\mathcal{I}$-measurable, since it is a vector of variables easily observable at the time of choice. In the context of major choice, the information set involves some knowledge of individual talent for mathematics and non mathematics intensive activities, as well as some anticipation of future labor market conditions and the prices of talent.

Assumption 7 (Imperfect Foresight)**.** *The selection indicator D satisfies*

$$\mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}] > \mathbb{E}[Y_0|\mathcal{I}] \Rightarrow D = 1 \text{ and } \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}] < \mathbb{E}[Y_0|\mathcal{I}] \Rightarrow D = 0,$$

where the conditional expectations are well defined, $\mathcal{I}$ is the sigma-algebra characterizing the agent's information set at the time of sector choice, Z is $\mathcal{I}$-measurable, and C is a non negative measurable function on $\mathcal{Y} \times \mathcal{Z}$, such that $y \mapsto y - C(y,z)$ is continuous and increasing in y for each $z \in \mathcal{Z}$.

As in the case of perfect foresight, we derive a crucial implication of Assumptions 1 and 7. Under Assumptions 1 and 7, when $D = 1$, $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}] \geq \mathbb{E}[Y_0|\mathcal{I}]$, and when $D = 0$, $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \mathbb{E}[Y_0|\mathcal{I}] \geq \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}]$. Hence we have

$$\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \max\{\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}]\}. \tag{3.1}$$

This relation is the key to deriving testable implications of restrictions on the joint distribution of potential outcomes. This relation also provides a direct proof of the identification of the cost function under the assumptions of d'Haultfœuille and Maurel [2013], as described in Appendix C.

The imperfect foresight version of the Assumption 4 is the following:

Assumption 8 (Imperfect Foresight Monotonicity)**.** *The random vector* $(\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1 - C(Y_1, Z)|\mathcal{I}])$ *is stochastically monotone non decreasing with respect to* Z*.*

We now discuss some sufficient conditions for Assumption 8. Writing $\mathbb{E}[Y_d - dC(Y_1, Z)|\mathcal{I}] = \mathbb{E}[Y_d - dC(Y_1, Z)|Z] + V_d$, for $d \in \{0, 1\}$, the case $(V_0, V_1) \perp Z$, coupled with monotonicity in z of $\mathbb{E}[Y_d - dC(Y_1, Z)|Z = z]$, for $d \in \{0, 1\}$, is a special case of Assumption 8. Another set of sufficient conditions mirrors Lemma 1: Assumption 8 holds if the cost function C is only a function of z and is non increasing in z, and the random vector $(\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1|\mathcal{I}])$ is stochastically monotone with respect to Z. For example, in the isoelastic utility case in Appendix B, if $\mu_d(z)$ is a non decreasing function of z, then $(\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1|\mathcal{I}])$ is stochastically monotone with respect to Z. Moreover, if $\sigma_1^2(z) - \sigma_0^2(z)$ is a non increasing function of z, then C is a non increasing function of z. Hence Assumption 7 holds.

As in the perfect foresight case, a simple implication of Assumption 8 is the following.

Assumption $8'$. *The random variable* $\max\{\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1 - C(Y_1, Z)|\mathcal{I}]\}$ *is stochastically monotone non decreasing with respect to* Z*.*

Consider Assumptions 7 and 8 instead of Assumptions 2 and 4. As before, the object of inference is the hidden cost function C. The next result characterizes the set of such functions that can rationalize the data under Assumptions 1, 7 and 8.

Definition 4 (Identified Set under Imperfect Foresight)**.** For any $\pi \in \Pi$, we call $\mathcal{C}^i(\pi)$ the collection of functions $C : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}^+$ such that there exists a σ-algebra $\mathcal{I}$ and a random vector (Y_0, Y_1, D, Z) where $((1 - D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 7 and 8 are satisfied. Denote by $\tilde{\mathcal{C}^i}(\pi)$ the set of functions in $\mathcal{C}^i(\pi)$ that are independent of y.

The next theorem provides a characterization of the identified set $\mathcal{C}^i(\pi)$ with moment inequalities.

Theorem 2 (Characterization under Imperfect Foresight)**.** *For any* $\pi \in \Pi$*, the identified set* $\mathcal{C}^i(\pi)$ *is equal to the set of non negative measurable functions* C *on* $\mathcal{Y} \times \mathcal{Z}$ *such that* $y \mapsto y - C(y, z)$ *is continuous and increasing for each* z*, and for all ordered pairs* $z \geq \tilde{z}$ *on the support of* Z*,* $\mathbb{E}(Y - DC(Y, Z)|Z = z) \geq \mathbb{E}(Y - DC(Y, Z)|Z = \tilde{z})$*, where* (Y, D, Z) *is any vector with distribution* π*.*

We observe a similar relation between Assumptions 8 and $8'$ as between Assumptions 4 and $4'$. Indeed, if a cost function rationalizes the data under Assumptions 1, 7 and the weaker $8'$, the characterization in Theorem 2 holds. Conversely, if the characterization in Theorem 2 holds, then the cost function rationalizes the data under Assumptions 1, 7 and the stronger Assumption 8.

As before, we now seek to characterize a minimal cost function that rationalizes the data under the imperfect foresight Roy model in closed form. Following the reasoning of the perfect foresight case, we start from Theorem 2, which characterizes cost functions that rationalize the data with the monotonicity in z of the conditional expectation $\mathbb{E}[Y-DC(Y,Z)|Z=z]$. Since C is non negative, we seek functions C such that $\mathbb{E}[Y-DC(Y,Z)|Z=z]=\inf\{\mathbb{E}[Y|Z=\tilde{z}];\ \tilde{z}\geq z\}$, the lower monotone envelope of $\mathbb{E}[Y|Z=z]$. This would yield the constraint $\mathbb{E}[C(Y,Z)|D=1,Z=z]=\underline{C}(z)$, where $\underline{C}$ is defined in (3.2) below. Cost functions satisfying the constraint share the same expectation and would not be ordered, leading to a multiplicity of minimal cost functions. We therefore limit our search to cost functions that are functions of z only, as in the case, where sector selection is based on the comparison of expected quasilinear utilities. The following corollary of Theorem 2 establishes that $\underline{C}$ of (3.2) is the minimum of the set $\tilde{\mathcal{C}}^i(\pi)$ of functions of z that rationalize the data under the imperfect foresight extended Roy model. The maximum cost function is obtained straightforwardly using the worst case bound, as in Section 1.3.

Corollary 3. For any $\pi\in\Pi$ such that $\tilde{\mathcal{C}}^i(\pi)$ is non empty, the minimum and maximum of $\tilde{\mathcal{C}}^i(\pi)$ are the function $\underline{C}$ and $\bar{C}$ defined for any z on the support of Z by[1]

$$\begin{aligned}\underline{C}(z) &= \mathbb{P}(D=1|Z=z)^{-1}\left(\mathbb{E}[Y|Z=z]-\inf_{\tilde{z}\geq z}\mathbb{E}[Y|Z=\tilde{z}]\right)1\{\mathbb{P}(D=1|Z=z)>0\},\\ \bar{C}(z) &= \mathbb{P}(D=1|Z=z)^{-1}\left(\mathbb{E}[Y|Z=z]-\sup_{\tilde{z}\leq z}\mathbb{E}[Y(1-D)+\underline{b}D|Z=\tilde{z}]\right).\end{aligned}\tag{3.2}$$

Remark (Testability)**.** Note that $\tilde{\mathcal{C}}^i(\pi)$ is non empty if $\mathcal{C}^i(\pi)$ is. Indeed, it follows from Theorem 2 that $C^*(z):=\mathbb{E}(C(Y,Z)|D=1,Z=z)1\{P(D=1|Z=z)>0\}$ is in $\mathcal{C}^i(\pi)$ if C is. Moreover, any π that satisfies $\mathbb{P}(D=1|Z)>0$ on the support of Z has non empty $\tilde{\mathcal{C}}^i(\pi)$, which implies that the model is not testable in that case. Indeed, in that case, $\underline{C}$ defined in Corollary 3 is in $\tilde{\mathcal{C}}^i(\pi)$. However, if π is such that $\mathbb{P}(D=1|Z=z)=\mathbb{P}(D=1|Z=\tilde{z})=0$ and $\mathbb{E}[Y|Z=z]<\mathbb{E}[Y|Z=\tilde{z}]$ for $z\geq\tilde{z}$ on the support of Z, then $\mathcal{C}^i=\varnothing$, so that Assumptions 1, 7 and 8 are jointly rejected.

3.2. **Random Cost.** In an extension of the selection model of Assumption 2, consider the case where the cost function is not restricted to deterministic functions of potential outcome Y_1 and observable covariate Z, but may depend on unobservable heterogeneity as well.

Assumption GRM. The selection indicator satisfies $Y_d-dC>Y_{1-d}-(1-d)C\Rightarrow D=d$, for $d=0,1$, C is a non negative random variable, and (Y_0,Y_1-C) is stochastically monotone non decreasing with respect to Z.

If choices are driven by the comparison of utilities in both sectors $u_d(Y_d,Z,\eta)$, for $d=0,1$, where η is a vector of heterogeneous drivers of utility, that are observed by the agents but

[1] We use the convention $x/0=+\infty$ for any $x\geq 0$.

not the analyst, then Assumption GRM holds with $C = Y_1 - u_0^{-1}(u_1(Y_1, Z, \eta), Z, \eta)$, where the inverse of u_0 is taken relative to the first argument. In the context of women's major choices, the unobserved heterogeneity component η may embody unobserved preferences for a major and unobserved quality of the different programs under consideration. If choices are driven by the comparison of expected utilities in both sectors $\mathbb{E}[u_d(Y_d, Z, \eta)|\mathcal{I}]$, for $d = 0, 1$, where $\mathcal{I}$ is the σ-algebra characterizing the agent's information set at the time of major choice, then Assumption GRM holds with $C = Y_1 - Y_0 + \mathbb{E}[u_0(Y_0, Z, \eta) - u_1(Y_1, Z, \eta)|\mathcal{I}]$. The Assumption is also compatible with different priors about the (lack of) rationality of decision makers.

Definition 5 (Identified Set for Random Cost)**.** For any $\pi \in \Pi$, we call $\mathcal{C}^r(\pi)$ the collection of non negative random variables C such that there exists a random vector (Y_0, Y_1, D, Z) where $((1-D)Y_0 + DY_1, D, Z)$ has distribution π and Assumption GRM is satisfied.

In analogy with the exercise of the previous section, we seek bounds on the distribution of costs C that rationalize the data. Consider the envelope distribution $\underline{F}$ of (1.3). A minimal cost in first order stochastic dominance must satisfy $Y - DC = Y^e \sim \underline{F}$, if the latter is feasible. This is equivalent to $C = Y - Y^e$ under the constraints $Y^e \sim \underline{F}$ and $(1-D)(Y - Y^e) = 0$. If the second constraint is feasible, the lower bound on the distribution of the difference $Y - Y^e$ with given fixed marginals will provide the lower bound on the distribution of random costs that rationalize the data under Assumption GRM. Note that, unlike the case of the extended Roy model of Section 3, the cost is no longer restricted to be a deterministic function of Y_d and Z, but can also depend on unobservable heterogeneity as well. Hence the joint distribution of (Y, Y^e) is only restricted by the marginals, and the copula is unrestricted, when obtaining the lower bound on the distribution for $Y - Y^e$ (equivalently the upper bound on the cumulative distribution function). A similar reasoning applies for maximal costs. We formalize this in the following theorem.

Theorem 3. *Let $\pi \in \Pi$ be a distribution for observable variables (Y, D, Z). If $C \in \mathcal{C}^r(\pi)$, then for each y and z, $F_L(y|z,1) \leq \mathbb{P}(C \leq y|Z = z, D = 1) \leq F_U(y|z,1)$, where*

$$\begin{aligned} F_L(y|z,1) &:= \sup_{\tilde{y}} \max\left\{0, F(\tilde{y}|Z = z, D = 1) - \bar{F}(\tilde{y} - y|z,1)\right\}, \\ F_U(y|z,1) &:= 1 + \inf_{\tilde{y}} \min\left\{0, F(\tilde{y}|Z = z, D = 1) - \underline{F}(\tilde{y} - y|z,1)\right\}, \\ \underline{F}(y|z,1) &:= p(z)^{-1}\left[\underline{F}(y|z) - F(y|z)\right] 1\{p(z) > 0\} + F(y|z,1), \\ \bar{F}(y|z,1) &:= p(z)^{-1}\left[\bar{F}(y|z) - F(y|z)\right] 1\{p(z) > 0\} + F(y|z,1), \end{aligned}$$

where $p(z) := \mathbb{P}(D = 1|Z = z)$ is the propensity score.

4. Empirical investigation of major choices in Germany

There is ample evidence that women are severely under-represented in STEM university majors and even more so in STEM fields (see for instance Beede et al. [2011], Zafar [2013] and Hunt et al. [2013]). Evidence on women's reasons for shunning STEM fields (see for instance Kahn and Ginther [2017]) include mathematics gender stereotypes and gender biased amenities, such as family friendliness and work/life balance. Our objective is to document the amplitude of such non pecuniary motivations as revealed in the form of a gender-specific cost of choosing STEM fields. The revealed cost of STEM fields is a function of the rate of feminization of the STEM faculty in the region at the time of choice. We therefore also shed light on the importance of role models in the determination of major choice (Kahn and Ginther [2017]).

4.1. **Data.** Our empirical analysis relies on surveys of German nationally representative university graduates. The data are collected by the German Centre for Higher Education Research and Science Studies (DZHW) as part of the DZHW Graduate Survey Series. Data and methodology are described in Baillet et al. [2017]. The waves we consider include graduates who obtained their highest degree during the academic years 2004-2005 and 2008-2009 respectively. Graduates were interviewed 1 year and 5 years after graduation[2]. At that point, extensive information was collected on their educational experience, employment history, including wages and hours worked, along with detailed socio-economic variables and geographical information about the region where the *Abitur* (high school final exam) was completed. We merge the fields of study into two categories. We call STEM the category, which consists of mathematics, physical, life and computer sciences, as well as engineering and related fields. The remaining majors are merged in the non-STEM-degree category. We only consider graduates from institutions in the country of the survey, who are active on their respective country's labor market at the time of the interview. We exclude all respondents who are still in education, have never worked or are currently inactive, unemployed, in part-time employment or self-employed. We keep only graduates who hold a "Bachelor", "Magister" or "Diplom", excluding those with "Staatsexamen" and "Lehramt" degrees, which are specific tracks mainly for teachers. Our stochastically monotone instrumental variable (SMIV), which we call "feminization of STEM," is defined as the proportion of women among faculty members by field of study in universities in the individual's region (Land) of residence at the time of choice. This variable is calculated for each individual in the sample from data on gender distribution of faculty by field and by Land provided by the Federal Statistical Office of Germany (DESTATIS). The data set provides for each

[2] The response rate was 25% for the 2005 cohort, with 39.5% attrition in the second wave, and 20% for the 2009 cohort with 14% attrition in the second wave.

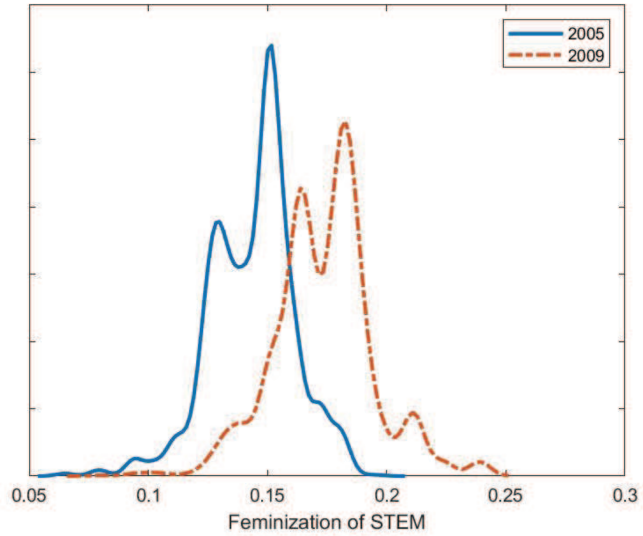
year between 1998 and 2010, the count of faculty members (Scientific and artistic staff "*Wissenschaftliches und Künstlerisches Personal*") by gender in ten fields of study. The variable is computed from the aggregation of mathematics, science and engineering.

4.2. **Descriptive statistics.** The category of individuals we consider consists of women from the former West Germany[3]. The variables we consider are average income during the first year after graduation, which serves as outcome variable Y, the choice of major D, which takes value 1 if the chosen sector is STEM and 0 otherwise, and the feminization of STEM, which serves as our SMIV instrumental variable Z. Table 1 gives proportions of STEM majors among women from the former West Germany in 2005 and 2009, as compared to proportions of STEM majors among men of the same category. Figure 1 shows the

Table 1. Sample of women and comparison with men

	Women		Men	
	2005	2009	2005	2009
STEM	423	247	1,018	712
Other	1,226	1,276	780	690

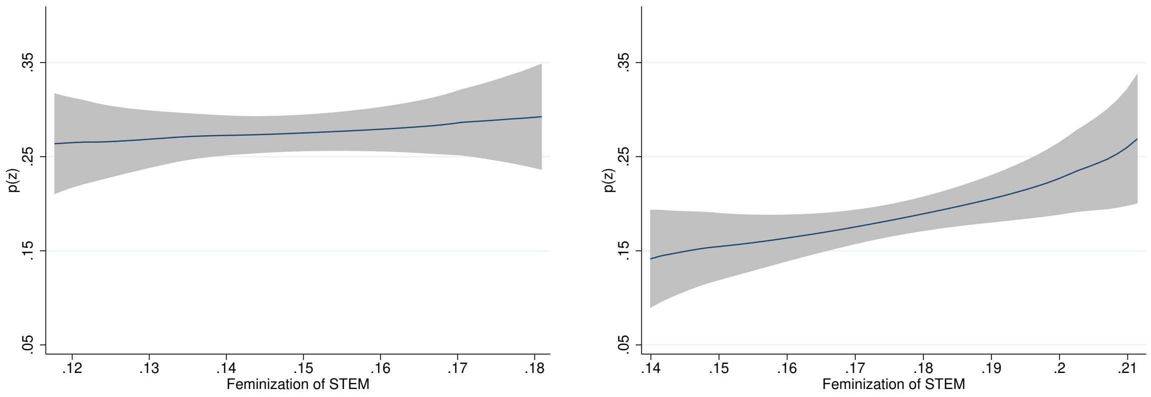
Figure 1. Distribution of Proportion of female faculty by field



[3] Mourifié et al. [2020] find no rejection of the Roy model in the sample of women from the former East Germany, hence a zero cost of STEM rationalizes choices. The difference in behavior between the former East and West may be partly attributable to differences in gender stereotypes, as evidenced in Lippmann and Senik [2018]

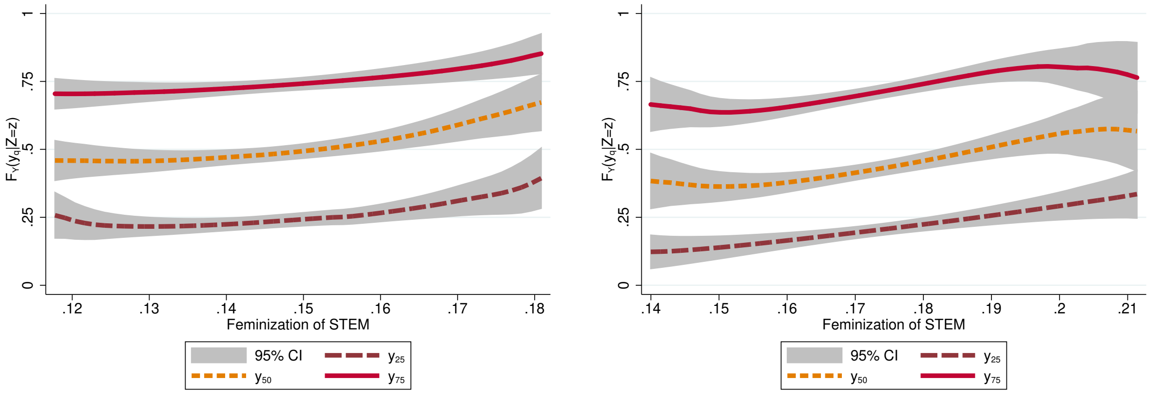
distribution of the feminization of STEM variable for cohorts graduating in 2005 and 2009. The relation between income, field of study and feminization of STEM is investigated with an estimation of the propensity score (probability of choosing STEM) as a function of feminization of STEM in 2005 and 2009 in Figure 2 and quartile regressions of income as a function of the feminization of STEM for 2005 and 2009 in Figure 3. Figure 2 shows a positive relationship between major choice and feminization of STEM (which may or may

Figure 2. Relation between D and Z



(a) STEM proportion, 2005 **(b)** STEM proportion, 2009

Figure 3. Relation between Y and Z



(a) 2005 **(b)** 2009

not be causal). Figure 3 shows a violation of stochastic monotonicity of income relative to the feminization of STEM, which explains the rejection of the pure Roy model for this category of individuals in Mourifié et al. [2020].

4.3. **Empirical methodology.** We now propose an inference procedure for the minimal cost function of Corollary 2. The upper bound can be treated symmetrically. However, it is uninformative in our data. For any given value of $z \in \mathcal{Z}$, we seek a data driven function $y \mapsto C_n(y,z)$ such that for each $y \in \mathcal{Y}$,

$$\lim_{n\to\infty} \mathbb{P}(C_n(y,z) \leq \underline{C}(y,z)) \geq 1-\alpha, \tag{4.1}$$

for some pre-determined level of significance α. Define $G(y|z,\tilde{z}) := \mathbb{P}(Y \leq y|\tilde{z}) - \mathbb{P}(Y \leq y, D=0|z)$. Call $\hat{G}$ a non parametric estimator for G, and define

$$\hat{G}_-(x|z,\tilde{z}) := \sup\left\{y \in \mathcal{Y} \ : \ \hat{G}(y|z,\tilde{z}) \leq x\right\}.$$

Finally, let $\hat{F}_1$ be a nonparametric estimator for $F_1(y|z) := \mathbb{P}(Y \leq y, D=1|z)$. In practice, we use nonparametric estimation procedures in Li and Racine [2008]. Lemma 2 (proved in the appendix) shows the applicability of the methodology in Chernozhukov et al. [2013].

Lemma 2. Under Assumption 5 and assuming $G(y|z,\tilde{z}) := \mathbb{P}(Y \leq y|\tilde{z}) - \mathbb{P}(Y \leq y, D=0|z)$ is continuous and increasing in y for all $z, \tilde{z}$, we have

$$\underline{C}(y,z) \geq y - \inf_{\tilde{z}\geq z} G_-(F_1(y|z)|z,\tilde{z}), \quad \text{where} \quad G_-(x|z,\tilde{z}) := \sup\{y \in \mathcal{Y} : G(y|z,\tilde{z}) \leq x\}.$$

Let $s_n(y;z,\tilde{z})$ be a standard error for the estimator $\hat{G}_-(\hat{F}_1(y|z)|z,\tilde{z})$ and $c_n^\alpha(y;z)$ be the critical value of Definition 3 in Chernozhukov et al. [2013]. Then, under the assumptions of Theorem 6 of Chernozhukov et al. [2013],

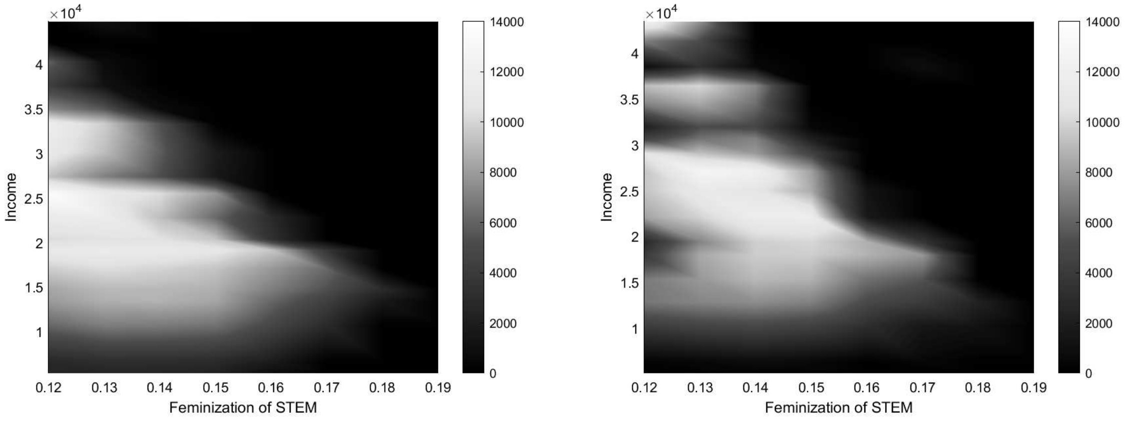
$$C_n(y,z) := y - \inf_{\tilde{z}\leq z}\left\{\hat{G}_-(\hat{F}_1(y|z)|z,\tilde{z}) + c_n^\alpha(y;z)s_n(y;z,\tilde{z})\right\}$$

satisfies requirement (4.1). Details of the procedure are given in Appendix D.

5. Findings

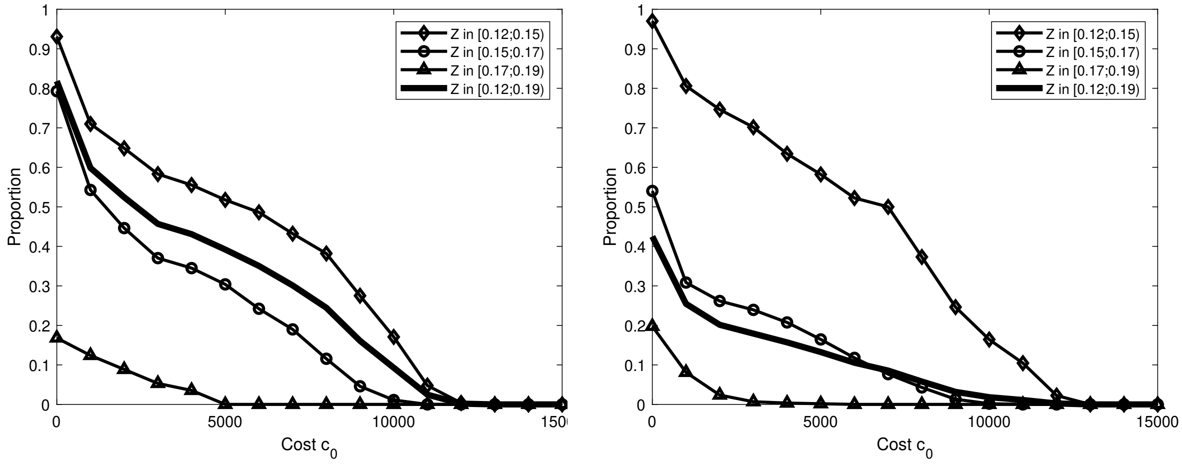
Figure 4 shows the lower bound of a one-sided 95% confidence region for the cost function. More precisely, it represents a function $C_n(y,z)$ of income and feminization of STEM, such that $\lim_{n\to\infty} \mathbb{P}(C_n(y,z) \leq \underline{C}(y,z)) \geq 0.95$, where $\underline{C}$ is the lower bound of the identified set. It is lower in 2009 than in 2005, which is in line with the increased feminization of STEM over time (seen in Figure 1), but the shape is similar. We observe that costs tend to be high for low income individuals and remain high for high income individuals, when the rate of feminization of STEM is low. Figures 5-7 show different ways to visualize the data from Figure 4. Figure 5 shows, for each cost level c_0, the proportion of individuals with $C_n \geq c_0$,

Figure 4. One-sided confidence region for minimum cost



(a) 2005 **(b)** 2009

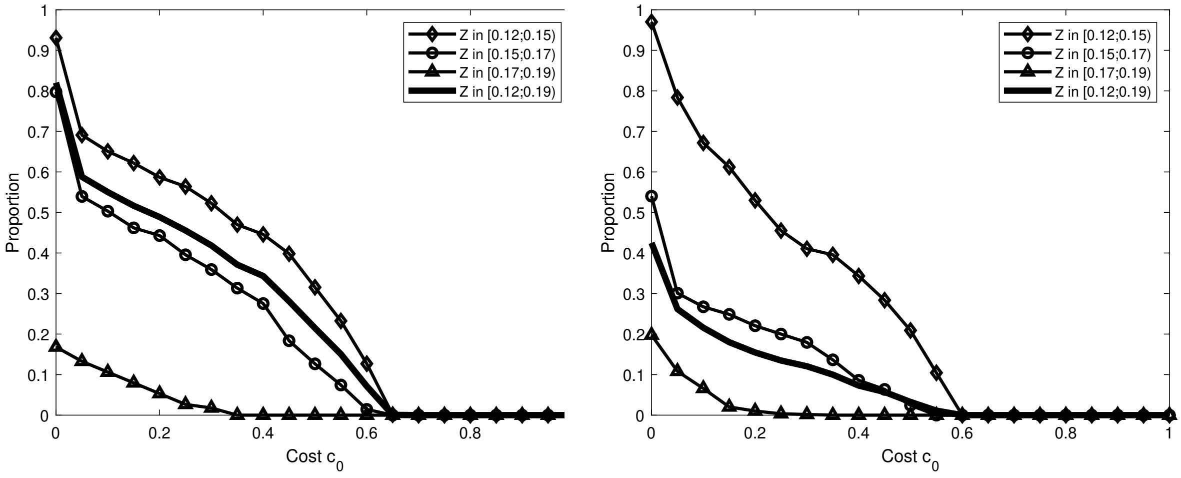
Figure 5. Distribution of costs in the population



(a) 2005 **(b)** 2009

averaged over three different bins of the feminization of STEM variable Z (diamond, circle and triangle curves) and their union (thick curve). Panels (a) and (b) show this for 2005 and 2009 respectively. Figure 6 is identical to Figure 5, except that the x axis is now cost level divided by income. So in Panel (a) of Figure 6 for instance, the triangle curve tells us that about 5% of individuals with $Z \in [0.17, 0.19)$ have C_n/Y at least as large as 0.2.

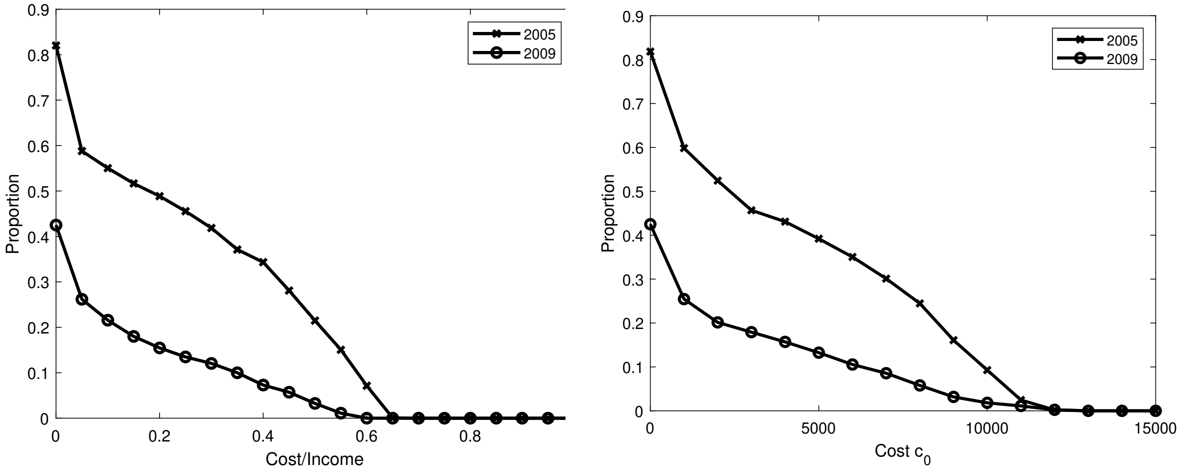
Figure 6. Distribution of costs as proportion of income in the population



(a) 2005 **(b)** 2009

Panel (a) of Figure 7 places the thick curves of Figures 6(a) and (b) together, and similarly,

Figure 7. Comparison 2005 - 2009



(a) Cost/Income **(b)** Cost

Panel (b) of Figure 7 places the thick curves of Figures 5(a) and (b) together for an easier comparison of 2005 and 2009. In Figure 5 and 7(b), we observe that among the 2005 graduation cohort, 9 out of 10 have positive minimum cost function, and 4 out of 10 have minimum cost larger than $5,000$ euros. This falls to 6 in 10 and 2 in 10 respectively, for the 2009 cohort. In Figure 6 and 7(a), we observe that among the 2005 graduation cohort, 5 out of 10 have minimum cost function larger than 20% of their income, and 3 out of 10 have minimum cost larger than 40% of their income. This falls to 2 in 10 and 1 in 10 respectively, for the 2009 cohort. Costs decrease sharply as the feminization rate increases in both years. These results suggest that policies promoting a higher representation of female faculty in STEM fields could reduce the perceived cost of STEM for the next generation.

6. Discussion

The objective of this paper was to uncover gender specific costs of choosing STEM fields based on minimal behavioral assumptions. We assumed that the choice of (STEM versus non STEM) major is determined on the basis of the maximization of $Y_d - dC(Y_d, Z)$, for $d = 1$ (STEM) or $d = 0$ (non STEM), where Y_d is potential income in Sector d, Z is the proportion of women on the STEM faculty in the individual's region, and $C(y, z)$ is the cost function of interest. As the choice $C(y, z) = 0$ can rationalize any joint distribution for (Y, D, Z), we rely on the additional identification restriction that potential incomes (Y_0, Y_1) are stochastically monotone in Z, as in Mourifié et al. [2020], and $C(y, z)$ is non increasing in z. With our choice of instrument Z, this restriction is interpreted as a positive (or neutral) effect of role models during university education on future prospects. Under this assumption, we characterized the (sharp) identified region for the function C as well as sharp lower and upper bounds in closed form. Confidence regions, derived using existing inference methods (see Chernozhukov et al. [2013]), reveal large costs of STEM fields for German female university graduates, especially for low realized income levels, and low presence of role models (as measured by the regional rate of feminization of the STEM faculty).

We interpret the revealed cost function C as a compensating wage differential, but our current methodology cannot disentangle the role of real or perceived gender biased disamenities of STEM fields, in terms of family friendliness and work/life balance, from behavioral and preference biases related to gender stereotypes. An area of concern regarding the validity of our identifying stochastic monotonicity assumption is the aggregation in the STEM category, of areas such as engineering, with extremely low feminization rates and high relative incomes, with areas such as life sciences, with high feminization rates and low relative incomes. Access to more disaggregated data on the proportions of female faculty in mathematics intensive fields (rather than STEM fields as traditionally classified) would considerably alleviate this concern.

Appendix A. Proofs of results in the main text

In the proof of Lemma 1 as well as Theorem 1, we rely on a fundamental characterization of first order stochastic dominance, which is stated in Proposition 2 below, and which can be found, for instance, in Shaked and Shanthikumar [2007], Section 6.B.1.

Definition 6 (Upper Sets)**.** A set $U \subseteq \mathbb{R}^k$ is called an *upper set* if $y \in U$ implies $\tilde{y} \in U$ for all $\tilde{y} \geq y$.

Proposition 2 (Characterization of First Order Stochastic Dominance)**.** A random vector X_1 is first order stochastically dominated by a random vector X_2 if and only if $\mathbb{P}(X_1 \in U) \leq \mathbb{P}(X_2 \in U)$ for all upper sets U in $\mathbb{R}^k$.

Proof of Lemma 1. Assume (Y_0, Y_1) to be stochastically monotone non decreasing with respect to Z. Fix z and $\tilde{z}$ in the support of Z, such that $z \geq \tilde{z}$. Fix an upper set U in $\mathbb{R}^2$. $\mathbb{P}((Y_0, Y_1 - C(Y_1, Z)) \in U | Z = z) = \mathbb{P}((Y_0, Y_1) \in G^{-1}(U, Z) | Z = z)$, where G is defined by $G((y_0, y_1), z) = (y_0, y_1 - C(y_1, z))$. Since G is increasing in the componentwise order of $\mathbb{R}^2$, $G^{-1}(U, z)$ is also an upper set, and Proposition 2 yields $\mathbb{P}((Y_0, Y_1) \in G^{-1}(U, z) | Z = z) \geq \mathbb{P}((Y_0, Y_1) \in G^{-1}(U, z) | Z = \tilde{z})$. Since C is non increasing in z, the right-hand-side of the latter inequality is at least as large as $\mathbb{P}((Y_0, Y_1) \in G^{-1}(U, \tilde{z}) | Z = \tilde{z}) = \mathbb{P}((Y_0, Y_1 - C(Y_1, \tilde{z})) \in U | Z = \tilde{z})$. We have therefore proved that for any upper set U, $\mathbb{P}((Y_0, Y_1 - C(Y_1, z)) \in U | Z = z) \geq \mathbb{P}((Y_0, Y_1 - C(Y_1, \tilde{z})) \in U | Z = \tilde{z})$, for any $z \geq \tilde{z}$, which is equivalent to stochastic monotonicity of $(Y_0, Y_1 - C(Y_1, Z))$ with respect to Z, as required. □

Proof of Theorem 1. Fix any $\pi \in \Pi$, and any $C \in \mathcal{C}(\pi)$. By Definition 3, there exists a random vector (Y_0, Y_1, D, Z) where $((1 - D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 1, 2, 3(2) and 4 are satisfied. Call $Y := (1 - D)Y_0 + DY_1$. For any $y \in \mathcal{Y}$, by Assumption 2, $Y - DC(Y, Z) \leq y$ if and only if $Y_0 \leq y$ and $Y_1 - C(Y_1, Z) \leq y$, so that $\mathbb{P}(Y - DC(Y, Z) \leq y | Z) = \mathbb{P}(Y_0 \leq y, Y_1 - C(Y_1, Z) \leq y | Z)$. By Assumption 4 and Proposition 2, the latter is monotone non increasing in z. Hence, for any $y \in \mathcal{Y}$, $\mathbb{P}(Y - DC(Y, Z) > y | Z = z)$ is non decreasing in z, so for any ordered pair $z \geq \tilde{z}$, of the support of Z, (1.2) holds. In addition, $\mathbb{P}(Y - DC(Y, Z) \geq \underline{b} | Z = z) \geq \mathbb{P}(Y_0 \geq \underline{b} | Z = z) = 1$.

Conversely, suppose the non negative measurable function C on $\mathcal{Y} \times \mathcal{Z}$ satisfies (1.2) for all ordered pairs $z \geq \tilde{z}$ of the support of Z. Then, by the characterization of first order stochastic dominance in Proposition 2, $Y - DC(Y, Z)$ is stochastically monotone with respect to Z. Thus, if we define $Y_0 := Y - DC(Y, Z)$ and take $Y_1 := g^{-1}(Y - DC(Y, Z), Z)$ with $g(y, z) := y - C(y, z)$ and g^{-1} its inverse with respect to the first argument, the vector $(Y_0, Y_1 - C(Y_1, Z))$ is equal to $(Y - DC(Y, Z), Y - DC(Y, Z))$, which is also stochastically monotone with respect to Z. Hence Assumption 4 holds for such a pair (Y_0, Y_1). Since $Y_0 = Y - DC(Y, Z) = Y_1 - C(Y_1, Z)$, Assumption 2 is also trivially satisfied, since it places non constraint on D in case of a tie. Finally, $(1 - D)Y_0 + DY_1 = Y$. Indeed, when $D = 0$, we have $Y_0 = Y - DC(Y, Z) = Y$, and when $D = 1$, we have $Y_1 = g^{-1}(Y - DC(Y, Z)) = g^{-1}(Y - C(Y, Z)) = Y$ (by definition of g) as required. Hence, there exists a random vector (Y_0, Y_1, D, Z) where $((1 - D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 2, 3(2) and 4 are satisfied, and C therefore belongs to the identified set $\mathcal{C}(\pi)$. □

Lemma 3 (Monotone Envelope)**.** The following hold for $\underline{F}$ and $\bar{F}$ defined in (1.3):

(1) For each $z \in \mathcal{Z}$, $y \mapsto \underline{F}(y|z)$ and $y \mapsto \bar{F}(y|z)$ are cdfs.
(2) If $\tilde{F}(\cdot|z)$ is a cdf satisfying $\mathbb{P}(Y \leq y|z) \leq \tilde{F}(y|z) \leq \mathbb{P}(Y \leq y, D = 0|z) + \mathbb{P}(D = 1|z)1\{y \geq \underline{b}\}$ for all (y, z), and $z \mapsto \tilde{F}(y|z)$ non increasing in z for all y, then $\underline{F}(y|z) \leq \tilde{F}(y|z) \leq \bar{F}(y|z)$ for all (y, z).

Proof of Lemma 3. We prove the results relating to $\underline{F}$. The results relating to $\bar{F}$ are treated symmetrically. Let $\mathcal{M}$ be the set of non decreasing real valued functions from $\mathbb{R}$ to $[0, 1]$. Define the operator S by

$$\begin{array}{rccc} S: & 2^{\mathcal{M}} & \to & \mathcal{M} \\ & K & \mapsto & S(K), \end{array}$$

where $S(K)$ is defined for each $y \in \mathbb{R}$ by $S(K)(y) = \sup\{F(y); F \in K\}$. Note that $S(K)$ is uniquely defined and belongs to $\mathcal{M}$ as required. Next, define the operator R by

$$\begin{array}{rccc} R: & \mathcal{M} & \to & \mathcal{M} \\ & F & \mapsto & R(F), \end{array}$$

where $R(F)$ is defined for each $y \in \mathbb{R}$ by $R(F)(y) = \lim_{\tilde{y} \downarrow y} F(\tilde{y})$. Note that $R(F)$ is uniquely defined and belongs to $\mathcal{M}$ as required. In addition, R is idempotent: $R(F)$ is right-continuous by construction, so that $R(R(F)) = R(F)$, and monotone: if $F_1 \leq F_2$, then $R(F_1) \leq R(F_2)$.

Define $F_K := R(S(K))$. We first show that F_K is a cdf if K is a collection of cdfs bounded above by the cdf $\bar{F} := 1\{\cdot \geq \underline{b}\}$. Indeed, F_K is non decreasing by construction of S and T, F_K is right-continuous because R is idempotent, F_K tends to 1 at $+\infty$ since each element of K is a cdf. Finally, F_K tends to 0 at $-\infty$ since $0 \leq F_K \leq R(\bar{F})$. Define $K_z := \{F(\cdot|\tilde{z}) : \tilde{z} \geq z\}$. Then, F^e of Lemma 3 is equal for each $z \in \mathcal{Z}$ to $F^e(\cdot|z) = R(S(K_z))$. By construction, $S(K_z) \leq S(K_{\tilde{z}})$ for $\tilde{z} \geq z$. Hence, by monotonicity of R, $F^e(y|z)$ is monotone non increasing in z. Putting it all together, we obtain the first result of Lemma 3.

Next, if $(y,z) \mapsto \tilde{F}(y|z)$ is a conditional cdf, such that $\tilde{F}(y|z) \geq F(y|z)$ for each (y,z) and $\tilde{F}$ is monotone non increasing in z for each y, then $\tilde{F}(\cdot|z) \geq \sup_{\tilde{z} \geq z} \tilde{F}(\cdot|z) \geq \sup_{\tilde{z} \geq z} F(\cdot|z) = S(K_z)$. Now, $R(\tilde{F}(\cdot|z)) = \tilde{F}(\cdot|z)$ since the latter is right-continuous. Finally, monotonicity of R yields $\tilde{F}(\cdot|z) \geq R(S(K_z)) := F^e(\cdot|z)$, which is the second result of Lemma 3. □

Proof of Corollary 1. From Equation (1.5) proved in the main text for all $y \in \mathcal{Y}$, we get $L(y - C(y,z)|z) \leq \mathbb{P}(Y - C(Y,Z) \leq y - C(y,z), D = 1|z) \leq U(y - C(y,z)|z)$. By the definition of U^- and by the invertibility of $y \mapsto y - C(y,z)$, the second inequality implies $y - C(y,z) \geq U^-(F_1(y|z)|z)$ as desired. Similarly, by the definition of L_- and by the invertibility of $y \mapsto y - C(y,z)$, the first inequality implies $y - C(y,z) \leq L_-(F_1(y|z)|z)$. □

Proof of Corollary 2. (1) First, we show that $y \mapsto y - \underline{C}(y,z)$ is increasing, hence invertible since it is continuous. The function $y \mapsto \underline{F}(y|z) - \mathbb{P}(Y \leq y, D = 0|z)$ is increasing and continuous and thus, so is $y \mapsto L(y|z) := \underline{F}(y|z) - \mathbb{P}(Y \leq y, D = 0|z)$. So $y - \underline{C}(y,z) = L_-\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right) = L^{-1}\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right)$ is increasing in y for all $z \in \mathcal{Z}$ because both $L^{-1}(y|z)$ and $\mathbb{P}(Y \leq y, D = 1|z)$ are increasing. (2) Second, we prove that $\underline{C}(y,z)$ is non negative. We have $L(y|z) = \underline{F}(y|z) - \mathbb{P}(Y \leq y, D = 0|z) \geq \mathbb{P}(Y \leq y|z) - \mathbb{P}(Y \leq y, D = 0|z) = \mathbb{P}(Y \leq y, D = 1|z)$. So we have that $L(y|z) \geq \mathbb{P}(Y \leq y, D = 1|z)$ and we already know that $L^{-1}(y|z)$ is increasing and continuous, so that $L^{-1}\left(L(y|z)|z\right) \geq L^{-1}\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right)$, hence $y \geq L_-\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right)$ or $0 \leq y - L_-\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right) = \underline{C}(y,z)$, as desired. (3) We now show that $\mathbb{P}(Y - \underline{C}(Y,z) \leq y, D = 1|z) = \underline{F}(y|z) - \mathbb{P}(Y \leq y, D = 0|z)$. As we have shown before, $\underline{C}(y,z) = y - L^{-1}\left(\mathbb{P}(Y \leq y, D = 1|z)|z\right)$. Hence $\mathbb{P}(Y - \underline{C}(Y,z) \leq y, D = 1|z) = \mathbb{P}(L^{-1}(F_1(Y|z)|z) \leq y, D = 1|z) = \mathbb{P}(F_1(Y|z) \leq L(y|z), D = 1|z) = L(y,z)$, since F_1 is continuous and increasing. (4) Finally, by Theorem 1, it suffices to show that $Y - D\underline{C}(Y,Z)$ is stochastically monotone with respect to Z. We have

$$\begin{aligned}
\mathbb{P}(Y - D\underline{C}(Y,Z) \leq y|z) &= \mathbb{P}(Y - D\underline{C}(Y,Z) \leq y, D = 1|z) + \mathbb{P}(Y - D\underline{C}(Y,Z) \leq y, D = 0|z) \\
&= \mathbb{P}(Y - \underline{C}(Y,Z) \leq y, D = 1|z) + \mathbb{P}(Y \leq y, D = 0|z) \\
&= \underline{F}(y|z) - \mathbb{P}(Y \leq y, D = 0|z) + \mathbb{P}(Y \leq y, D = 0|z) \;=\; \underline{F}(y|z)
\end{aligned}$$

and $\underline{F}(y|z)$ is stochastically monotone by construction.

□

Proof of Proposition 1. We show Proposition 1(2). Inverses are with respect to the first variable only. Differentiating $u_0(u_0^{-1}(u,z),z) = u$ with respect to z yields

$$\nabla_z u_0^{-1}(u,z) = -\left(\frac{\partial u_0}{\partial y}(u_0^{-1}(u,z),z)\right)^{-1} \nabla_z u_0(u_0^{-1}(u,z),z). \tag{A.1}$$

Plugging (A.1) taken at $u = u_1(y,z)$ into the expression for $\nabla_z C(y,z)$ yields:

$$\nabla_z C(y,z) = \left(\frac{\partial u_0}{\partial y}(u_0^{-1}(u_1(y,z),z),z)\right)^{-1} \left(\nabla_z u_0(u_0^{-1}(u_1(y,z),z),z) - \nabla_z u_1(y,z)\right). \tag{A.2}$$

The result follows. □

Proof of Theorem 2. Fix any $\pi \in \Pi$, and any $C \in \mathcal{C}^i(\pi)$. By Definition 4, there exists a random vector (Y_0, Y_1, D, Z) where $((1-D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 7 and 8 are satisfied. Call $Y := (1-D)Y_0 + DY_1$. By Assumption 7, if $D = 1$, then $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}] \geq \mathbb{E}[Y_0|\mathcal{I}]$. Similarly, if $D = 0$, then $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \mathbb{E}[Y_0|\mathcal{I}] \geq \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}]$. Hence, $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}] = \max\{\mathbb{E}[Y_0|\mathcal{I}], \mathbb{E}[Y_1 - C(Y_1,Z)|\mathcal{I}]\}$. By Imperfect Foresight Monotonicity (Assumption 8), $\mathbb{E}[Y - DC(Y,Z)|\mathcal{I}]$ is therefore stochastically monotone with respect to Z. Since Z is $\mathcal{I}$-measurable, iterated expectations yields monotonicity of the function $z \mapsto \mathbb{E}[Y - DC(Y,Z)|Z = z]$ as required.

Conversely, suppose the non negative measurable function C on $\mathcal{Y} \times \mathcal{Z}$ is such that $\mathbb{E}[Y - DC(Y,Z)|Z = z]$ is monotonic non decreasing in z. Thus, if we define $Y_0 := Y - DC(Y,Z)$ and take $Y_1 := g^{-1}(Y - DC(Y,Z))$ with $g(y,z) := y - C(y,z)$, the vector $(Y_0, Y_1 - C(Y_1,Z))$ is equal to $(Y - DC(Y,Z), Y - DC(Y,Z))$, and therefore satisfies Assumption 8. Assumption 2 is also trivially satisfied, since it places non constraint on D in case of a tie. Finally, $(1-D)Y_0 + DY_1 = Y$. Hence, there exists a random vector (Y_0, Y_1, D, Z) where $((1-D)Y_0 + DY_1, D, Z)$ has distribution π and Assumptions 7 and 8 are satisfied, and C therefore belongs to the identified set $\mathcal{C}^i(\pi)$. □

Proof of Corollary 3. Consider a π such that $\tilde{\mathcal{C}}^i(\pi)$ is non empty and consider an arbitrary function C in $\tilde{\mathcal{C}}^i(\pi)$. By Theorem 2, $\mathbb{E}[Y - DC(Z)|Z = z] = \mathbb{E}[Y|Z = z] - p(z)C(z)$ is non decreasing on the support of Z. Hence $\mathbb{E}[Y|Z = z] - p(z)C(z) \leq \inf_{\tilde{z} \geq z} \mathbb{E}[Y|Z = \tilde{z}]$. So $C(z) \geq \underline{C}(z)$, by the definition of $\underline{C}$ in Equation (3.2). Conversely, we now show that $\underline{C}(z)$ rationalizes the data under the imperfect foresight model. To see this, define $Y_1 = Y + (1-D)\underline{C}(Z)$ and $Y_0 = Y - D\underline{C}(Z)$. We have $Y = DY_1 + (1-D)Y_0$ and $\mathbb{E}[Y_1 - \underline{C}(Z)|Z] = \mathbb{E}[Y_0|Z]$, so that Assumption 7 is trivially satisfied. Finally, $\mathbb{E}[Y_0|Z = z] = \mathbb{E}[Y_1 - \underline{C}(Z)|Z = z] = \mathbb{E}[Y - D\underline{C}(Z)|Z = z] = \mathbb{E}[Y|Z = z] - p(z)\underline{C}(z) = \inf_{\tilde{z} \geq z} \mathbb{E}[Y|Z = \tilde{z}]$, which is non decreasing by construction, so that Assumption 8 is also satisfied, as required. □

Proof of Theorem 3. Given a distribution $\pi \in \Pi$, let C be in $\mathcal{C}^r(\pi)$. We give the proof for the upper bound on the distribution of costs. The lower bound is treated symmetrically. By Assumption GRM, $Y - DC = \max\{Y_0, Y_1 - C\}$, which is stochastically monotone with respect to Z. Call $\tilde{F}(\cdot|z)$ the cumulative distribution function of $Y - DC$ conditionally on $Z = z$ and write $\tilde{Y} := Y - DC$. Since $\tilde{Y} = Y$ when $D = 0$, the cumulative distribution of $\tilde{Y}$ conditional on $Z = z$ and $D = 1$ is

$$\tilde{F}(y|z,1) := \frac{1}{p(z)}\left[\tilde{F}(y|z) - F(y|z)\right]1\{p(z) > 0\} + F(y|z,1).$$

By construction, $C = Y - \tilde{Y}$ when $D = 1$, so that conditionally on $D = 1$, C is the difference between Y with cumulative distribution function $F(\cdot|z,1)$ and $\tilde{Y}$ with cumulative distribution function $\tilde{F}(\cdot|z,1)$. Hence, by Proposition 1 in Rüschendorf [1982], $\mathbb{P}(C \leq y|Z = z, D = 1) \leq 1 + \inf_{\tilde{y}} \min\{0, F(\tilde{y}|z), 1) - \tilde{F}(\tilde{y} - y|z,1)\}$, which is the upper bound on the cumulative distribution of the difference between two random variables with given marginals.

By the construction of $\underline{F}(\cdot|z)$ in (1.3,) and the fact that $\tilde{Y}$ is stochastically monotone with respect to Z and that $\tilde{Y} = Y - DC \leq Y$, we have $\tilde{F}(y|z) \geq \underline{F}(y|z)$ for any y by Lemma 3. Hence

$$\begin{aligned}\tilde{F}(y|z,1) &= \frac{1}{p(z)}\left[\tilde{F}(y|z) - F(y|z)\right]1\{p(z) > 0\} + F(y|z,1)\\ &\geq \frac{1}{p(z)}\left[\underline{F}(y|z) - F(y|z)\right]1\{p(z) > 0\} + F(y|z,1) = \underline{F}(y|z,1),\end{aligned}$$

which in turn implies $\mathbb{P}(C \leq y|Z = z, D = 1) \leq 1 + \inf_{\tilde{y}} \min\{0, F(\tilde{y}|z,1) - \underline{F}(\tilde{y} - y|z,1)\}$ as required. □

Proof of Lemma 2. Under Assumption 5, $L(y|z)$ is continuous and increasing. Hence, we have:

$$\begin{aligned}
L(y|z) &= \sup_{\tilde{y}\leq y}\{\underline{F}(\tilde{y}|z)-\mathbb{P}(Y\leq\tilde{y},D=0|z)\}\\
&= \underline{F}(\tilde{y}|z)-\mathbb{P}(Y\leq\tilde{y},D=0|z)\\
&= \lim_{\tilde{y}\downarrow y}\sup_{\tilde{z}\leq z}\mathbb{P}(Y\leq\tilde{y}|\tilde{z})-\mathbb{P}(Y\leq y,D=0|z)\\
&= \lim_{\tilde{y}\downarrow y}\left\{\sup_{\tilde{z}\leq z}\mathbb{P}(Y\leq\tilde{y}|\tilde{z})-\mathbb{P}(Y\leq\tilde{y},D=0|z)\right\}\\
&= \sup_{\tilde{z}\leq z}G(y|z,\tilde{z}).
\end{aligned}$$

Since L is continuous and increasing, it can also be written

$$L(y|z)=\sup_{\tilde{y}\leq y}\sup_{\tilde{z}\leq z}G(\tilde{y}|z,\tilde{z})=\sup_{\tilde{z}\leq z}\sup_{\tilde{y}\leq y}G(\tilde{y}|z,\tilde{z}).$$

Defining $\tilde{G}(y|z,\tilde{z}):=\sup_{\tilde{y}\leq y}G(\tilde{y}|z,\tilde{z})$, $\tilde{G}(y|z,\tilde{z})\leq\sup_{\tilde{z}\leq z}\tilde{G}(y|z,\tilde{z})$ implies

$$L_-(x|z):=\sup\{y:L(y|z)\leq x\}\leq\tilde{G}_-(x|z,\tilde{z}):=\sup\{y:\tilde{G}(y|z,\tilde{z})\leq x\},$$

for all $\tilde{z}\leq z$. Hence $L_-(x|z)\leq\inf_{\tilde{z}\leq z}\tilde{G}_-(x|z,\tilde{z})$. Finally, by continuity of G and by the definition of $\tilde{G}$, we have $\tilde{G}_-(x|z,\tilde{z})\leq G_-(x|z,\tilde{z}):=\sup\{y:G(y|z,\tilde{z})\leq x\}$. The upper bound is treated similarly. □

Appendix B. Structural underpinnings of the imperfect foresight model

The selection rule of Assumption 7 can be interpreted as a transformation of a decision rule based on relative expected utilities in both sectors, assuming very specific utility functions. Assumption 2 is the special case of Assumption 7, where the vector (Y_0,Y_1) of potential outcomes is $\mathcal{I}$-measurable. Other examples include the following:

(1) Suppose utility in Sector d is $u_d(Y_d,Z):=Y_d+g_d(Z)$, and Sector d is chosen when $\mathbb{E}[u_d(Y_d,Z)|\mathcal{I}]>\mathbb{E}[u_{1-d}(Y_{1-d},Z)|\mathcal{I}]$. Then Assumption 7 holds with $C(y,z):=g_0(z)-g_1(z)$, as long as the latter is non negative.

(2) Suppose utility in Sector d, is $u_d(Y_d,Z):=g_d(Z)Y_d$, with $g_d>0$ for $d=0,1$, and Sector d chosen if $\mathbb{E}[u_d(Y_d,Z)|\mathcal{I}]>\mathbb{E}[u_{1-d}(Y_{1-d},Z)|\mathcal{I}]$. Then Assumption 7 holds with $C(y,z):=y(1-g_1(z)/g_0(z))$, as long as the latter is non negative.

(3) Suppose utility in Sector d is quadratic, i.e.,

$$u_d(Y_d,Z)=Y_d-\eta_d(Z)Y^2,\ \text{with } Y\leq\inf_z\min_d\{1/(2\eta_d(Z))\}\text{ a.s.},$$

and suppose $\mathbb{E}[Y_0^2|\mathcal{I}]/\mathbb{E}[Y_1^2|\mathcal{I}]=f(Z)\leq 1$, (given relative proportions of women in STEM and non STEM, the information from historical data about the price of women's talent in STEM is noisier), then Assumption 7 holds with $C(y,z):=(\eta_1(z)-\eta_0(z)f(z))y^2$, as long as the latter is non negative.

(4) Suppose utility in Sector d is isoelastic, i.e.,

$$u_d(Y_d)=\frac{Y_d^{1-\rho}-1}{1-\rho},$$

with relative risk aversion coefficient $\rho>1$. Suppose also that $Y_d=\exp(X_d)$, where X_d, conditionally on $Z=z$, is distributed according to $\mathrm{N}\big(\mu_d(z),\sigma_d^2(z)\big)$. Then expected utility is

$$\mathbb{E}[u_d(Y_d,Z)|Z=z]=\frac{\exp\{(1-\rho)\left(\mu_d(z)-(\rho-1)\sigma_d^2(z)/2\right)\}}{1-\rho},$$

and sector selection is based on the comparison of $\mathbb{E}[Y_1 - C(Y_1, Z)|\mathcal{I}]$ and $\mathbb{E}[Y_0|\mathcal{I}]$, with

$$C(y,z) = \left(1 - \exp\left(\frac{\sigma_0^2(z) - \sigma_1^2(z)}{2}\rho\right)\right) y, \text{ when } \sigma_0^2(z) < \sigma_1^2(z).$$

Hence, Assumption 7 holds if the variance of Sector 1 outcomes is higher than the variance of Sector 0 outcomes.

Appendix C. Point identification of the cost function

In this section, we describe how Equation (3.1) can be used to prove the identification result of d'Haultfœuille and Maurel [2013]. The extended Roy model in this section does not rely on the restriction that the cost function associated with Sector 1 be non negative. However, it assumes that the cost function is a function of exogenous observables only (here the vector Z) and that the potential outcomes have a separable representation summarized in the following statement of Assumptions in d'Haultfœuille and Maurel [2013].

Assumption 9. *Observable and potential outcomes are related by $Y = DY_1 + (1-D)Y_0$. The selection indicator D satisfies $\mathbb{E}[Y_d - dC(Z)|\mathcal{I}] > \mathbb{E}[Y_{1-d} - (1-d)C(Z)|\mathcal{I}] \Rightarrow D = d$, where Z is measurable with respect to the agent's information σ-algebra $\mathcal{I}$, and $\mathbb{E}[Y_1 - Y_0|\mathcal{I}] = \mathbb{E}[Y_1 - Y_0|Z] + V$ with $V \perp Z$.*

If C is a function of z only and $\mathbb{E}[Y_1 - Y_0|\mathcal{I}] = \mathbb{E}[Y_1 - Y_0|Z] + V$ with $V \perp Z$, then (3.1) yields $\mathbb{E}[Y - DC(Z)|\mathcal{I}] = \max\{0, \mathbb{E}[Y_1 - Y_0 - C(Z)|\mathcal{I}]\} + \mathbb{E}[Y_0|\mathcal{I}]$, hence $\mathbb{E}[Y - Y_0|\mathcal{I}] - C(Z)\mathbb{P}[D = 1|\mathcal{I}] = \max\{0, \mathbb{E}[Y_1 - Y_0|Z] - C(Z) + V\}$. Iterated expectations then yields

$$\mathbb{E}[Y - Y_0|Z] - C(Z)\mathbb{P}[D = 1|Z] = \mathbb{E}[\max\{0, \mathbb{E}[Y_1 - Y_0|Z] - C(Z) + V\}|Z]. \tag{C.1}$$

To differentiate both terms with respect to the first component of Z, we make the following regularity assumptions, where Z_{-1} (resp. z_{-1}) is the vector of components of Z (resp. z) excluding Z_1 (resp. z_1).

Assumption 10. *For all z_{-1} on the support of Z_{-1}, the functions $z_1 \mapsto C(z)$, $z_1 \mapsto \mathbb{E}[Y_d|D = d, Z = z]$, $d = 0, 1$, and $z_1 \mapsto \mathbb{E}[D|Z = z]$ are continuously differentiable on the support of Z_1 conditional on $Z_{-1} = z_{-1}$.*

By Assumption 10 and the dominated convergence theorem, the partial derivative of the right-hand side of (C.1) is

$$\frac{\partial}{\partial z_1}\mathbb{E}[\max\{0, \mathbb{E}[Y_1 - Y_0|Z] - C(Z) + V\}|Z] = \mathbb{P}[D = 1|Z = z]\frac{\partial}{\partial z_1}\left(\mathbb{E}[Y_1 - Y_0|Z] - C(Z)\right),$$

whereas the differentiating the left-hand side yields

$$\begin{aligned}(\mathbb{E}[Y - Y_0|Z = z] - C(z))&\frac{\partial}{\partial z_1}\mathbb{P}[D = 1|Z = z] \\ &+ \mathbb{P}[D = 1|Z = z]\frac{\partial}{\partial z_1}\left(\mathbb{E}[Y - Y_0|Z = z] - C(z)\right).\end{aligned}$$

Finally, we have

$$C(z)\ \frac{\partial}{\partial z_1}\mathbb{P}[D = 1|Z = z] \ = \ \frac{\partial}{\partial z_1}(\mathbb{E}[Y - Y_0|Z = z] - \mathbb{P}[D = 1|Z = z]\ \frac{\partial}{\partial z_1}(\mathbb{E}[Y_1 - Y_0|Z = z],$$

which can be rearranged into Equation (2.6) in d'Haultfœuille and Maurel [2013], and which identifies the cost function if conditional means of potential outcomes are identified.

Appendix D. Additional details on the inference procedure

Inference for the lower bound proceeds according to the following steps:

(1) Estimate the empirical counterparts of distribution functions $\mathbb{P}(Y \leq y|z)$, $\mathbb{P}(Y \leq y, D = 0|z)$ and $\mathbb{P}(D = 1|z)$ on a fine grid of Y, $\mathcal{G}_Y$ and a predefined grid on Z, $\mathcal{G}_Z = [0.12, 0.13, ..., 0.19]$, by local linear regression, using Epanechnikov kernels. Results presented are for $h = 0.066$. Qualitatively similar results obtained for $h/2$, and $h/3$.

(2) For each pair $(z, \tilde{z}) \in \mathcal{G}_Z \times \mathcal{G}_Z$, and $\tilde{y} \in \mathcal{G}_Y$, calculate:

$$L(\tilde{y}, \tilde{z}, z) = \hat{P}(Y \leq \tilde{y}|\tilde{z}) - \hat{P}(Y \leq \tilde{y}, D = 0|z).$$

and apply the monotonic transformation:

$$L^*_\varepsilon(\tilde{y}, \tilde{z}, z) = \max\left\{L(\tilde{y}, \tilde{z}, z), \max\left\{L(y, \tilde{z}, z) : y \in \mathcal{G}_Y, y \leq \tilde{y}\right\} + \varepsilon\right\}.$$

The parameter ε ensures the strict monotonicity of the empirical counterpart.

(3) For each $y \in \mathcal{G}_Y$ invert L to obtain:

$$\hat{L}_-(F_1(y|z)|z, \tilde{z}) = \sup\left\{\tilde{y} : L^*_\varepsilon(\tilde{y}, \tilde{z}, z) \leq \hat{F}_1(y|z)\right\}$$

(4) Bootstrap $\hat{L}_-(F_1(y|z)|z, \tilde{z})$ to obtain the covariance matrix of estimator evaluated at each $(y, z, \tilde{z}) \in \mathcal{G}_Y \times \mathcal{G}_X \times \mathcal{G}_Y$. With this, calculate a uniform (in y and z) critical value of level α for the test:

$$H_0 : \inf\left\{\tilde{z} \geq z : L_-(F_1(y|z)|z, \tilde{z}) \leq 0\right\}$$

using the methodology of Chernozhukov, Lee and Rosen (2013), applying the Adaptive Inequality selection. In practice, considering all values of $y \in \mathcal{G}_Y$ results in many redundant inequalities and decreases the power of the test. We consider a subset of $\mathcal{G}_Y$ for the computation of the critical value. Invert the test to obtain a one sided confidence interval for $L_-(F_1(y|z)|z))$.

(5) The one sided confidence interval for $\underline{C}(y, z)$ follows immediately.

Inference for the upper bound is similar, adapting steps (2) to (4).

References

J. Altonji, P. Arcidiacono, and A. Maurel. The analysis of field choice in college and graduate schools: determinants and wage effects. *Handbook of the Economics of Education*, 5:305–396, 2016.

P. Arcidiacono, J. H. A. Maurel, and T. Romano. Ex ante returns and occupational choice. preprint, 2019.

F. Baillet, A. Franken, and A. Weber. DZHW graduate panel 2009: Data and methods report on the graduate panel 2009 (1st and 2nd survey waves). Technical report, German Centre for Higher Education Research and Science Studies, 2017.

P. Bayer, S. Khan, and C. Timmins. Nonparametric identification and estimation in a Roy model with common nonpecuniary returns. *Journal of Business and Economic Statistics*, 29:201–215, 2011.

D. Beede, T. Julian, D. Langdon, G. McKittrick, B. Khan, and M. Doms. Women in stem: a gender gap to innovation. Department of Commerce, Economics and Statistics Administration, 2011.

R. Blundell, A. Gosling, H. Ichimura, and C. Meghir. Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, 75:323–336, 2007.

T. Breda, J. Grenet, M. Monnet, and C. van Effenterre. Can female role models reduce the gender gap in science? evidence from classroom intervention in french high schools. halshs-01713068, 2018.

D. Card. Estimating the return to schooling: progress on some persistent econometric problems. *Econometrica*, 69:1127–1160, 2001.

V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96:559–575, 2009.

V. Chernozhukov, S. Lee, and A. Rosen. Inference on intersection bounds. *Econometrica*, 81:667–737, 2013.

F. Cunha, J. Heckman, and S. Navarro. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers*, 57:191–261, 2005.

T. N. Daymont and P. J. Andrisani. Job preferences, college major, and the gender gap in earnings. *Journal of Human Resources*, pages 408–428, 1984.

A. Delavande and B. Zafar. University choice: the role of expected earnings, nonpecuniary outcomes, and financial constraints. *Journal of Political Economy*, 127:2343–2393, 2019.

X. d'Haultfœuille and A. Maurel. Inference on an extended Roy model, with an application to schooling decisions in france. *Journal of Econometrics*, 174:95–106, 2013.

P. Eisenhauer, J. Heckman, and Mosso. Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments. *International Economic Review*, 56:331–357, 2015a.

P. Eisenhauer, J. Heckman, and E. Vytlacil. The generalized Roy model and the cost-benefit analysis of social programs. *Journal of Political Economy*, 123:413–443, 2015b.

J. Heckman and E. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96:4730–4734, 1999.

J. Heckman, J. Humphries, and G. Veramendi. Dynamic treatment effects. *Journal of Econometrics*, 191:276–292, 2016.

J. Heckman, J. Humphries, and G. Veramendi. Returns to education: the causal effects of education on earnings, health, and smoking. *Journal of Political Economy*, 126:S197–S246, 2018.

J. Hunt, J.-P. Garant, H. Herman, and D. Munroe. Why are women underrepresented amongst patentees? *Research Policy*, 42:831–843, 2013.

S. Kahn and D. Ginther. Women and stem. NBER Working Paper No. 23525, 2017.

G. Kaplan and S. Schulhofer-Wohl. The changing (dis-)utility of work. *Journal of Economic Perspectives*, 32:239–258, 2018.

Q. Li and J. Racine. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, 31:57–65, 2008.

Q. Lippmann and C. Senik. Math, girls and socialism. *Journal of Comparative Economics*, 46:874–888, 2018.

C. Manski and J. Pepper. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica*, 68:997–1010, 2000.

A. Mas and A. Pallais. Valuing alternative work arrangements. *American Economic Review*, 107:3722–3759, 2017.

I. Mourifié, M. Henry, and R. Méango. Sharp bounds and testability of a Roy model of STEM major choices. forthcoming, *Journal of Political Economy*, 2020.

C. Riegle-Crumb and C. Moore. The gender gap in high school physics: considering the context of local communities. *Social Science Quarterly*, 95:253–268, 2014.

L. Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14:623–632, 1982.

F. Saltiel. What's math got to do with it? multidimensional ability and the gender gap in stem. unpublished manuscript, 2018.

M. Shaked and G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

C. Sloane, E. Hurst, and D. Black. A cross-cohort analysis of human capital specialization and the college gender wage gap. Becker Friedman Institute working paper number 2019-121, 2019.

R. Thaler. Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, 1:39–60, 1980.

A. Tversky and D. Kahneman. Loss aversion in riskless choice: a reference-dependent model. *Quarterly Journal of Economics*, 106:1039–1061, 1991.

M. Wiswal and B. Zafar. Determinants of college major choice: identification using an information experiment. *Review of Economic Studies*, 82:791–824, 2015.

M. Wiswal and B. Zafar. Preference for the workplace, investment in human capital, and gender. *Quarterly Journal of Economics*, 133:457–507, 2018.

Y. Xie, M. Fang, and K. Shauman. Stem education. *Annual Review of Sociology*, 41: 331–357, 2015.

B. Zafar. College major choice and the gender gap. *Journal of Human Resources*, 48(3): 545–595, 2013.

The Pennsylvania State University

Munich Center for the Economics of Aging

University of Toronto