LCERPA

Laurier Centre for Economic Research & Policy Analysis

LCERPA Working paper No. 2025-9

Updated May 2026

A Dynamic Factor Model of Skill Formation and Mental Health: Counterfactual Analysis of Interventions for Social Mobility

Cecilia S. Diaz-Campo
Olin Business School, Washington University in St. Louis

M. Antonella Mancino
Wilfrid Laurier University

Salvador Navarro
University of Western Ontario

A Dynamic Factor Model of Skill Formation and Mental Health: Counterfactual Analysis of Interventions for Social Mobility*

Cecilia S. Diaz-Campo† M. Antonella Mancino‡ Salvador Navarro§

May 12, 2026

Abstract

This paper develops a dynamic latent factor framework to characterize the joint evolution of latent skills and outcomes under treatment and selection. Building on the skill-formation literature, we introduce explicit innovation terms in the law of motion, allowing the distribution of latent factors to evolve endogenously over time rather than being re-normalized period by period, and establish nonparametric identification under standard measurement and independence conditions. We then specialize the framework to study the long-run mental health consequences of incarceration for justice-involved youth. We leverage longitudinal data from the Pathways to Desistance study, modeling incarceration as a three-category treatment (no, early, recent). Our estimates reveal that incarceration tends to have substantial adverse effects on mental health on average, particularly for youths with stronger baseline skills who are rarely incarcerated, with marked heterogeneity across individuals. We also show that mental health mobility across counterfactual incarceration scenarios exhibits a U-shaped relationship with baseline cognitive skills.

Keywords: Bayesian Analysis, Dynamics, Mental Health, Incarceration, Counterfactuals, Mobility

JEL Code: C11, I10, K00

*The authors acknowledge funding from the Social Sciences and Humanities Research Council of Canada. Data used for this project were supported by the National Institute on Drug Abuse through a cooperative agreement that calls for scientific collaboration between the grantees and the National Institute on Drug Abuse staff.

†Olin Business School, Washington University in St. Louis. Email: cdiazcampo@wustl.edu.

‡Department of Economics, Wilfrid Laurier University. Email: amancino@wlu.ca.

§Department of Economics, University of Western Ontario. Email: snavarr@uwo.ca.

1 Introduction

Mental health is a central determinant of individual well-being and a key driver of life outcomes, with well-documented effects on educational attainment and labor market outcomes (see, e.g., Currie and Stabile, 2006; Currie et al., 2010; Wang et al., 2023; Biasi et al., 2025; Diaz-Campo et al., 2025; Jolivet and Postel-Vinay, 2025). Concurrently, mental health is shaped by a range of socioeconomic experiences and institutional exposures, among which interactions with the justice system, and incarceration in particular, play a prominent role (see, e.g., Turney et al., 2012; Haglund et al., 2014; Salazar, 2020; Bhuller et al., 2025).[1] Understanding the consequences of incarceration for mental health therefore represents a critical area of social scientific inquiry and public policy concern.

Following Cunha et al. (2006), this paper develops a methodology to construct joint distributions of counterfactuals, and applies it to study how incarceration affects mental health for a sample of youth offenders. Counterfactual analysis allows us to ask what would have happened to an individual's mental health trajectory had they, for instance, not experienced incarceration, or had their pre-existing cognitive and mental health vulnerabilities been different. Such analyses are essential for disentangling the effects of incarceration itself from pre-existing individual heterogeneity and other life circumstances. Moreover, understanding the heterogeneous impacts of incarceration, i.e., recognizing that its effects may vary substantially across individuals with different baseline observed and *unobserved* characteristics, is crucial for evaluating the effects of incarceration policy. Traditional causal inference design-based approaches provide a clean, robust understanding of average causal effects, at the price of often limiting our ability to move beyond aggregate statistics to see who is most affected and how.

Our goal is to move beyond estimating average effects to identify how incarceration shifts the entire distribution of mental health outcomes for different individuals. By recovering the joint distribution of counterfactual outcomes, we can move beyond average summaries of effects and answer questions like which groups, as determined by their observed and unobserved characteristics, are affected by incarceration and how they are affected.

[1] For evidence on the consequences of incarceration on labor market outcomes, see e.g., Kling (2006); Mueller-Smith (2015); Gordon et al. (2023); Garin et al. (2025); Mancino (2025); on recidivism, see e.g., Durlauf and Nagin (2010); Kuziemko (2013); Bhuller et al. (2020); Rose and Shem-Tov (2021); and on overall health, see e.g., Binswanger et al. (2007, 2009); Hjalmarsson and Lindquist (2022); Norris et al. (2024).

We introduce a dynamic factor model designed to address these challenges by analyzing the evolution of mental health among justice-involved youth from the Pathways to Desistance study. We explicitly model latent cognitive skills and mental health at baseline, their role in selection into incarceration over a seven-year period, and their influence on subsequent mental health outcomes. Our approach builds upon the framework of Cunha et al. (2010) by incorporating multiple, interacting latent traits within a dynamic setting but makes a key departure. Rather than re-normalizing the latent factor distribution in each period, we introduce explicit innovation terms in the law of motion. This allows factor means and variances to evolve endogenously according to their law of motion, without requiring normalizations on the evolved factor distribution (see Agostinelli and Wiswall, 2025; Freyberger, 2025).

We also allow treatment effects to depend directly on baseline latent skills through interaction terms, capturing essential heterogeneity (Heckman et al., 2006). Combined with nonparametric identification of factor distributions via characteristic function arguments, this framework yields the joint distribution of counterfactual outcomes conditional on the entire vector of latent skills and treatment, allowing for rich patterns of unobserved heterogeneity and dependence.

By leveraging this econometric structure, we aim to: (1) estimate the distribution of causal impacts of incarceration on mental health at the seven-year follow-up, accounting for selection on observed and unobserved characteristics; (2) explore heterogeneous and dynamic treatment effects (Heckman and Navarro, 2007; Fruehwirth et al., 2016), particularly how the impact of incarceration varies with baseline cognitive skills and mental health status as well as the timing of treatment; and (3) analyze the distributional dynamics of mental health by examining both positional mobility, how individuals shift their relative standing within the mental health distribution under incarceration versus no incarceration; and absolute growth, measuring changes in individuals' mental health levels within the overall distribution. In doing so, we seek to provide a comprehensive understanding of the interplay between individual vulnerabilities, justice system involvement, and mental health trajectories. The methodology allows us to move beyond simple descriptive statements to a more robust analysis of the distributional consequences of incarceration, in the spirit of constructing and evaluating policy counterfactuals as advocated by Cunha et al. (2006).

Our analysis reveals that incarceration has substantial and heterogeneous negative effects on

mental health among justice-involved youth, even after accounting for selection on observed and unobserved factors. We uncover how incarceration shifts the entire distribution of mental health outcomes, with individuals experiencing incarceration generally showing worse mental health trajectories. We identify notable heterogeneity: the impact of incarceration varies with initial cognitive skills and mental health status, with certain subgroups disproportionately affected. Beyond average effects, we document significant mobility in mental health rankings induced by incarceration, with individuals in the middle of the mental health distribution exhibiting the greatest rank changes, while those at the extremes showing greater persistence. Our findings highlight how incarceration limits upward mobility in mental health, especially among lower-ranked individuals.

Beyond the empirical application, the paper also makes a methodological contribution. Section 2 develops a general dynamic factor framework with treatment and evolving latent factors and provides a nonparametric identification result for the joint distribution of these factors and potential outcomes. Establishing nonparametric identification at this general level is valuable on its own, since it clarifies which assumptions are used for identification, independently of any particular parametric specification used in applications.[2] This result is of independent interest: it shows that, by introducing explicit innovation terms in the law of motion, one can identify the evolving factor distribution without imposing restrictions on its period-by-period form, and it applies more generally than the empirical specification in Section 3, where only one factor is allowed to evolve.

The paper proceeds as follows. Section 2 develops the general econometric framework for identifying joint distributions of latent factors, including the treatment assignment model and dynamic evolution across time. Section 3 tailors this general framework to the Pathways to Desistance data, specifying measurement systems for cognitive skills and mental health, and treatment rules relevant to incarceration and mental health dynamics. Section 4 presents estimation results and assesses model fit through graphical and statistical checks. Section 5 conducts counterfactual analyses of incarceration's distributional impacts on mental health, exploring heterogeneity, mobility, and growth in outcomes. Finally, Section 6 concludes with a discussion of the implications of our findings and avenues for future research.

[2]See Roehrig (1988); Matzkin (2007); Hauck and Woutersen (2026).

2 Identifying Joint Distributions of Counterfactuals using Dynamic Factor Models

This section develops a general dynamic factor model and establishes nonparametric identification of the joint distribution of latent factors, treatment, and outcomes over time. The focus is on the identification problem itself. We work with a linear law of motion augmented with explicit innovation terms, and show how to recover the evolving factor distribution without imposing additional restrictions on its functional form. The empirical model in Section 3 can be viewed as a special case of this underlying nonparametrically identified structure.

We begin by assuming the analyst has access to a collection of observed variables, such as psychometric scores or standardized tests, that depend on a smaller set of latent factors, for example, cognitive skills or socioemotional traits. Some of these measurements may be observed repeatedly over time. The analyst may be interested in the joint dynamics of these measurements and/or of the latent variables.

To simplify, we assume two periods, $t \in \{1, 2\}$, two latent factors, $(\theta_{a,i,t}, \theta_{b,i,t})$, and a binary treatment, $D_i \in \{0, 1\}$. For individual i and period t, the analyst observes a vector of measurements $\{Y_{i,j,t}\}_{j=1}^{J_t}$. All measures are assumed continuous. We omit covariates for clarity; they can be reintroduced with standard modifications. Extensions to allow for additional factors, discrete or mixed outcomes, multiple dynamic treatments, or richer panel structures can be done following the literature (see e.g., Jöreskog and Goldberger, 1975; Jöreskog, 1977; Carneiro et al., 2003; Heckman and Navarro, 2007; Cunha et al., 2010; Fruehwirth et al., 2016; Williams, 2020) and we pursue some of them in our application.

2.1 Measurement Model in Period 1

In period 1, each measurement satisfies:

$$Y_{i,j,1} = \gamma_{j,1} + \theta_{a,i,1}\alpha_{a,j,1} + \theta_{b,i,1}\alpha_{b,j,1} + \varepsilon_{i,j,1}, \quad j = 1, ..., J_1. \tag{1}$$

The θ's are latent factors, α's are loadings, and ε's are measurement errors or "uniquenesses."

We invoke the following set of assumptions to identify the system

Assumption 1.

Location: $\mathbb{E}[\theta_{a,1}] = \mathbb{E}[\theta_{b,1}] = \mathbb{E}[\varepsilon_{j,1}] = 0, \quad j = 1, ..., J_1.$

Dedicated Measures: $\alpha_{b,1,1} = \alpha_{b,2,1} = 0$; $\alpha_{a,3,1} = \alpha_{a,4,1} = 0.$

Scale and Sign: $\alpha_{a,1,1} = 1$, $\alpha_{b,3,1} = 1.$

Independence: $(\theta_{a,1}, \theta_{b,1}) \perp\!\!\!\perp \{\varepsilon_{j,1}\}_{j=1}^{J_1}$, *and* $\varepsilon_{j,1} \perp\!\!\!\perp \varepsilon_{j',1}$ *for* $j \neq j'$.

These assumptions are sufficient but not necessary. The location restrictions help identify the means $\gamma_{j,1}$. The dedicated measures condition ensures that at least two items load exclusively on a single factor; weaker versions are also valid (e.g., Williams, 2020). Scale and sign normalizations determine the magnitude and direction of the factors. For example, if measure 1 was an IQ score, more of factor $\theta_{a,1}$ would be associated with a higher IQ score (direction), and the factor would be measured in IQ units (magnitude). Independence between factors and measurement errors facilitates identification. Some form of independence between measures and the factors is needed, although it does not need to be as strong as the one we impose here (e.g., Cunha et al., 2010). Similarly, it is possible to allow for some forms of correlation between the ε's.

From observed covariances, we can identify the covariance between factors from

$$\text{cov}(Y_{1,1}, Y_{3,1}) = \text{cov}(\theta_{a,1}, \theta_{b,1}) \equiv \rho_1. \tag{2}$$

We can identify the loadings from

$$\alpha_{a,2,1} = \frac{\text{cov}(Y_{2,1}, Y_{3,1})}{\rho_1}, \alpha_{b,4,1} = \frac{\text{cov}(Y_{1,1}, Y_{4,1})}{\rho_1}, \tag{3}$$

and the factor variances

$$\sigma_a^2 = \frac{\text{cov}(Y_{1,1}, Y_{2,1})}{\alpha_{a,2,1}}, \sigma_b^2 = \frac{\text{cov}(Y_{3,1}, Y_{4,1})}{\alpha_{b,4,1}}. \tag{4}$$

Provided a rank condition holds, the remaining loadings ($j > 4$) are identified from

$$\begin{aligned} \text{cov}(Y_{1,1}, Y_{j,1}) &= \alpha_{a,j,1}\sigma_a^2 + \alpha_{b,j,1}\rho_1, \\ \text{cov}(Y_{3,1}, Y_{j,1}) &= \alpha_{a,j,1}\rho_1 + \alpha_{b,j,1}\sigma_b^2. \end{aligned} \tag{5}$$

Finally, we can recover the variance of the uniqueness terms from

$$\text{var}(Y_{j,1}) = \alpha_{a,j,1}^2\sigma_a^2 + \alpha_{b,j,1}^2\sigma_b^2 + 2\alpha_{a,j,1}\alpha_{b,j,1}\rho_1 + \sigma_{\varepsilon_{j,1}}^2. \tag{6}$$

2.1.1 Identification of Distributions

Once the parameters of the measurement system are identified, we turn to identifying the joint distribution of the latent factors $(\theta_{a,1}, \theta_{b,1})$ and the uniquenesses nonparametrically. We begin by applying Kotlarski's theorem (Kotlarski, 1967; Rao, 1992), which shows that the marginal distribution of a latent variable and two independent measurement errors can be recovered from their sum under independence assumptions.

To proceed, define the demeaned variables $\tilde{Y}_{j,1} = Y_{j,1} - \gamma_{j,1}$. From the measurement system and dedicated measures assumptions, we know that

$$\begin{aligned} \tilde{Y}_{i,1,1} &= \theta_{a,i,1} + \varepsilon_{i,1,1}, \\ \frac{\tilde{Y}_{i,2,1}}{\alpha_{a,2,1}} &= \theta_{a,i,1} + \frac{\varepsilon_{i,2,1}}{\alpha_{a,2,1}}. \end{aligned} \tag{7}$$

Since these expressions are two independent noisy measurements of the same latent variable $\theta_{a,1}$, and the noise terms are mutually independent and mean-zero, Kotlarski's result applies, and the marginal distributions of $\theta_{a,1}, \varepsilon_{1,1}, \varepsilon_{2,1}$ are identified nonparametrically. The same reasoning applies for $\theta_{b,1}$ using measurements $\tilde{Y}_{3,1}$ and $\tilde{Y}_{4,1}$.

To identify the joint distribution of $(\theta_{a,1}, \theta_{b,1})$, take the characteristic function of $\left(\tilde{Y}_{1,1}, \tilde{Y}_{3,1}\right)$

$$\begin{aligned} \phi_{\tilde{Y}_{1,1},\tilde{Y}_{3,1}}(y_1, y_3) &= \mathbb{E}\left[e^{i\left((\theta_{a,1}+\varepsilon_{1,1})y_1 + \left(\theta_{b,1}+\varepsilon_{3,1}\right)y_3\right)}\right] \\ &= \mathbb{E}\left[e^{i\left(\theta_{a,1}y_1 + \theta_{b,1}y_3\right)} e^{i(\varepsilon_{1,1}y_1 + \varepsilon_{3,1}y_3)}\right] \\ &= \phi_{\left(\theta_{a,1},\theta_{b,1}\right)}(y_1, y_3) \times \phi_{(\varepsilon_{1,1},\varepsilon_{3,1})}(y_1, y_3). \end{aligned} \tag{8}$$

The left-hand side is observable from data, and $\phi_{(\varepsilon_{1,1},\varepsilon_{3,1})}(y_1, y_3)$ is identified from marginal distributions via earlier steps. We recover the joint characteristic function of the latent variables by

$$\phi_{\left(\theta_{a,1},\theta_{b,1}\right)}(y_1, y_3) = \frac{\phi_{\tilde{Y}_{1,1},\tilde{Y}_{3,1}}(y_1, y_3)}{\phi_{(\varepsilon_{1,1},\varepsilon_{3,1})}(y_1, y_3)}, \tag{9}$$

which identifies the joint distribution of $(\theta_{a,1}, \theta_{b,1})$, assuming regularity conditions and non-vanishing characteristic functions.

Higher-order moments provide an alternative route. For example:

$$\begin{aligned}
\mathbb{E}\left[\tilde{Y}_{1,1}^2 \tilde{Y}_{2,1}\right] &= \alpha_{a,2,1}\mathbb{E}\left[\theta_{a,1}^3\right], \\
\mathbb{E}\left[\tilde{Y}_{3,1}^2 \tilde{Y}_{4,1}\right] &= \alpha_{b,4,1}\mathbb{E}\left[\theta_{b,1}^3\right], \\
\mathbb{E}\left[\tilde{Y}_{1,1}^2 \tilde{Y}_{3,1}\right] &= \mathbb{E}\left[\theta_{a,1}^2 \theta_{b,1}\right], \\
\mathbb{E}\left[\tilde{Y}_{1,1} \tilde{Y}_{3,1}^2\right] &= \mathbb{E}\left[\theta_{a,1} \theta_{b,1}^2\right].
\end{aligned} \tag{10}$$

Hence, if all moments exist, taking higher-order moments identifies the joint distribution of the latent factors (Billingsley, 1995).

This identification result is a core feature of the methodology, distinguishing it from others that make stronger distributional assumptions or rely on rank preservation or perfect dependence.

2.2 Treatment Assignment

We now consider a binary treatment that occurs between periods 1 and 2. The probability of being treated is modeled as a threshold-crossing rule driven by latent factors. Define the selection index

$$T_i = \gamma_T\left(X_i\right) + \theta_{a,i,1}\alpha_{a,T} + \theta_{b,i,1}\alpha_{b,T} + \varepsilon_{i,T}, \tag{11}$$

where X_i is a vector of observables that affect the probability of receiving treatment. The individual is treated if $T_i > 0$, so the treatment indicator is

$$D_i = \mathbb{1}\left\{T_i > 0\right\}. \tag{12}$$

We assume $\varepsilon_{i,T}$ is mean-zero and independent of everything else. To normalize the scale in this latent variable model, we impose $\mathrm{Var}\left(\varepsilon_T\right) = 1$.

Under standard identification arguments from discrete choice models (Manski, 1988; Matzkin, 1992), $\gamma_T\left(X_i\right)$ and the distribution of the sum $\left(\theta_{a,1}\alpha_{a,T} + \theta_{b,1}\alpha_{b,T} + \varepsilon_T\right)$ are identified. Using conditional distributions of outcomes and treatment assignment

$$F_{Y_{j,1}|D=d}\left(y \mid X = x\right)\Pr\left(D = d \mid X = x\right) \tag{13}$$

and varying the evaluation points via y, x, we can identify the joint distribution of

$$\theta_{a,1}\alpha_{a,j,1} + \theta_{b,1}\alpha_{b,j,1} + \varepsilon_{j,1}, \text{ and } \theta_{a,1}\alpha_{a,T} + \theta_{b,1}\alpha_{b,T} + \varepsilon_T. \tag{14}$$

This identifies the cross-covariances of these random variables, which in turn identify $\alpha_{a,T}, \alpha_{b,T}$ (see e.g., Cunha et al., 2006; Fruehwirth et al., 2016).

2.3 Identification in Period 2: Dynamics and Evolving Latent Structure

Let $Y_{i,j,2}$ denote the measurements in the second period. The measurement system is assumed to be structurally similar to that in period 1

$$Y_{i,j,2} = \gamma_{j,2} + \theta_{a,i,2}\alpha_{a,j,2} + \theta_{b,i,2}\alpha_{b,j,2} + \varepsilon_{i,j,2}, \quad j = 1, ..., J_2. \tag{15}$$

The latent factors evolve according to a first-order law of motion:[3]

$$\theta_{a,i,2} = D_i\delta_{a,T} + \theta_{a,i,1}\lambda_{a,a} + \theta_{b,i,1}\lambda_{b,a} + \theta_{a,i,1}D_i\lambda_{a,a,T} + \theta_{b,i,1}D_i\lambda_{b,a,T} + U_{a,i}, \tag{16}$$

$$\theta_{b,i,2} = D_i\delta_{b,T} + \theta_{a,i,1}\lambda_{a,b} + \theta_{b,i,1}\lambda_{b,b} + \theta_{a,i,1}D_i\lambda_{a,b,T} + \theta_{b,i,1}D_i\lambda_{b,b,T} + U_{b,i}. \tag{17}$$

We maintain the following assumptions

Assumption 2.

Location: $\mathbb{E}[U_a] = \mathbb{E}[U_b] = \mathbb{E}[\varepsilon_{j,2}] = 0, \quad j = 1, ..., J_2.$

Dedicated Measures: $\alpha_{b,1,2} = \alpha_{b,2,2} = 0$; $\alpha_{a,3,2} = \alpha_{a,4,2} = 0.$

Scale and Sign: $\alpha_{a,1,2} = 1$, $\alpha_{b,3,2} = 1.$

Independence: $(\theta_{a,1}, \theta_{b,1}) \perp\!\!\!\perp \{\varepsilon_{j,2}\}_{j=1}^{J_2}$, $(\theta_{a,1}, \theta_{b,1}) \perp\!\!\!\perp (U_a, U_b)$, $(U_a, U_b) \perp\!\!\!\perp \{\varepsilon_{j,2}\}_{j=1}^{J_2}$, *and* $\varepsilon_{j,2} \perp\!\!\!\perp \varepsilon_{j',2}$ *for* $j \neq j'$.

Assumption 2 imposes the same measurement restrictions as Assumption 1 (dedicated measures, scale and sign normalizations, location restrictions, and independence), but serves a different purpose in the dynamic structure. Rather than re-normalizing the latent skill distribution in each period and treating $(\theta_{a,2}, \theta_{b,2})$ as period-specific factors (Cunha et al., 2010), we maintain a common latent structure $(\theta_{a,t}, \theta_{b,t})$ across time and introduce explicit innovation terms U_a and U_b in the law of motion.[4] The approach we follow imposes that the law of motion is linear in the period 1

[3] Notice that this model allows for what Heckman et al. (2006) refer to as essential heterogeneity.

[4] The canonical approach in Cunha et al. (2010) normalizes the factor distribution period by period. Agostinelli and Wiswall (2025) develops an estimation strategy that highlights how such re-normalizations affect technology parameters, while Freyberger (2025) analyzes when scale and location restrictions in skill formation models are innocuous versus when they induce misspecification. Our specification aligns with this literature by separating the latent stock from innovations rather than re-scaling the factor each period.

factors, whereas the alternative approach accommodates more general functional forms (e.g., CES production functions as in Cunha et al., 2010).

The critical distinction is that all location and independence restrictions in period 2 apply to the innovations (U_a, U_b) rather than to the factors $(\theta_{a,2}, \theta_{b,2})$ themselves. This has two implications. First, the period-2 factors need not be mean zero; we impose $\mathbb{E}[U_a] = \mathbb{E}[U_b] = 0$ instead, so the means of $(\theta_{a,2}, \theta_{b,2})$ are determined endogenously by the law of motion and the period-1 distribution of $(\theta_{a,1}, \theta_{b,1}, D)$. Second, while we normalize the dedicated loadings $\alpha_{a,1,2}$ and $\alpha_{b,3,2}$ to one, the same scale could equivalently be fixed by imposing sign restrictions on these loadings together with a variance normalization on (U_a, U_b) (for example, $\text{var}(U_a) = \text{var}(U_b) = 1$). In both cases, the scale and dispersion of $(\theta_{a,2}, \theta_{b,2})$ are pinned down by the identified law of motion and the distribution of (U_a, U_b). No additional ad hoc normalizations on the period-2 factor distribution are required beyond those implied by the measurement system and the innovation process.

2.3.1 Identifying the Law of Motion for θ_2

We now show how the parameters

$$\lambda_{a,a},\ \lambda_{b,a},\ \lambda_{a,b},\ \lambda_{b,b},\ \lambda_{a,a,T},\ \lambda_{b,a,T},\ \lambda_{a,b,T},\ \lambda_{b,b,T},\ \delta_{a,T},\ \delta_{b,T}$$

are identified from the joint distribution of $(Y_{1,1}, Y_{3,1}, Y_{1,2}, Y_{3,2}, D)$.

Baseline coefficients $\lambda_{a,a}, \lambda_{b,a}$

First, consider the untreated group $D = 0$. We have that

$$(Y_{1,2} \mid D = 0) = \gamma_{1,2} + \theta_{a,1}\lambda_{a,a} + \theta_{b,1}\lambda_{b,a} + U_a + \varepsilon_{1,2}. \tag{18}$$

Using the dedicated period–1 measures,

$$Y_{1,1} = \gamma_{1,1} + \theta_{a,1} + \varepsilon_{1,1}, \qquad Y_{3,1} = \gamma_{3,1} + \theta_{b,1} + \varepsilon_{3,1}, \tag{19}$$

with $\varepsilon_{1,1}, \varepsilon_{3,1}$ independent of $(\theta_{a,1}, \theta_{b,1}, U_a, \varepsilon_{1,2})$, we obtain

$$\text{cov}(Y_{1,2}, Y_{1,1} \mid D = 0) = \lambda_{a,a}\,\text{var}(\theta_{a,1} \mid D = 0) + \lambda_{b,a}\,\text{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 0), \tag{20}$$

$$\text{cov}(Y_{1,2}, Y_{3,1} \mid D = 0) = \lambda_{a,a}\,\text{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 0) + \lambda_{b,a}\,\text{var}(\theta_{b,1} \mid D = 0). \tag{21}$$

The left-hand sides are observable covariances of Y, and the conditional variances and covariance of $(\theta_{a,1}, \theta_{b,1})$ are identified from the period–1 measurement system. Under the rank condition that the conditional covariance matrix of $(\theta_{a,1}, \theta_{b,1}) \mid D = 0$ is nonsingular, (20)–(21) forms a 2×2 linear system in $(\lambda_{a,a}, \lambda_{b,a})$ with a unique solution.

Baseline coefficients $\lambda_{a,b}, \lambda_{b,b}$

The same argument applied to $\theta_{b,2}$ identifies $(\lambda_{a,b}, \lambda_{b,b})$. For $D = 0$, combining (17) and (**??**) gives

$$(Y_{3,2} \mid D = 0) = \gamma_{3,2} + \theta_{a,1}\lambda_{a,b} + \theta_{b,1}\lambda_{b,b} + U_b + \varepsilon_{3,2}. \tag{22}$$

Analogous calculations yield

$$\begin{aligned} \operatorname{cov}(Y_{3,2}, Y_{1,1} \mid D = 0) &= \lambda_{a,b} \operatorname{var}(\theta_{a,1} \mid D = 0) + \lambda_{b,b} \operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 0), \\ \operatorname{cov}(Y_{3,2}, Y_{3,1} \mid D = 0) &= \lambda_{a,b} \operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 0) + \lambda_{b,b} \operatorname{var}(\theta_{b,1} \mid D = 0). \end{aligned} \tag{23}$$

Under the same rank condition as before, the coefficients $(\lambda_{a,b}, \lambda_{b,b})$ are uniquely determined.

Interaction coefficients $\lambda_{\cdot,\cdot,T}$

We now exploit the treated group $D = 1$. From the dedicated period–2 measure we have,

$$(Y_{1,2} \mid D = 1) = \gamma_{1,2} + \theta_{a,1}(\lambda_{a,a} + \lambda_{a,a,T}) + \theta_{b,1}(\lambda_{b,a} + \lambda_{b,a,T}) + \delta_{a,T} + U_a + \varepsilon_{1,2}. \tag{24}$$

Therefore,

$$\operatorname{cov}(Y_{1,2}, Y_{1,1} \mid D = 1) = (\lambda_{a,a} + \lambda_{a,a,T})\operatorname{var}(\theta_{a,1} \mid D = 1) + (\lambda_{b,a} + \lambda_{b,a,T})\operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 1), \tag{25}$$

$$\operatorname{cov}(Y_{1,2}, Y_{3,1} \mid D = 1) = (\lambda_{a,a} + \lambda_{a,a,T})\operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 1) + (\lambda_{b,a} + \lambda_{b,a,T})\operatorname{var}(\theta_{b,1} \mid D = 1). \tag{26}$$

Subtracting (20)–(21) from (25)–(26), and using that $(\lambda_{a,a}, \lambda_{b,a})$ are already known, we can write

$$\begin{aligned} \Delta_{11} &\equiv \operatorname{cov}(Y_{1,2}, Y_{1,1} \mid D = 1) - \operatorname{cov}(Y_{1,2}, Y_{1,1} \mid D = 0) \\ &= K_{11} + \lambda_{a,a,T} \operatorname{var}(\theta_{a,1} \mid D = 1) + \lambda_{b,a,T} \operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 1), \\ \Delta_{31} &\equiv \operatorname{cov}(Y_{1,2}, Y_{3,1} \mid D = 1) - \operatorname{cov}(Y_{1,2}, Y_{3,1} \mid D = 0) \\ &= K_{31} + \lambda_{a,a,T} \operatorname{cov}(\theta_{a,1}, \theta_{b,1} \mid D = 1) + \lambda_{b,a,T} \operatorname{var}(\theta_{b,1} \mid D = 1), \end{aligned}$$

where K_{11} and K_{31} collect the known baseline terms

$$K_{11} \equiv \lambda_{a,a}\big(\mathrm{var}(\theta_{a,1} \mid D=1) - \mathrm{var}(\theta_{a,1} \mid D=0)\big) + \lambda_{b,a}\big(\mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=1) - \mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=0)\big),$$

$$K_{31} \equiv \lambda_{a,a}\big(\mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=1) - \mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=0)\big) + \lambda_{b,a}\big(\mathrm{var}(\theta_{b,1} \mid D=1) - \mathrm{var}(\theta_{b,1} \mid D=0)\big).$$

Defining $\tilde{\Delta}_{11} \equiv \Delta_{11} - K_{11}$ and $\tilde{\Delta}_{31} \equiv \Delta_{31} - K_{31}$, we obtain the linear system

$$\begin{cases} \tilde{\Delta}_{11} = \lambda_{a,a,T}\,\mathrm{var}(\theta_{a,1} \mid D=1) + \lambda_{b,a,T}\,\mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=1) \\ \tilde{\Delta}_{31} = \lambda_{a,a,T}\,\mathrm{cov}(\theta_{a,1}, \theta_{b,1} \mid D=1) + \lambda_{b,a,T}\,\mathrm{var}(\theta_{b,1} \mid D=1) \end{cases},$$

which has a unique solution for $(\lambda_{a,a,T}, \lambda_{b,a,T})$ under the rank condition that the conditional covariance matrix of $(\theta_{a,1}, \theta_{b,1}) \mid D=1$ is nonsingular.

Repeating the same argument for $Y_{3,2}$ and $\theta_{b,2}$ yields two additional equations identifying $(\lambda_{a,b,T}, \lambda_{b,b,T})$. Thus, all interaction parameters $\lambda_{\cdot,\cdot,T}$ are identified.

Identifying the intercepts $\gamma_{1,2}, \gamma_{3,2}$ and treatment intercepts $\delta_{a,T}, \delta_{b,T}$

We begin by forming

$$\mathbb{E}[Y_{1,2} \mid D=0] = \gamma_{1,2} + \lambda_{a,a}\,\mathbb{E}[\theta_{a,1} \mid D=0] + \lambda_{b,a}\,\mathbb{E}[\theta_{b,1} \mid D=0]. \tag{27}$$

The conditional mean $\mathbb{E}[Y_{1,2} \mid D=0]$ is observable, the period–1 means $\mathbb{E}[\theta_{a,1} \mid D=0]$ and $\mathbb{E}[\theta_{b,1} \mid D=0]$ are identified from the period–1 measurement system, and $(\lambda_{a,a}, \lambda_{b,a})$ have been identified in the previous subsection. Equation (27) therefore identifies $\gamma_{1,2}$.

For the treated group $D=1$ we get

$$\begin{aligned} \mathbb{E}[Y_{1,2} \mid D=1] = \gamma_{1,2} &+ \lambda_{a,a}\,\mathbb{E}[\theta_{a,1} \mid D=1] + \lambda_{b,a}\,\mathbb{E}[\theta_{b,1} \mid D=1] \\ &+ \delta_{a,T} + \lambda_{a,a,T}\,\mathbb{E}[\theta_{a,1} \mid D=1] + \lambda_{b,a,T}\,\mathbb{E}[\theta_{b,1} \mid D=1]. \end{aligned} \tag{28}$$

All terms other than $\delta_{a,T}$ in (28) are now known. Consequently, (28) identifies $\delta_{a,T}$.

An entirely analogous argument applied to the dedicated b–item $Y_{3,2} = \gamma_{3,2} + \theta_{b,2} + \varepsilon_{3,2}$, together with the law of motion for $\theta_{b,2}$ in (17), first identifies $\gamma_{3,2}$ from $\mathbb{E}[Y_{3,2} \mid D=0]$ and then identifies $\delta_{b,T}$ from $\mathbb{E}[Y_{3,2} \mid D=1]$.

2.3.2 Factor means and joint distributions in period 2

Taking unconditional expectations of the law of motion in (16)–(17) and using Assumption 2 together with the already identified joint distribution of $(\theta_{a,1}, \theta_{b,1}, D)$ from Section 2.1 and the treatment assignment model, we obtain the unconditional means $\mathbb{E}[\theta_{a,2}]$ and $\mathbb{E}[\theta_{b,2}]$.

Given Assumption 2, the period–2 measurement system in (15) is a standard two–factor model with dedicated measures, scale and sign normalizations, and independent uniquenesses. With the location of $(\theta_{a,2}, \theta_{b,2})$ now pinned down, the same covariance–based arguments as in Section 2.1 imply that the period–2 factor loadings $\{\alpha_{a,j,2}, \alpha_{b,j,2}\}_{j=1}^{J_2}$, the covariance matrix of $(\theta_{a,2}, \theta_{b,2})$, and the variances of the measurement errors $\{\varepsilon_{j,2}\}_{j=1}^{J_2}$ are identified.

Applying the nonparametric deconvolution arguments used in the "Identification of Distributions" subsection for period 1 to the period–2 system then yields the joint distribution of $(\theta_{a,2}, \theta_{b,2})$ and the period–2 uniquenesses. Combining this joint distribution with the already identified joint distribution of $(\theta_{a,1}, \theta_{b,1}, D)$ and the law of motion (16)-(17) identifies the joint distribution of $(\theta_{a,1}, \theta_{b,1}, \theta_{a,2}, \theta_{b,2}, D)$ and, in particular, the joint distribution of the innovation vector (U_a, U_b), which is a known affine transformation of these latent variables.

2.4 Counterfactual Analysis

With the assumptions and procedures above, we identify all factor loadings across periods and selection equations, all factor and uniqueness distributions nonparametrically, the law of motion for latent variables across periods, and the joint distributions of factors, treatment decisions, and outcomes. These results provide the full structure necessary for counterfactual analysis and recovery of potential outcomes. In the next section, we formalize this structure to generate and evaluate counterfactual outcomes, including nonparametric identification of standard and non-standard treatment effect parameters.

2.4.1 Potential Outcomes and Counterfactual Distributions

An alternative interpretation for the law of motion in equation (16) is that it describes a model of potential outcomes. If we explicitly write subscripts 0 and 1 to refer to the potential untreated

and treated states, the factors in period 2 as functions of the treatment would be $(\theta_{a,i,2,0}, \theta_{a,i,2,1})$. It follows that the realized factor for individual i is

$$\theta_{a,i,2} = D_i \theta_{a,i,2,1} + (1 - D_i)\, \theta_{a,i,2,0}, \tag{29}$$

and similarly for $\theta_{b,i,2}$. This gives us a potential outcomes representation of observed measurements $(Y_{i,j,2,0}, Y_{i,j,2,1})$ as well.

Given nonparametric identification, one can go beyond the representation. We can fully characterize the distribution of $(\theta_{a,2,0}, \theta_{a,2,1})$ and $(\theta_{b,2,0}, \theta_{b,2,1})$ and, by mapping through the measurement system, the joint distribution of $(Y_{j,2,0}, Y_{j,2,1})$.

Having recovered the joint distribution of potential outcomes, we can define the return

$$\begin{aligned} \Delta_{\theta_{a,i,2}} &= \theta_{a,i,2,1} - \theta_{a,i,2,0} \\ &= \delta_{a,T} + \theta_{a,i,1} \lambda_{a,a,T} + \theta_{b,i,1} \lambda_{b,a,T}, \end{aligned} \tag{30}$$

which can be used to compute standard mean treatment parameters.

2.4.2 Distributional and Mobility Counterfactuals

Beyond means, nonparametrically recovering the joint distribution of potential outcomes enables analysis of more detailed questions. For example, we can estimate the fraction of the population that would benefit from universal treatment

$$\Pr\left(\Delta_{\theta_{a,i,2}} > 0\right), \tag{31}$$

or the fraction amongst the treated who benefit from treatment

$$\Pr\left(\Delta_{\theta_{a,i,2}} > 0 \mid D_i = 1\right). \tag{32}$$

We can also calculate counterfactual mobility parameters. For example, we can calculate a mobility table where we look at the probability that an individual who starts in the p^{th}-percentile of the untreated distribution, $\theta_{a,2,0}$, ends in the q^{th}-percentile of the treated distribution, $\theta_{a,2,1}$.

While these examples illustrate the advantage of recovering the joint distribution of potential outcomes, they are not exhaustive (see e.g., Cunha et al., 2006). More generally, parameters specified as functions of $\theta_{a,2}$ may also be formulated in terms of $\theta_{b,2}$ or in terms of the observed measures

$Y_{j,2}$.

2.4.3 Policy Counterfactuals

We are not restricted to working with the counterfactual distribution of the observed treatment. The structure also supports simulation and evaluation of alternative treatment assignment rules. For two distinct policies (I and II) resulting in different treatment rules,[5] label the respective treatment indicators D^I and D^{II}. The approach allows calculation of, for example, the proportion of individuals whose treatment status changes between policies, or of the proportion of these "switchers" who benefit from the alternative policy

$$\Pr\left(\Delta_{\theta_{a,i,2}} > 0 \mid D^I \neq D^{II}\right). \tag{33}$$

The framework we just described allows us to obtain a richer understanding of the inequality and social mobility consequences of existing policies, and of policy changes, than are available from measures based on mean treatment parameters. Furthermore, we can analyze these changes holding the period fixed, i.e., without the need to wait for time to elapse. Finally, we can use our analysis to generate various counterfactual states even if they have never been observed.

3 Empirical Application: Incarceration and Mental Health in the Pathways to Desistance Data

In this section, we apply our methodology to examine how incarceration shapes mental health trajectories among young offenders. We begin by describing the data used in the analysis. We then specialize the model in Section 2 to account for specific features of our data.

3.1 Data

This study utilizes data from the Pathways to Desistance (PD) project, a longitudinal study that tracks the development and life trajectories of adolescent offenders as they transition into early

[5] For example, because there is a change in how one of the $X's$ is treated between policies.

adulthood. The sample comprises youth who were adjudicated for serious offenses in either juvenile or adult court in Maricopa County, Arizona, or Philadelphia County, Pennsylvania, between November 2000 and January 2003. Eligible participants were between 14 and 18 years old at the time of their offense and were required to provide informed assent or consent. A total of 1,354 individuals enrolled in the study, corresponding to a 67% enrollment rate.[6]

Participants completed a baseline interview shortly after their adjudication (within 75 days for those in the juvenile system, and within 90 days after decertification hearings or arraignments for those in the adult system, depending on the site). The study includes six follow-up interviews conducted every six months and four annual follow-ups thereafter. Interviews were typically held in participants' homes or secure facilities if they were incarcerated. The panel data span up to seven years for each individual. To encourage participation and reduce attrition, individuals received $50 for the baseline interview, with payments increasing for subsequent follow-ups (Monahan et al., 2009). The study maintained high retention rates, exceeding 90% for the first six waves and remaining at or above 83% for the later annual waves.

An important feature of the PD dataset is its detailed longitudinal information on criminal activity and incarceration. In addition, it includes rich data on mental health and a wide range of individual characteristics that may be relevant for predicting incarceration. These features are crucial for analyzing the relationship between incarceration and mental health outcomes.

The baseline survey includes basic demographic information, such as age, gender, ethnicity, and location. It also contains data on incarceration before the baseline. Specifically, the survey records whether individuals were ever detained in a juvenile facility or jail before the event that led them into the PD survey. Incarceration data are subsequently collected in each follow-up survey wave. These data are based on self-reports gathered during each wave of the survey.[7] At each follow-up, individuals report whether they were incarcerated during the recall period, the type of facility (e.g., prison, jail, or detention center), and the number of months spent in each facility.

The baseline survey includes a rich set of self-reported measures of mental health symptoms from the Brief Symptom Inventory (BSI). The BSI is a self-report tool where participants assess

[6]For more information on the PD study see Schubert et al. (2004).

[7]To encourage accurate self-reporting, responses are kept confidential, and participants are provided a certificate of confidentiality from the U.S. Department of Justice.

the degree to which they have been disturbed by any of 53 possible symptoms over the past week. Responses are rated on a scale of 0 ("Not at All"), 1 ("A Little Bit"), 2 ("Moderately"), 3 ("Quite a Bit"), and 4 ("Extremely"), with higher scores reflecting more severe mental health symptoms. The BSI includes nine subscales, each targeting a specific symptom group. These include: somatization, obsessive-compulsive behaviors, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. Each subscale is an average of between 4 and 7 individual symptoms, and hence can take more than 5 values for each subscale. The BSI measures are collected repeatedly in the follow-up surveys.

Lastly, the baseline survey also incorporates results from a comprehensive array of standardized psychometric tests administered to each individual to measure cognitive skills. Cognitive measures include scores from the Wechsler Abbreviated Scale of Intelligence (WASI), which provides an estimate of overall intellectual functioning (IQ) derived from two subtests: Vocabulary and Matrix Reasoning. In addition, the data include results from two neurological tests designed to assess cognitive dysfunction associated with impairment of the brain's frontal cortex: the Stroop Color-word test and the Trail Making Test. The Stroop Test examines cognitive interference in reading ability through three timed tasks involving color words, colored symbols, and mismatched word-color combinations. Participants are instructed to read words or name ink colors as quickly and accurately as possible. The Trail Making Test assesses overall brain function and potential neurological impairment through two parts. Part A requires individuals to connect a sequence of numbered circles in order, while Part B involves alternating between numbers and letters in sequential order. Relevant to our model, follow-up surveys do not collect information on these cognitive measures.

We restrict our sample to individuals who have not been treated before the beginning of the study (i.e., we exclude those ever detained before the baseline survey). The final sample comprises 686 individuals with baseline information on at least one cognitive or mental health measure.

Table 1 presents descriptive statistics for our sample. Men make up the majority of the sample, with women comprising 17.1% of participants. The sample is almost evenly divided between Phoenix (44.6%) and Philadelphia (55.4%). A substantial share of the sample consists of minority individuals, with 42.3% identifying as Black and 30.5% as Hispanic. Seven years after the baseline survey, 34.1% of individuals have never been incarcerated, 45.0% have experienced at least

one month of incarceration during that period, of which 31.2% were first incarcerated early (i.e., years 1 to 2 after baseline) and 13.8% more recently (i.e., years 3 to 7 after baseline), and 20.8% have missing information on incarceration. In Section 3.2, we discuss how we control for potential nonrandom attrition in incarceration.

Table 1: Pathways to Desistance Study - Summary Statistics - Demographics

	Mean	SD	N
Age at Baseline			
Share Age 14	0.134	0.341	686
Share Age 15	0.203	0.402	686
Share Age 16	0.325	0.469	686
Share Age 17	0.265	0.442	686
Share Age 18	0.073	0.260	686
Share Female	0.171	0.376	686
Ethnicity			
Share White	0.226	0.419	686
Share Hispanic	0.305	0.461	686
Share Black	0.423	0.494	686
Share Other	0.047	0.211	686
Share Phoenix	0.446	0.497	686
First Incarceration			
Have Not Been Incarcerated	0.341	0.474	686
Early Incarceration (Years 1-2 After Baseline)	0.312	0.464	686
Recent Incarceration (Years 3-7 After Baseline)	0.138	0.346	686
Missing Information	0.208	0.407	686

Notes: The descriptive statistics reported in this table correspond to data from the baseline survey, except for the incarceration indicator, which is based on both baseline and follow-up surveys.

Table 2 presents descriptive statistics for self-reported mental health symptoms at baseline and seven years later. At baseline, for each of the nine measures, only a minority of individuals report experiencing symptoms to at least a mild degree on average (i.e., a little bit, moderately, quite a bit, or extremely), with rates ranging from 8.5% for phobic anxiety to 33.5% for paranoid ideation. Seven years after the baseline survey, 55.2% of individuals have missing information on mental health.[8] Among those with observed outcomes, all nine measures indicate a modest reduction in the proportion of individuals reporting symptoms at or above a mild level on average. For instance, among individuals without missing information, the share reporting at least mild somatization symptoms decreases from 12.3% to 10.0%. This pattern holds across all measures, suggesting a

[8]In Section 3.2, we discuss how we control for potential nonrandom attrition in mental health seven years after.

slight improvement in mental health over time.

Table 2: Pathways to Desistance Study - Summary Statistics - Mental Health Measures

	Mean - Baseline	Mean - Year 7	N
1. BSI Somatization			
Not At All (0)	0.380	0.249	686
Between 0 and 1	0.421	0.153	686
1 or More	0.112	0.045	686
Missing	0.086	0.552	686
2. BSI Depression			
Not At All (0)	0.310	0.239	686
Between 0 and 1	0.417	0.137	686
1 or More	0.187	0.071	686
Missing	0.086	0.552	686
3. BSI Anxiety			
Not At All (0)	0.347	0.203	686
Between 0 and 1	0.411	0.187	686
1 or More	0.156	0.058	686
Missing	0.086	0.552	686
4. BSI Hostility			
Not At All (0)	0.236	0.140	686
Between 0 and 1	0.415	0.233	686
1 or More	0.262	0.074	686
Missing	0.086	0.552	686
5. BSI Obsessive-Compulsive			
Not At All (0)	0.213	0.146	686
Between 0 and 1	0.443	0.203	686
1 or More	0.258	0.099	686
Missing	0.086	0.552	686
6. BSI Interpersonal Sensitivity			
Not At All (0)	0.411	0.299	686
Between 0 and 1	0.353	0.093	686
1 or More	0.150	0.055	686
Missing	0.086	0.552	686
7. BSI Phobic Anxiety			
Not At All (0)	0.570	0.313	686
Between 0 and 1	0.259	0.108	686
1 or More	0.085	0.026	686
Missing	0.086	0.552	686
8. BSI Paranoid Ideation			
Not At All (0)	0.156	0.138	686
Between 0 and 1	0.423	0.187	686
1 or More	0.335	0.122	686
Missing	0.086	0.552	686
9. BSI Psychoticism			
Not At All (0)	0.334	0.255	686
Between 0 and 1	0.418	0.141	686
1 or More	0.162	0.051	686
Missing	0.086	0.552	686

Notes: This table reports descriptive statistics for mental health measures from the Brief Symptom Inventory (BSI) at the baseline survey and the last follow-up survey (i.e., seven years after baseline). The BSI comprises nine subscales, each corresponding to a specific grouping of symptoms: somatization, depression, anxiety, hostility, obsessive-compulsive behaviors, interpersonal sensitivity, phobic anxiety, paranoid ideation, and psychoticism. For each symptom, responses are rated on a scale of 0 ("Not At All"), 1 ("A Little Bit"), 2 ("Moderately"), 3 ("Quite a Bit"), and 4 ("Extremely"), with higher scores indicating more severe mental health symptoms.

Table 3 reports descriptive statistics for the tests designed to measure cognitive skills. The average IQ score among participants is notably lower than the general population mean of 100, with only 10% of individuals scoring above that benchmark. For each of the Stroop Measures, scores above 40 are typically regarded as within the normal range. At the baseline, 53.1%, 36.0%, and 20.6% have scores below normal for the Color, Word, and Color/Word tests, respectively. For cognitive impairment, the Trail-Making assessments classify individuals into one of four categories, with the two lowest indicating some degree of impairment. In our sample, 20.3% show signs of mild, moderate, or severe impairment based on Trail-Making A, and 35.1% based on Trail-Making B.

3.2 An Empirical Model of Skills and Mental Health Dynamics

In this section, we adapt the general framework developed in Section 2 to an empirical setting focused on the joint dynamics of cognitive skills and mental health. While generalizing the model along some dimensions, we also tailor it to the specific features of our data. The two latent factors $\theta_{a,t}$ and $\theta_{b,t}$ in the theoretical model correspond here to latent cognitive skills C and latent mental health M (and its seven-year evolution M_7), respectively, and the treatment indicator D refers to incarceration over the seven-year period. All measurement equations, treatment-selection equations, and laws of motion introduced below are special cases of the more general dynamic factor structure developed in Section 2.

At baseline, each individual i is endowed with a level of cognitive skill, denoted by C_i, and a level of mental health, denoted by M_i. We also observe a set of background characteristics X_i, which are allowed to correlate with observed outcomes but are assumed to be independent of latent factors and measurement errors. Associated with these latent states are vectors of measurements: $\{\mathcal{C}_{i,j}\}_{j=1}^{N_C}$ for cognitive skill and $\{\mathcal{M}_{i,j}\}_{j=1}^{N_M}$ for mental health.

For mental health, we employ the nine subscales from the BSI ($N_M = 9$). All nine are treated as ordered discrete measures. For each subscale j, the number of values (K_j) corresponds to the number of distinct values where we group the "1 or More" into a single category. We use six measures of cognitive skills ($N_C = 6$). There are four continuous measures: the WASI IQ and the three Stroop scores, and two Trail-Making scores, which are measured on an ordered discrete scale.[9] The vector

[9]The Trail-Making scores are reverse-coded so that higher scores reflect less cognitive impairment, aligning them

Table 3: Pathways to Desistance Study - Summary Statistics - Cognitive Skills

(a) Panel A: WASI IQ

	Percentile	N
1%	55	681
5%	63	681
10%	67	681
25%	77	681
50%	86	681
75%	94	681
90%	102	681
95%	107	681
99%	116	681

(b) Panel B: Stroop Measures

	Share of Sample	N
Stroop Color		
Share Below or Equal 40	0.531	686
Share Above 40	0.453	686
Share Missing	0.016	686
Stroop Word		
Share Below or Equal 40	0.360	686
Share Above 40	0.624	686
Share Missing	0.016	686
Stroop Color/Word		
Share Below or Equal 40	0.206	686
Share Above 40	0.778	686
Share Missing	0.016	686

(c) Panel C: Trail Making

	Share of Sample	N
Trail Making Part A		
Perfectly Normal	0.402	686
Normal	0.388	686
Mild/Moderately Impaired	0.134	686
Moderately/Severely Impaired	0.069	686
Missing	0.007	686
Trail Making Part B		
Perfectly Normal	0.362	686
Normal	0.281	686
Mild/Moderately Impaired	0.261	686
Moderately/Severely Impaired	0.090	686
Missing	0.006	686

Notes: The descriptive statistics reported in this table correspond to data from the baseline survey. The estimate of general intellectual ability (IQ) is based on two subtests: Vocabulary and Matrix Reasoning. The Stroop Color/Word Test assesses the effects of interference on reading ability through three parts, measuring interference from words, colors, and combined word-color. Scores on the Stroop tests are continuous, with values above 40 considered within the normal range. The Trail-Making Test measures general brain function and consists of two parts: A and B. Scores take one of four values, with the lowest two indicating mild/moderate or moderate/severe impairment.

with the remaining cognitive measures.

of individual characteristics, X_i, includes dummies for age, gender, ethnicity, and location.

We define a latent structure that links each measurement to the latent factors. For the cognitive measurements, we assume

$$\mathcal{C}^*_{i,j} = X_i'\gamma_{\mathcal{C},j} + C_i\psi_j + \varepsilon_{i,\mathcal{C},j}, \quad j \in \{1, ..., N_C\}. \tag{34}$$

Analogously, for the mental health measurements

$$\mathcal{M}^*_{i,j} = X_i'\gamma_{\mathcal{M},j} + M_i\mu_j + \varepsilon_{i,\mathcal{M},j}, \quad j \in \{1, ..., N_M\}. \tag{35}$$

When the corresponding observed measure is continuous, we set the measured outcome equal to the latent one, e.g., $\mathcal{C}_{i,j} = \mathcal{C}^*_{i,j}$. For discrete, ordered outcomes taking values $k = 1, ..., K_j$, we impose a standard ordered threshold structure:

$$\mathcal{M}_{i,j} = k \quad \text{if} \quad o_{\mathcal{M},j,k-1} < \mathcal{M}^*_{i,j} \leq o_{\mathcal{M},j,k}, \quad k = 1, ..., K_j, \tag{36}$$

with $o_{\mathcal{M},j,0} = -\infty$, $o_{\mathcal{M},j,1} = 0$, and $o_{\mathcal{M},j,K_j} = \infty$. Similar conventions are applied to other measures.[10]

Identification proceeds analogously to the system described in Section 2,[11] with factors now encompassing cognitive skill and mental health. The measurement errors ε represent idiosyncratic deviations from the factor structure and are assumed to be mean-zero and mutually independent. As before, we assume that C and M are independent of the measurement errors. For scale and sign normalization, we set $\psi_1 = 1$ and $\mu_4 = -1$. Namely, the cognitive skills factor is normalized to have a loading of one on the WASI IQ score, while the mental health factor is normalized to have a loading of negative one on the BSI hostility measure.[12]

Treatment is defined as a three-category measure of incarceration timing over the seven-year period following baseline: no incarceration ($D_i = 0$), early incarceration ($D_i = 1$), and recent incarceration ($D_i = 2$). Early incarceration refers to individuals whose *first* incarceration spell occurred within the first two years after baseline, while recent incarceration refers to those first

[10]The measurement equations we write are either linear-separable in the latent mental health factor, or based on a linear index in M that is mapped into categories via thresholds. It is possible to allow for nonseparable relationships between the latent factor and measurement errors, $M_j = g_j(M, \varepsilon_j)$, along the lines of the nonseparable measurement-error framework studied by Hu (2008).

[11]For the discrete case, identification follows from similar arguments to the ones we describe for the continuous case in Section 2. See Carneiro et al. (2003) for example.

[12]We normalize the mental health factor to negative one, rather than one, because higher BSI scores indicate more severe mental health symptoms.

incarcerated thereafter (i.e., years 3 to 7 after baseline). Treatment operates through a latent index model,

$$T_i = X_i'\gamma_T + C_i\psi_T + M_i\mu_T + \varepsilon_{i,T}, \tag{37}$$

where the observed treatment status is determined by threshold crossings of the latent variable,

$$D_i = \begin{cases} 0 & \text{if } T_i \leq \kappa_1, \\ 1 & \text{if } \kappa_1 < T_i \leq \kappa_2, \\ 2 & \text{if } T_i > \kappa_2. \end{cases} \tag{38}$$

We normalize the first cutoff to zero ($\kappa_1 = 0$) and estimate the second cutoff κ_2. We assume $\varepsilon_{i,T}$ is independent of C_i, M_i, and the measurement errors, and that X_i is independent of all latent variables and uniquenesses.

We observe repeated mental health measurements in period 7, denoted by $\mathcal{M}_{i,j,7}$, for $j = 1, ..., N_M$. Let the corresponding latent variables be

$$\mathcal{M}^*_{i,j,7} = X_i'\gamma_{\mathcal{M},j,7} + M_{i,7}\mu_{j,7} + \varepsilon_{i,\mathcal{M},j,7}, \quad j \in \{1, ..., N_M\}, \tag{39}$$

and their observed counterparts be

$$\mathcal{M}_{i,j,7} = k \quad \text{if} \quad o_{\mathcal{M},j,7,k-1} < \mathcal{M}^*_{i,j,7} \leq o_{\mathcal{M},j,7,k}, \quad k = 1, ..., K_j, \tag{40}$$

with $o_{\mathcal{M},j,7,0} = -\infty$, $o_{\mathcal{M},j,7,1} = 0$, $o_{\mathcal{M},j,7,K_j} = \infty$, and $\mu_{4,7} = -1$. Specifically, the mental health factor at year seven is normalized to have a loading of negative one on the BSI hostility measure.

Let $D_{i1} = \mathbb{1}\{D_i = 1\}$ and $D_{i2} = \mathbb{1}\{D_i = 2\}$. The law of motion for mental health allows treatment effects to depend directly on baseline factors

$$\begin{aligned} M_{i,7} \;=\; & D_{i1}\delta_{T1} + D_{i2}\delta_{T2} + C_i\lambda_C + M_i\lambda_M + C_iD_{i1}\lambda_{CT1} + C_iD_{i2}\lambda_{CT2} \\ & + M_iD_{i1}\lambda_{MT1} + M_iD_{i2}\lambda_{MT2} + U_i. \end{aligned} \tag{41}$$

U_i is a mean zero random variable that is independent of C_i, M_i, and the measurement errors. As in the previous sections, we assume all uniquenesses are mutually independent and independent of the factors. This dynamic formulation allows treatment effects to be heterogeneous, depending

on both baseline cognitive skills and baseline mental health.

Finally, to address potential biases arising from nonrandom sample attrition, we incorporate a multinomial model for missing data seven years after baseline, with categories corresponding to fully observed outcomes, missing mental health measures only, and missing both mental health measures and treatment. The model depends on baseline cognitive skills C and mental health M, as well as on the vector of individual characteristics, X.

3.3 Estimation

Although the model is semi-parametrically identified, we adopt parametric assumptions to facilitate estimation. We postulate normality for all uniquenesses and use mixtures of normals to approximate the distributions of latent variables.

For continuous measurements, we assume $\varepsilon_{i,j} \sim N\left(0, \sigma^2_{\varepsilon_j}\right)$. For discrete outcomes, including ordered measures and treatment, we assume standard normal errors with unit variance.

The joint distribution of baseline cognitive skill and mental health is approximated by a four-component mixture of normals

$$f_{C,M}\left(z_C, z_M\right) = \sum_{h=1}^{4} \pi_{CM,h} \phi\left(z_C; m_{C,h}, \sigma^2_{C,h}\right) \phi\left(z_M; m_{M,h}, \sigma^2_{M,h}\right) \tag{42}$$

subject to $\sum_{h=1}^{4} \pi_{CM,h} m_{C,h} = 0; \sum_{h=1}^{4} \pi_{CM,h} m_{M,h} = 0$, where $\phi(z; m, \sigma^2)$ is the pdf of a normal random variable with mean m and variance σ^2 evaluated at z. The innovation U_i is modeled analogously using a two-component mixture

$$f_U\left(u\right) = \sum_{h=1}^{2} \pi_{U,h} \phi\left(u; m_{U,h}, \sigma^2_{U,h}\right) \tag{43}$$

subject to $\sum_{h=1}^{2} \pi_{U,h} m_{U,h} = 0$.[13]

We estimate the model using Monte Carlo Markov Chain (MCMC) methods with a Gibbs sampler. Each iteration of the sampler sequentially draws from the full conditional distribution of every parameter and latent variable, conditioning on the current values of the other parameters. We either use non-informative priors or diffuse proper priors when a prior is called for. We briefly describe the sampler next, and leave a more detailed summary of it for Appendix A.1.

[13] In Online Appendix OA.1.3, we show that results are robust to increasing the number of mixture components.

For each continuous cognitive measurement, estimation is straightforward. Conditional on C_i, the equation becomes linear in parameters with a normally distributed error. The posterior distributions of regression coefficients and loadings are Gaussian, and the inverse error variances follow standard conjugate Gamma distributions.

The ordered models are completed by sampling the latent utilities, e.g., $\mathcal{M}^*_{i,j}$, which are conditionally normally distributed given their observed categories. Thresholds are assigned diffuse uniform priors, and given these, latent utilities are sampled from truncated normal distributions. The treatment assignment equation is treated analogously.

In year 7, measurement equations for mental health are handled similarly, conditional on the constructed values of $M_{i,7}$ from the law of motion. The joint presence of observed measurements, latent traits, treatment, and interactions allows us to recover the parameters of the dynamic system, including the heterogeneous treatment effects and propagation mechanisms.

To sample the parameters governing the law of motion, $\delta_{T1}, \delta_{T2}, \lambda_C, \lambda_M, \lambda_{CT1}, \lambda_{CT2}, \lambda_{MT1}, \lambda_{MT2}$, we construct regressors from transformations of observed variables and latent states

$$\begin{aligned} \mathcal{M}^{**}_{i,j,7} &= D_{i1,j}\delta_{T1} + D_{i2,j}\delta_{T2} + C_{i,j}\lambda_C + M_{i,j}\lambda_M + CD_{i1,j}\lambda_{CT1} + CD_{i2,j}\lambda_{CT2} \\ &\quad + MD_{i1,j}\lambda_{MT1} + MD_{i2,j}\lambda_{MT2} + U_{i,j} + \varepsilon_{i,\mathcal{M},j,7} \end{aligned} \tag{44}$$

with $\mathcal{M}^{**}_{i,j,7} \equiv \mathcal{M}^*_{i,j,7} - X_i'\gamma_{\mathcal{M},j,7}$, and $D_{i1,j} \equiv D_{i1}\mu_{j,7}$, $CD_{i1,j} \equiv C_i D_{i1}\mu_{j,7}$, and so on. Each parameter appears across measurement equations indexed by j, and we estimate them by pooling across individuals and outcomes.

Posterior computations for the mixture distributions follow standard Bayesian mixture updating. For each latent variable, posterior sampling alternates between updating mixture component parameters, reassigning individuals to components, and drawing component weights from Dirichlet distributions. Sampling latent factors proceeds by conditioning on all observed and latent objects in the complete data likelihood. Posterior distributions are normal given the structure of the system.

This estimation strategy delivers full posterior distributions for all parameters and latent variables. With these recovered, the model permits recovery of the counterfactual distributions and treatment effect functionals introduced in the preceding section.

Our Gibbs sampling procedure offers substantial computational advantages over traditional es-

timation approaches. In our experience, the serial MCMC algorithm we employ achieves orders-of-magnitude speed improvements relative to optimization-based methods such as maximum likelihood, even when the latter are run with parallelized objective function evaluations. The iterative structure and tractability of the Gibbs sampler enable efficient exploration of the posterior distribution, making it particularly well-suited to the high-dimensional latent variable models considered here.

4 Results

In this section, we present estimation results obtained using the Gibbs sampler described above. To ensure thorough exploration of the posterior distribution, we run the algorithm for a total of 140,000 iterations, discarding the initial 40,000 draws as burn-in to mitigate the influence of starting values. To further reduce autocorrelation in the retained samples, we save every 20th draw from the remaining iterations for a total of 5,000 draws. This approach balances computational efficiency with the need for a representative sample from the posterior distribution of the model parameters.

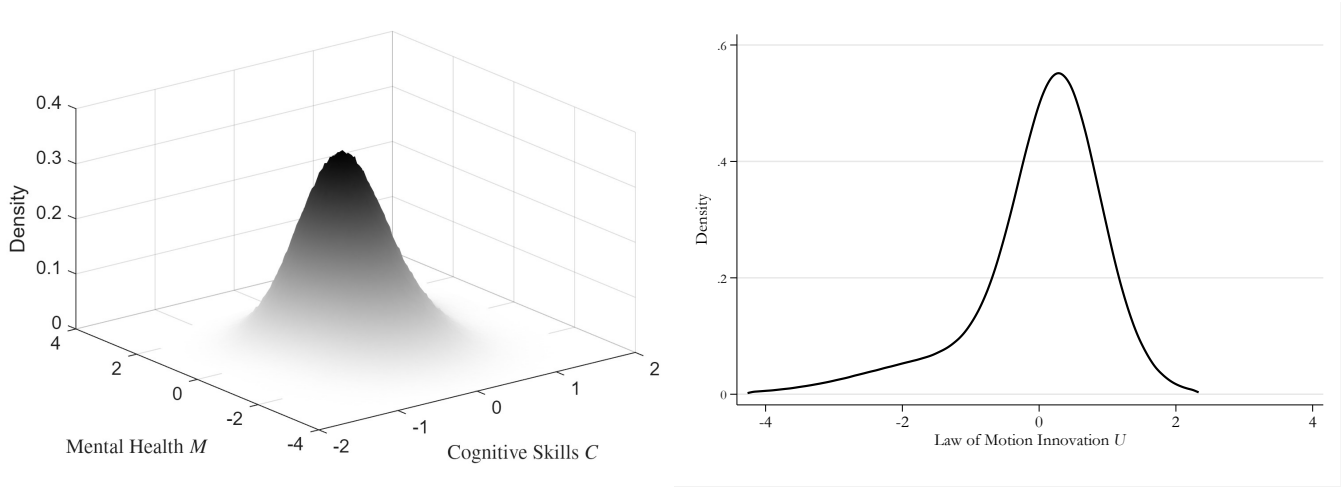
4.1 Summarizing Posterior Distributions and Variance Decomposition

We start by presenting the results from our factor analysis, in which we project our baseline and future measurements onto three factors: one related to cognitive skills, one related to baseline mental health, and one related to mental health in year 7. Figure 1a plots the average estimated joint posterior distributions of C and M. There is a small positive correlation between cognitive skills and mental health at baseline. Figure 1b shows the average estimated posterior distribution of the law of motion innovation to mental health in year 7 (U). As can be seen, the distributions are highly non-normal, highlighting the importance of allowing for our flexible mixture models.

The parameter estimates from the mixture model for the joint distribution of baseline latent skills (i.e., cognitive skills C and mental health M) and the subsequent innovation U to the mental health production function are presented in Table A.2.1. The means and standard deviations of the estimated posterior distributions of the factor measurement model are presented in Tables A.2.2, A.2.3, and A.2.4. The results in Table A.2.2 show that higher cognitive skills are associated with better scores across all cognitive measures we use. Likewise, Tables A.2.3 and A.2.4 show that

Figure 1: Estimated Posterior Distributions of the Factors

(a) Joint Distribution of C and M (b) Distribution of U



Notes: This figure reports the estimated posterior distributions of the factors. Figure 1a displays the estimated joint distribution of cognitive skills and mental health at baseline, while Figure 1b displays the estimated density of the law of motion innovation to mental health in year 7.

better mental health is associated with fewer BSI symptoms.[14]

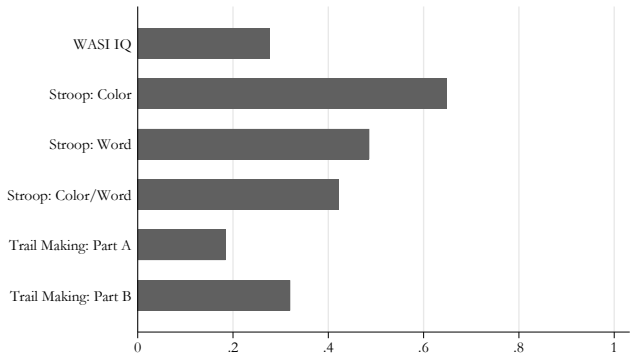
Figure 2 presents a variance decomposition to assess the extent to which measurement error influences our cognitive and mental health measures. Specifically, we break down the variance of the latent component of each measurement into two parts: the share explained by the underlying factor and the share attributable to measurement error. Figure 2a shows that the cognitive factor is more strongly related to the Stroop measures of cognitive dysfunction than to the WASI-IQ and Trail-Making measures. Nevertheless, even for the Stroop measures, cognitive skills explain on average (across posterior draws) only 51.9% of the unobserved variance. Figure 2b shows that the baseline mental health factor explains more than 48.4% of the variance for all BSI measures. Moreover, our mental health factor is more closely related to obsessive-compulsive behaviors, anxiety, and depression than to phobic anxiety and interpersonal sensitivity. A similar pattern is observed for the mental health factor at year 7 (see Figure 2c).

Table 4 summarizes the estimated posterior distributions for the treatment equation (37). As expected, higher cognitive skills are negatively associated with spending any time in jail, prison, or detention within the seven-year window after the baseline survey. Better mental health at baseline

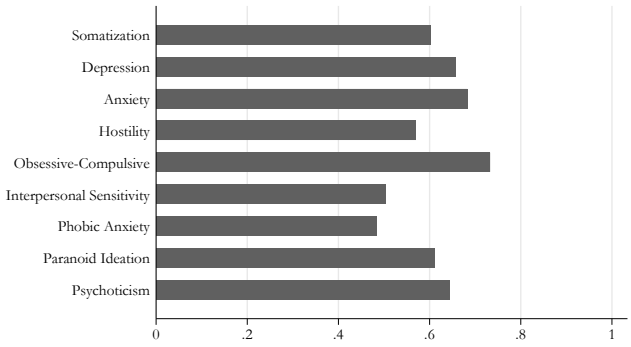
[14]The negative sign of the mental health factor in Tables A.2.3 and A.2.4 for all nine measures is consistent with the coding of BSI scores, where larger values reflect worse mental health.

Figure 2: Fraction of the Variance Explained by the Factor

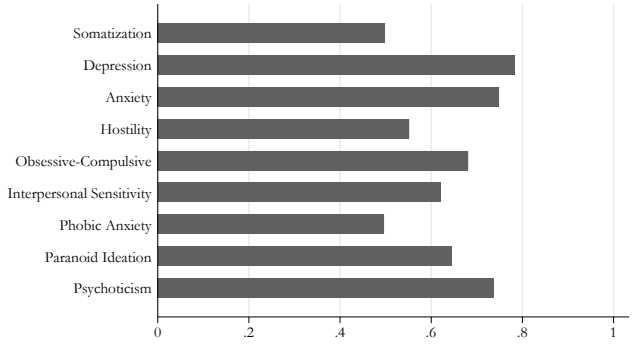
(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



(c) Mental Health (Year 7): M_7



Notes: These figures present the average fraction of the variance of each cognitive and mental health measure explained by the cognitive and mental health factors. For example, Figure 2a shows that 27.7% of the fraction of the variance of the residualized (against X) WASI IQ measure is explained by the cognitive skills C.

is also negatively associated with incarceration, although this association is imprecisely estimated. This selection is further illustrated in Figure 3, which displays the distributions of cognitive skills and baseline mental health conditional on treatment status. The negative selection is stronger for individuals first incarcerated between years three and seven after baseline (i.e., "recent").

Table 5 reports the estimates for the law of motion of mental health in equation (41). The results provide evidence of time dependence in mental health. On average, better baseline mental health is positively associated with better future mental health. The estimates also indicate that being incarcerated negatively affects mental health on average. The effect is more pronounced for those incarcerated more recently and larger for individuals with better baseline mental health, although the posterior distributions of these parameters have relatively large standard deviations.

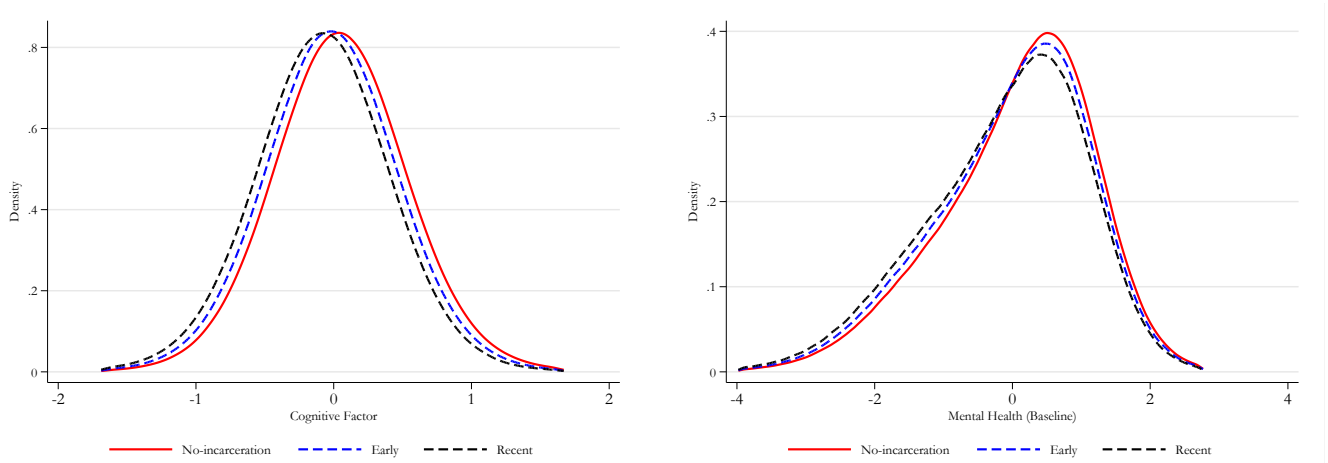
Table 4: Estimated Parameters from Model - Treatment Equation

	Mean	SD
Constant	-0.156	0.276
Age 15	0.326	0.178
Age 16	0.380	0.166
Age 17	0.166	0.171
Age 18	0.022	0.235
Female	-1.127	0.154
White	0.138	0.263
Hispanic	0.129	0.253
Black	0.429	0.257
Phoenix	0.055	0.126
Cognitive Skills (ψ_T)	-0.255	0.121
Mental Health (μ_T)	-0.060	0.050
Cutoff 1 (κ_1)	0.000	0.000
Cutoff 2 (κ_2)	1.199	0.070

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the treatment equation (37).

Figure 3: Distribution of Cognitive and Baseline Mental Health Factors By Treatment

(a) Cognitive Skills: C (b) Mental Health (Baseline): M



Notes: These figures show the estimated densities of the cognitive and baseline mental health factors conditional on incarceration status. For example, in Figure 3a, the red solid-line plots $f(C \mid D = 0)$, the blue dashed-line plots $f(C \mid D = 1)$, and the black dashed-line plots $f(C \mid D = 2)$.

Lastly, Table A.2.5 reports the estimates for the multinomial model for missing data seven years after baseline, with categories corresponding to fully observed outcomes, missing mental health only, and missing both mental health and treatment. The results indicate negative selection on baseline mental health: individuals with better mental health are less likely to be fully observed, relative to having all outcomes missing, and more likely to have missing mental health only after seven years.

Table 5: Estimated Parameters from Model - Law of Motion Equation

	Mean	SD
Early Incarceration (δ_{T1})	-0.189	0.122
Recent Incarceration (δ_{T2})	-0.448	0.162
Cognitive Skills (λ_C)	-0.235	0.167
Mental Health (λ_M)	0.379	0.083
Early Incarc. $\times$ Cognitive Skills (λ_{CT1})	0.221	0.281
Recent Incarc. $\times$ Cognitive Skills (λ_{CT2})	-0.522	0.345
Early Incarc. $\times$ Mental Health (λ_{MT1})	-0.189	0.108
Recent Incarc. $\times$ Mental Health (λ_{MT2})	-0.085	0.150

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the law of motion of mental health in equation (41).

4.2 Model Fit

In this section, we assess how our estimated model fits the observed data. Figures 4, 5, and 6 present graphical assessments of model fit for the cognitive measures, baseline mental health measures, and year 7 mental health measures, respectively. The model performs well overall. For the continuous cognitive measures (Figures 4a-4d), the predicted distributions closely match the observed ones in terms of dispersion and general shape. The observed distributions tend to exhibit slightly higher peaks around the mode compared to the predicted ones, suggesting a modest underestimation of concentration near the central tendency. For the discrete variables, both cognitive and mental health, the predicted frequencies align closely with the empirical data.

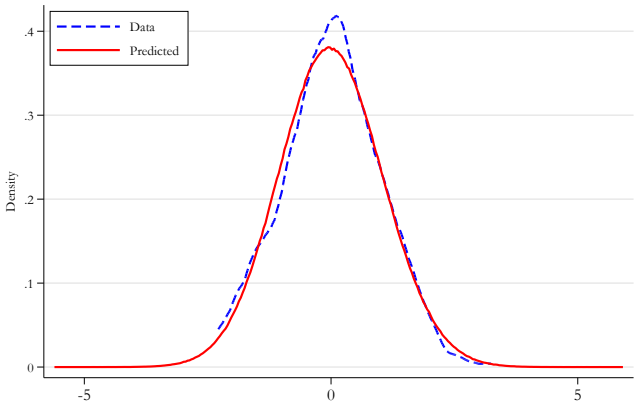
We formally measure the discrepancy between model and data by defining three test quantities. For ease of comparison, for our first test quantity, we use a classical frequentist interpretation of the Chi-Square test. Therefore, we use the Chi-Square test statistic, defined as the sum of differences between observed and predicted outcome frequencies, each squared and divided by the prediction

$$\chi^2_m = \sum_{b=1}^{B^m} \frac{(O_b^m - E_b^m)^2}{E_b^m}, \tag{45}$$

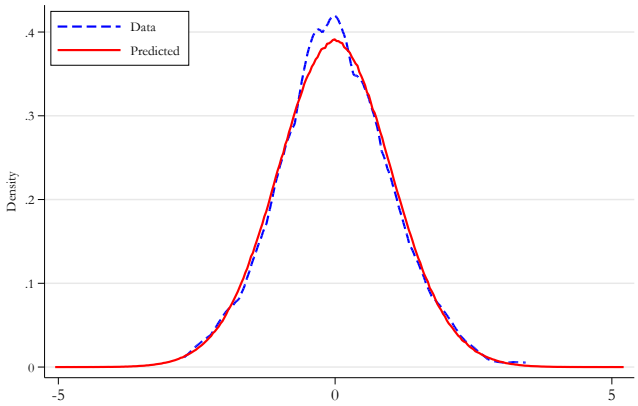
where outcome measures are indexed by m, bins are indexed by b with total number of bins indexed by B_m, observed count for bin b is indexed by O_b^m, and predicted count for bin b is indexed by E_b^m.

Figure 4: Posterior Predictive Fit - Cognitive Measures

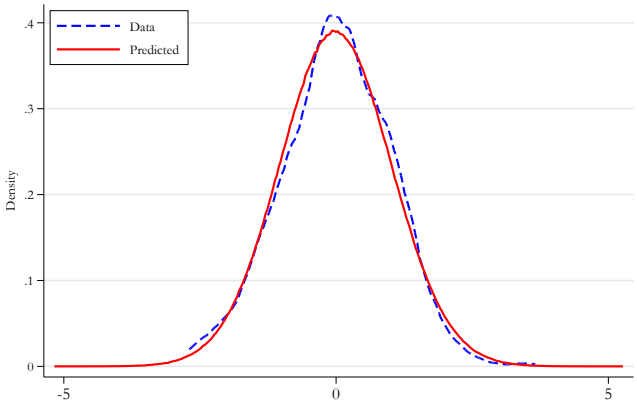
(a) WASI IQ



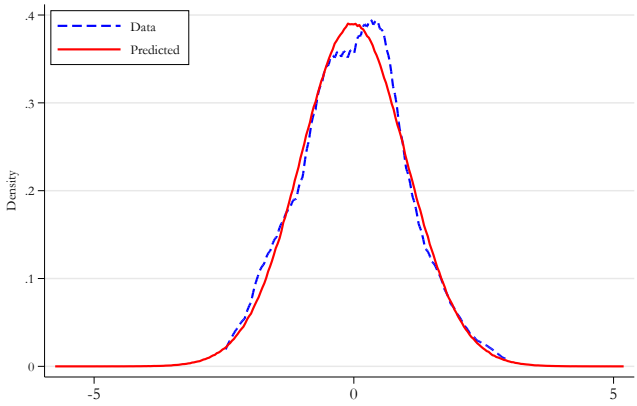
(b) Stroop Color



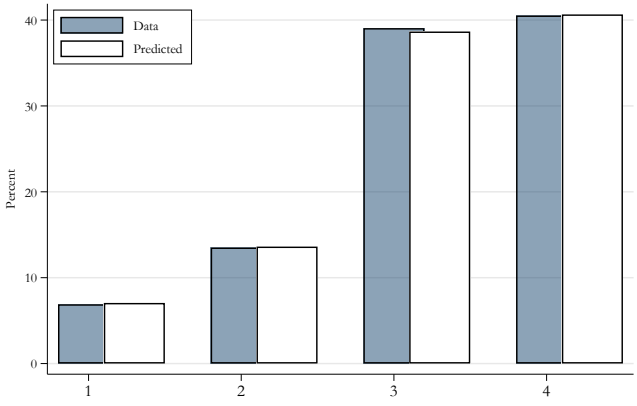
(c) Stroop Word



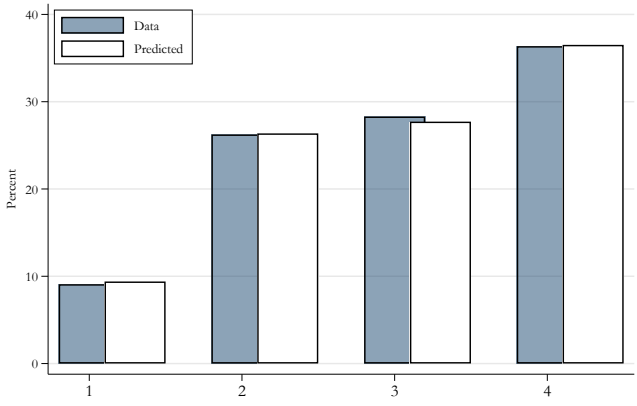
(d) Stroop Color/Word



(e) Trail Making A

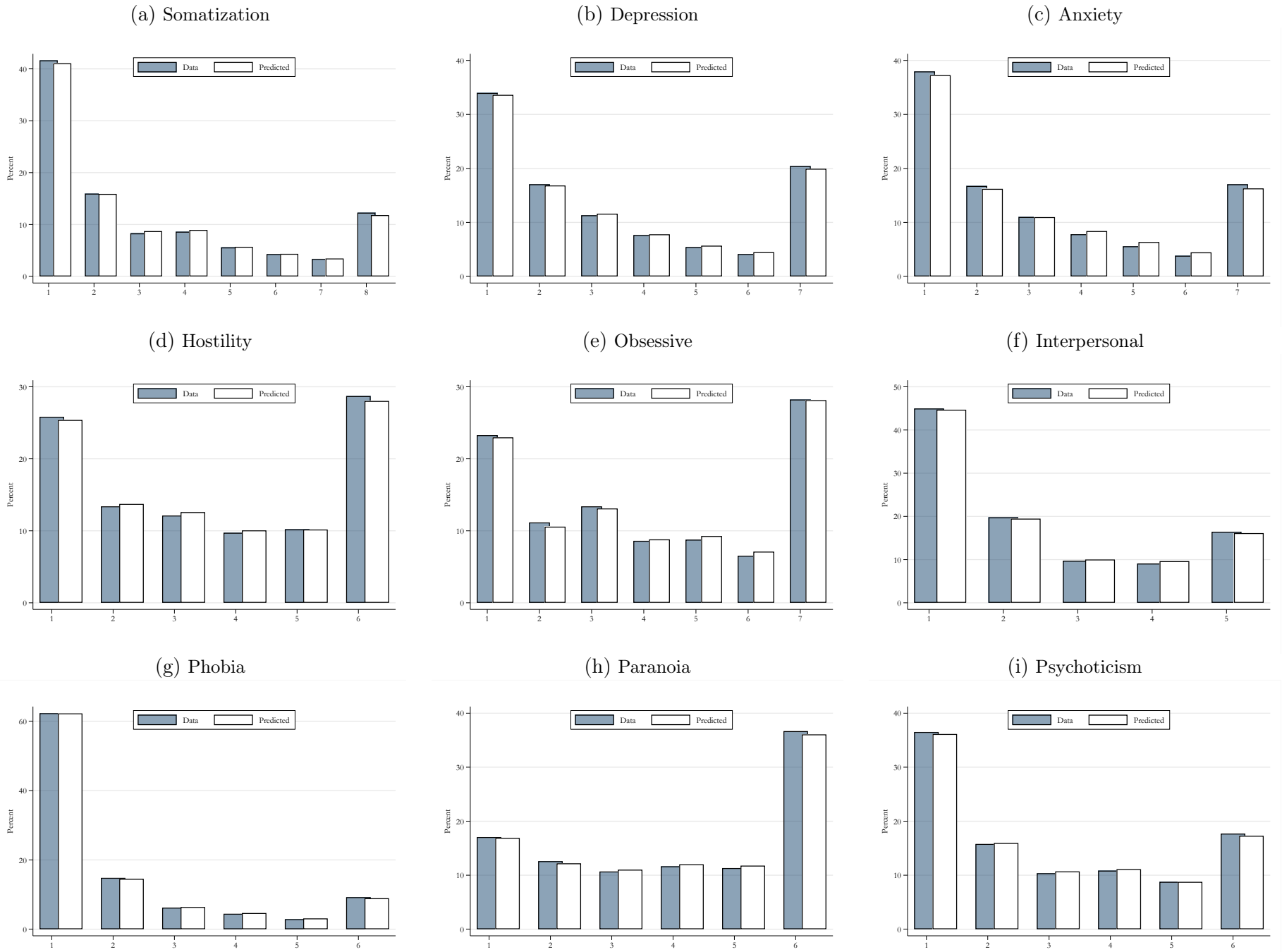


(f) Trail Making B



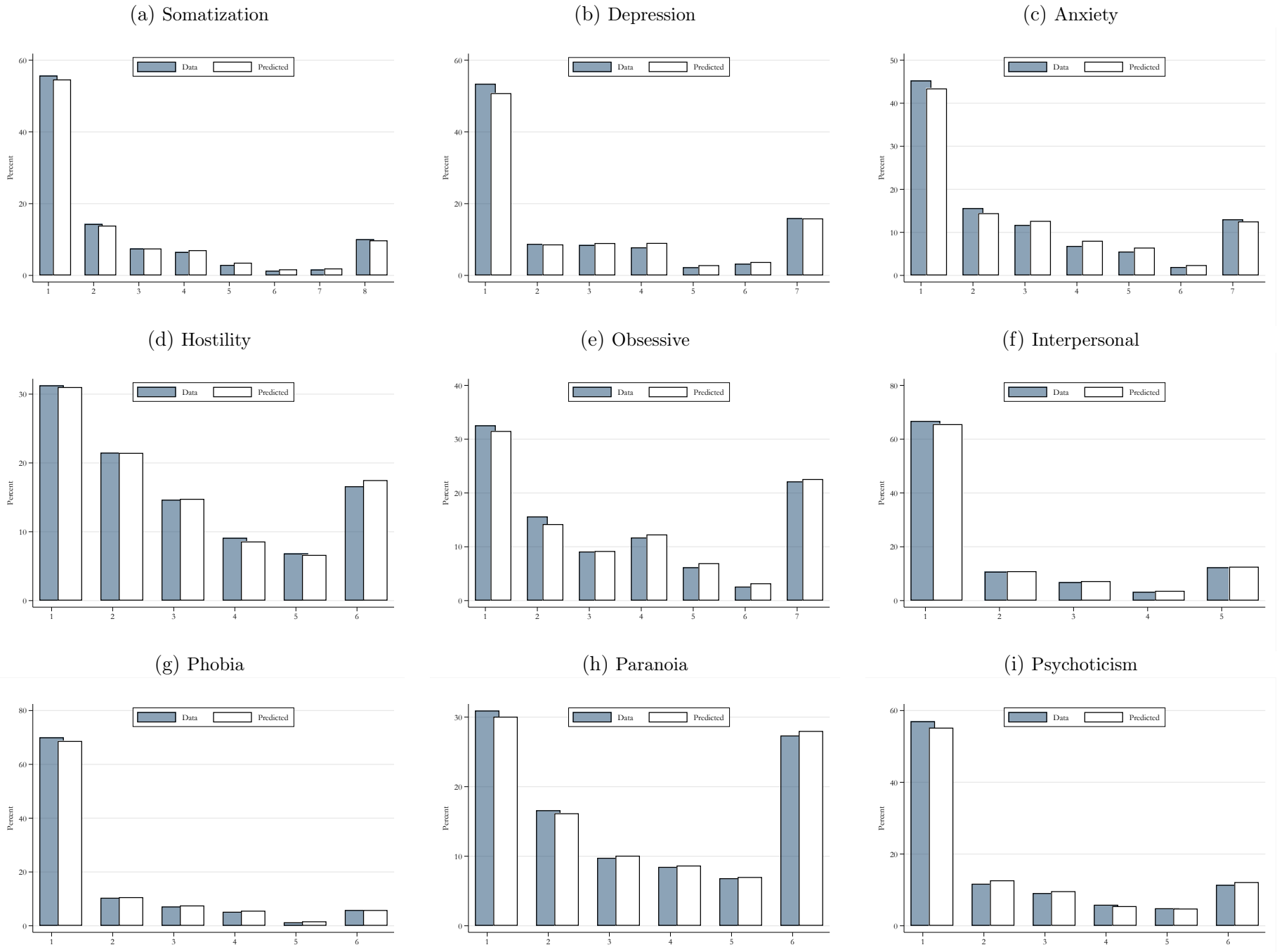
Notes: This figure compares observed and predicted distributions for each cognitive measure. For continuous variables (Figures 4a-4d), observed distributions appear as blue dashed lines and predicted as solid red lines. For discrete variables (Figures 4e and 4f), observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions.

Figure 5: Posterior Predictive Fit - BSI Mental Health Measures (Baseline)



Notes: This figure compares observed and predicted distributions for each baseline mental health measure. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions.

Figure 6: Posterior Predictive Fit - BSI Mental Health Measures (Year 7)



Notes: This figure compares observed and predicted distributions for each mental health measure seven years after the baseline. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions.

Column (1) in Table 6 reports the *Classical* p-value from the chi-square distribution for our test statistic for the cognitive and mental health measures displayed on the rows. Out of 24 measures, we reject the null (in a classical sense) of equality of frequencies in only 2 instances for a 5% significance level, and in both cases the p-values lie very close to the threshold.

Let $Y_{i,j}$ denote the observed value of measure j for individual i, and $\overline{Y}_j$ its mean. Define Θ_s to be the collection of parameters evaluated using the s-th parameter draw of the estimated posterior of the model. Then, $\widehat{Y}_{i,j,s}$ represents the predicted value of measure j evaluated at Θ_s. Finally, let $\overline{\widehat{Y}}_j = \mathbb{E}[\widehat{Y}_{i,j,s}]$ where the expectation is taken over both i and s. The Posterior Predictive P-value (PPP) is defined as

$$\frac{1}{S \times N} \sum_s \sum_i \mathbb{1}[\|\widehat{Y}_{i,j,s}, \overline{\widehat{Y}}_j\| > \|\overline{Y}_j, \overline{\widehat{Y}}_j\|], \tag{46}$$

the probability that the predicted data could be more extreme than the observed data, for a given distance measure $\|\cdot\|$. A model is suspect if its discrepancy has a tail-area probability near 0 or 1, indicating that the observed pattern would be unlikely to be seen in replications of the data if the model were true.

Columns (2) and (3) present two versions of the PPP as a *Bayesian* measure of fit for the measures on the rows. Column (2) uses the Mahalanobis distance, while column (3) uses a joint Euclidean distance of means and standard deviations. Only one measure (BSI Psychoticism in period 7) is close to 1, and in only a few instances are these values smaller than 0.05 (4 in Mahalanobis and 2 in Euclidean).

4.3 Robustness Checks

In this section, we present results from alternative specifications aimed at assessing the sensitivity of our estimates to key modeling choices. First, we estimate versions of the model in which treatment is defined using alternative specifications. Second, we re-estimate the model with an increased number of mixture components. Third, we consider a reduced number of BSI subscales for latent mental health, both at baseline and in year seven.

We begin by considering a specification in which treatment is defined as whether an individual goes to prison between the baseline period and the final observation, seven years later. This speci-

Table 6: Goodness Of Fit

	Chi-Square Classical p-value (1)	Mahalanobis Bayesian PPP (2)	Euclidean Bayesian PPP (3)
Cognitive Skills: C			
WASI IQ†	0.045	0.309	0.425
Stroop Color†	0.952	0.609	0.682
Stroop Word†	0.662	0.677	0.747
Stroop CW†	0.041	0.716	0.777
Trail Making A	0.997	0.754	0.679
Trail Making B	0.986	0.231	0.230
Mental Health: M			
BSI Somatization	1.000	0.486	0.288
BSI Depresion	0.999	0.426	0.473
BSI Anxiety	0.961	0.179	0.187
BSI Hostility	0.998	0.381	0.559
BSI Obsessive	0.996	0.773	0.725
BSI Interpersonal	0.993	0.849	0.643
BSI Phobia	0.999	0.853	0.860
BSI Paranoia	0.997	0.017	0.026
BSI Psychoticism	1.000	0.010	0.112
Mental Health: M_7			
BSI Somatization	0.998	0.020	0.036
BSI Depresion	0.977	0.807	0.885
BSI Anxiety	0.960	0.471	0.391
BSI Hostility	0.999	0.405	0.134
BSI Obsessive	0.988	0.260	0.219
BSI Interpersonal	0.993	0.023	0.144
BSI Phobia	0.997	0.224	0.100
BSI Paranoia	1.000	0.916	0.901
BSI Psychoticism	0.988	0.944	0.959

Notes: This table presents three measures of Goodness of Fit for our cognitive and mental health measures. Column (1) displays the *Classical* p-values from the chi-square distribution. Columns (2) and (3) report *Bayesian* Posterior Predictive P-values using the Mahalanobis distance and the joint Euclidean distance of means and standard deviations, respectively. For discrete variables, the total number of bins is the number of categories in the empirical distribution. For continuous variables (indexed by a dag), we set the number of bins equal to 4 and the cutoffs equal to the quartiles of the simulated distribution.

fication provides a benchmark for assessing the importance of the richer treatment structure in the baseline model. The results display broadly similar patterns, suggesting that our main findings are not driven by the specific treatment definition. See Online Appendix OA.1.1 for details.

We then consider an alternative treatment definition based on the timing of first incarceration over the same seven-year period. Specifically, the categories distinguish between no incarceration

($D_i = 0$) and first incarceration occurring in years 1 through 7 ($D_i = 1, ..., 7$). While the estimates are noisier, the overall patterns remain similar. See Online Appendix OA.1.2 for details.

We subsequently examine robustness to the number of mixture components. We re-estimate a model using 6 mixtures (instead of 4) to approximate the joint distribution of baseline cognitive skills C and mental health M, and 3 mixtures (instead of 2) to approximate the distribution of the innovation U. The resulting estimates are very similar, indicating that our findings are not sensitive to the specific approximation of the distributions. See Online Appendix OA.1.3 for details.

Finally, we assess the robustness of our results to the number of BSI subscales included in the factor model. To this end, we present results re-estimating the factor model using measures 2, 4, 6, and 8 of the BSI subscales (Depression, Hostility, Interpersonal Sensitivity, Paranoid Ideation), both at baseline and after seven years, instead of the full set of nine BSI subscales.[15] Online Appendix OA.1.4 displays the estimated posterior distributions of the factors under this alternative specification, showing that the distributions are largely robust to the choice of subscales used to construct the factors.

5 Counterfactuals

Overall, our estimated model fits the data really well. It demonstrates a strong ability to replicate the data patterns. These results support the model's suitability for counterfactual policy analysis.

5.1 Treatment Effects

We now use the potential outcomes interpretation of the law of motion of mental health introduced in Section 2.4.1 to present estimates of standard treatment effect parameters. Using (30), we define the return of incarceration on mental health in year 7 as:

$$\Delta_{M_{i,7,j',j}} = M_{i,7,j'} - M_{i,7,j}, \quad j \in \{0,1\}, \quad j' \in \{1,2\}, \quad j' > j.$$

[15] Alternative sets of measures give similar results.

The dynamic treatment effect parameters are defined as:[16]

$$
\begin{aligned}
ATE(M_7, j', j) &= \mathbb{E}[\Delta_{M_{i,7,j',j}}], \quad j \in \{0,1\}, \quad j' \in \{1,2\}, \quad j' > j \\
ATT(M_7, k, j', j) &= \mathbb{E}[\Delta_{M_{i,7,j',j}} \mid D_i = k], \quad k \in \{0,1,2\}, \quad j \in \{0,1\}, \quad j' \in \{1,2\}, \quad j' > j \\
MTE(M_7, k, j', j) &= \mathbb{E}[\Delta_{M_{i,7,j',j}} \mid T_i = \kappa_k], \quad k \in \{1,2\}, \quad j \in \{0,1\}, \quad j' \in \{1,2\}, \quad j' > j.
\end{aligned}
$$

Table 7 presents summary measures of the posterior distributions of some estimands of interest: the Average Treatment Effect, the Average Treatment on the Treated, and the Marginal Treatment Effect for $j = 0$ (i.e., relative to no incarceration). Column (1) presents the posterior mean of the treatment effect parameters. In addition to point summaries, in columns (2) to (6), we report measures of posterior uncertainty. Columns (2) and (3) present the lower and upper bounds of the 95% posterior credibility interval, respectively. Column (4) presents the posterior probability that the parameter is positive. Columns (5) and (6) report the posterior probability that the estimated parameters are close to zero. Specifically, Column (5) shows the posterior probability that the absolute value of the parameter is less than 0.01. Column (6) reports the probability for the analogous test using a threshold equal to 1% of the parameter's range across the relevant posterior distribution. In all cases, there is strong evidence of an average negative effect of incarceration on mental health, with the effect being larger (in absolute value) for individuals recently incarcerated (i.e., $|ATE(M_7, 1, 0)| < |ATE(M_7, 2, 0)|$), and those typically not incarcerated (i.e., $|ATT(M_7, k, 2, 0)| < |ATT(M_7, 0, 2, 0)|$, $k \in \{1, 2\}$).

These patterns can be seen more directly in the average treatment-on-the-treated effects reported in Table 7. The parameters $ATT(M_7, 0, 1, 0)$ and $ATT(M_7, 1, 1, 0)$ are virtually identical (-0.190 versus -0.189), indicating that the effect of early incarceration on mental health is very similar for individuals who are never incarcerated and for those who are incarcerated early. By contrast, the effect of recent incarceration is more negative for individuals who are typically not incarcerated than for those who are recently incarcerated, with $ATT(M_7, 0, 2, 0)$ equal to -0.481 and $ATT(M_7, 2, 2, 0)$ equal to -0.392. Taken together, these results suggest that higher baseline cognitive skills help individuals cope with incarceration when it occurs early in the period we study, partly offsetting the vulnerability associated with worse baseline mental health, but that such protective effects of

[16] See Heckman and Navarro (2007), Fruehwirth et al. (2016), and Heckman and Vytlacil (1999).

cognition are weaker when incarceration takes place later in the seven-year window.

Table 7: Mean Treatment Effects

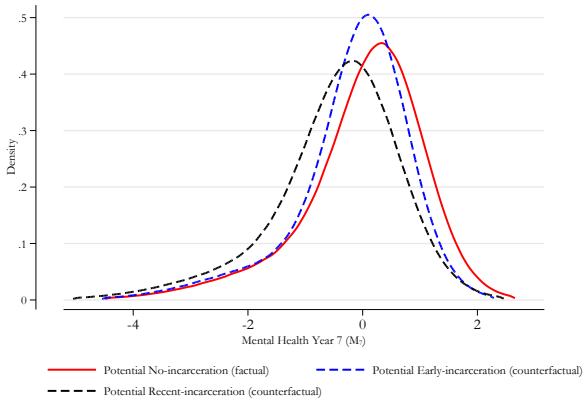
	Mean (1)	lb (2)	ub (3)	P($\cdot > 0$) (4)	P($\lvert\cdot\rvert < 0.01$) (5)	P($\lvert\cdot\rvert < 1\%$) (6)
ATE($M_7, 1, 0$)	-0.189	-0.435	0.044	0.059	0.018	0.018
ATE($M_7, 2, 0$)	-0.448	-0.773	-0.130	0.004	0.002	0.002
ATT($M_7, 1, 1, 0$)	-0.189	-0.438	0.041	0.059	0.019	0.019
ATT($M_7, 1, 2, 0$)	-0.437	-0.762	-0.121	0.005	0.002	0.002
ATT($M_7, 2, 1, 0$)	-0.187	-0.446	0.055	0.069	0.024	0.026
ATT($M_7, 2, 2, 0$)	-0.392	-0.721	-0.062	0.011	0.002	0.003
ATT($M_7, 0, 1, 0$)	-0.190	-0.439	0.052	0.060	0.021	0.020
ATT($M_7, 0, 2, 0$)	-0.481	-0.815	-0.153	0.002	0.000	0.001
MTE($M_7, 1, 1, 0$)	-0.192	-0.463	0.065	0.069	0.020	0.020
MTE($M_7, 1, 2, 0$)	-0.456	-0.786	-0.114	0.005	0.002	0.003
MTE($M_7, 2, 1, 0$)	-0.189	-0.473	0.081	0.085	0.023	0.024
MTE($M_7, 2, 2, 0$)	-0.418	-0.775	-0.071	0.010	0.003	0.004

Notes: Let Θ_s be the collection of parameters evaluated using the s-th parameter draw of the estimated posterior of the model. Column (1) presents $\frac{1}{S}\sum_{s=1}^{S} ATE(M_7, 1, 0; \Theta_s)$. Columns (2) and (3) present the 2.5 and 97.5 percentiles of $\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S}$. Columns (4), (5), and (6) are $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[ATE(M_7, 1, 0; \Theta_s) > 0]$, $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01]$, and $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01 \times Range(\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S})]$, respectively. Similar definitions apply to the remaining parameters.

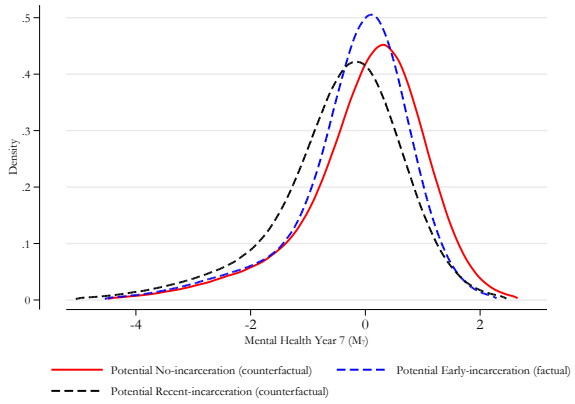
As discussed in Section 2.4.2, a key advantage of nonparametrically identifying the factor distributions is the ability to analyze features of the outcome distribution beyond mean treatment effects. This point is illustrated in Figure 7, which presents the counterfactual distributions of outcomes and gains for mental health in year 7. Figures 7a-7c show that the potential mental health distribution absent incarceration ($M_{i,7,0}$) first-order stochastically dominates the distribution under early and recent incarceration ($M_{i,7,1}$ and $M_{i,7,2}$) regardless of incarceration status (D). Nonetheless, as shown in Figure 7e, the estimated gains from recent incarceration are consistently larger for individuals who are actually recently incarcerated.

Figure 7: Factual and Counterfactual Distributions of Mental Health and Mental Health Gain (Year 7) conditional on Incarceration Status

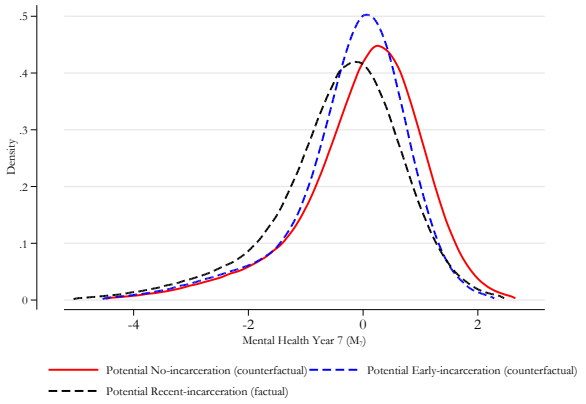
(a) M_7 Conditional on $D = 0$



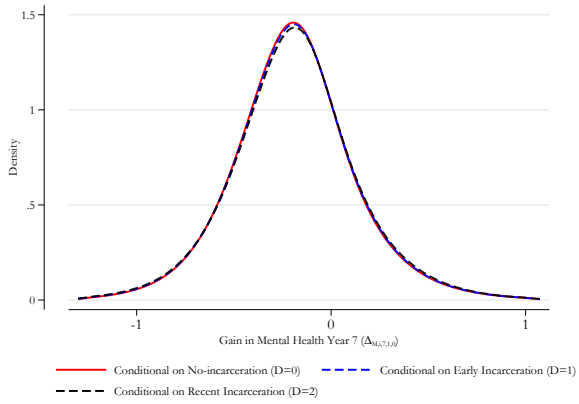
(b) M_7 Conditional on $D = 1$



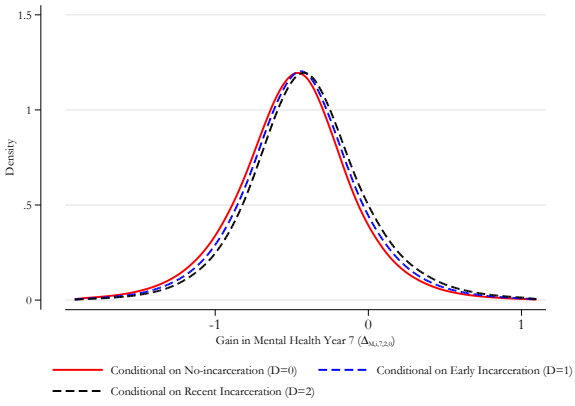
(c) M_7 Conditional on $D = 2$



(d) $\Delta_{M_i,7,1,0}$ by Incarceration Status (D)



(e) $\Delta_{M_i,7,2,0}$ by Incarceration Status (D)



Notes: Figure 7a displays the factual density $f(M_{i,7,0}|D_i = 0)$ and counterfactual densities $f(M_{i,7,1}|D_i = 0)$ and $f(M_{i,7,2}|D_i = 0)$, for people who do not go to prison. Similar definitions apply to Figures 7b and 7c. Figure 7d displays the distribution of mental health gain from early imprisonment, $\Delta_{M_i,7,1,0}$, conditional on no imprisonment (red solid line), early imprisonment (blue dashed line), and recent imprisonment (black dashed line). Similarly, Figure 7e displays the distribution of mental health gain from recent imprisonment, $\Delta_{M_i,7,2,0}$, by incarceration status.

These results align with the estimated mean treatment effects reported in Table 7. For most individuals, mental health deteriorates following incarceration, particularly when incarceration is more recent. At the same time, by estimating the full distribution of gains, we find that a nontrivial share of people would actually benefit from incarceration: 24.3% under universal early imprisonment and 10.9% under universal recent imprisonment. To give a sense of magnitude, the fraction of individuals with positive mental health gains ranges from at least 9.5% to at most 24.9%, depending on the counterfactual and subpopulation considered. Specifically, 9.5% of those who were never incarcerated would have experienced better mental health had they instead been recently incarcerated, while 24.9% of those who were recently incarcerated would have experienced better mental health had their incarceration occurred earlier.

Figure 8 further illustrates the sources of heterogeneity underlying the estimated treatment effects. The figures plot selected Marginal Treatment Effects on mental health in year 7 across deciles of baseline cognitive skills and baseline mental health. The deciles are computed using the full population distributions of baseline cognitive skills and mental health, while the estimated MTEs are evaluated for individuals located near the corresponding treatment threshold. For the marginal return to early incarceration for people at the margin of early incarceration, $MTE(M_7, 1, 1, 0)$, estimated returns tend to increase with baseline cognitive skills and decrease with baseline mental health, though the implied return surface is relatively flat. By contrast, the marginal return to recent incarceration for people on that margin, $MTE(M_7, 2, 2, 0)$, exhibits a much steeper gradient, with estimated returns decreasing along both dimensions.

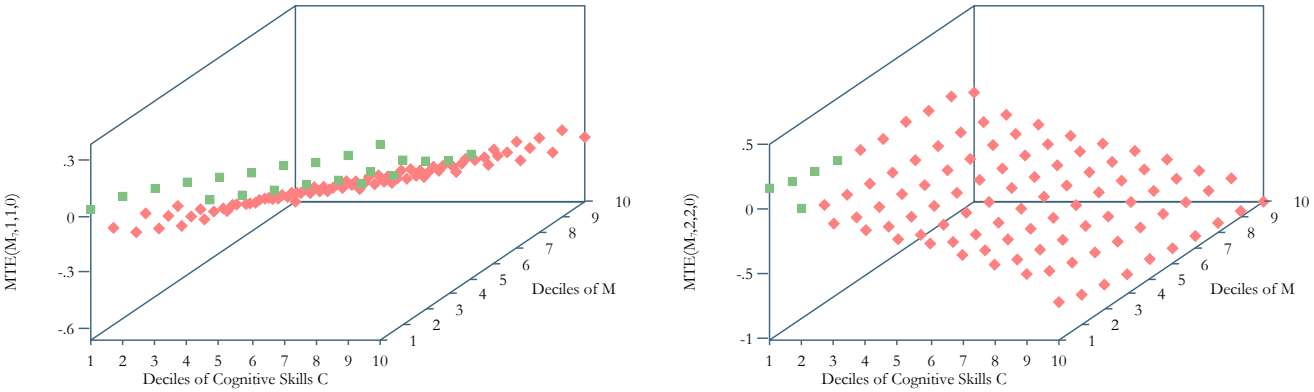
5.2 Mental Health: Mobility and Growth

Understanding the impact of incarceration on individual well-being requires not only examining average changes in mental health, but also investigating how it alters individuals' positions within the overall distribution. In this section, we present both detailed mobility matrices and summary measures that provide a comprehensive view of mobility patterns and their heterogeneity across the cognitive skill distribution.

We measure mobility by adapting the concept of *positional movement* from Fields (2006). This

Figure 8: Marginal Treatment Effects on Year-7 Mental Health Across Deciles of Baseline Cognitive Skills and Mental Health

(a) $MTE(M_7, 1, 1, 0)$ across deciles of C and M (b) $MTE(M_7, 2, 2, 0)$ across deciles of C and M



Notes: This figure plots selected estimated Marginal Treatment Effects of incarceration on mental health in year 7 across deciles of baseline cognitive skills (C) and baseline mental health (M). Panel 8a reports $MTE(M_7, 1, 1, 0)$, while Panel 8b reports $MTE(M_7, 2, 2, 0)$. Green markers denote positive estimated marginal treatment effects, indicating improvement in mental health under the counterfactual incarceration state, while red markers denote negative estimated marginal treatment effects, indicating deterioration in mental health.

approach measures changes in individuals' ranks across two counterfactual distributions, rather than changes in absolute levels. In our context, it captures how a person's position in the potential mental health distribution absent incarceration compares relative to their position in the potential distribution under incarceration. Positional mobility is particularly informative if individuals derive welfare from their relative standing in the mental health distribution rather than from absolute levels of mental health. Under this assumption, rank-based mobility provides a direct measure of changes in individual well-being.

Figure 9 presents a decile-by-decile mobility matrix comparing the joint distribution of two potential outcomes of mental health in year 7: no incarceration ($M_{7,0}$) and early incarceration ($M_{7,1}$).[17] Let $Q_{j,0}$ denote the jth quantile of the distribution of $M_{7,0}$, and let $Q_{k,1}$ denote the kth quantile of the distribution of $M_{7,1}$. The entry in cell (j, k) reports the probability

$$\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}]\right), \tag{47}$$

that is, each cell reports the probability that an individual who falls into a given decile of the mental health distribution under the no-incarceration scenario ($M_{7,0}$) would instead fall into a particular

[17] Appendix Figures A.3.1-A.3.3 replicate this mobility exercise on the potential mental health distributions conditional on incarceration status.

decile under the early incarceration scenario ($M_{7,1}$).[18] For example, we read from Figure 9a, row 1, column 1, that an individual with potential mental health absent incarceration drawn from the first decile of the distribution of $M_{7,0}$ has an 87.6% probability of drawing a potential mental health under early incarceration in the first decile of the distribution of $M_{7,1}$. In contrast, an individual in the sixth decile of the distribution of $M_{7,0}$ has only a 31.0% probability of being in the sixth decile of the distribution of $M_{7,1}$, indicating greater mobility in the middle of the distribution.

These patterns reflect a more general finding present in almost all panels of Figure 9: extreme deciles (such as the first and tenth) exhibit more *immobility* across counterfactual scenarios, while interior deciles are more fluid, independent of the cognition level. That is, individuals at the very top or bottom of the mental health distribution tend to retain their relative position regardless of incarceration and cognitive status, whereas individuals in the middle are more likely to shift ranks under the alternative scenario. This suggests that early incarceration has more heterogeneous effects in the center of the distribution and relatively less impact on those already at the tails. Within the extreme deciles of the mental health distribution, the lowest deciles exhibit more *immobility* than the highest deciles, particularly among individuals with lower cognitive skills. For example, going back to Figure 9a, individuals in the first decile of the distribution of $M_{7,0}$ have only a 12.4% probability of leaving their decile if they go to prison early, while this probability is more than doubled (26.1%) for individuals in the tenth decile of the distribution of $M_{7,0}$.

Furthermore, beginning from the lowest deciles of the counterfactual mental health distribution absent incarceration, upward mobility at the intensive margin is notably constrained for those who experience early incarceration. For example, among individuals initially in the first decile under the no-incarceration scenario, almost none transition above the fourth decile when subjected to early incarceration. This pattern holds across all strata of cognitive skills, and particularly among those with low cognitive skills. Even those with relatively high cognitive skills who occupy the bottom of the mental health distribution absent incarceration, exhibit limited mobility under early incarceration (see Figure 9d). In contrast, individuals in the tenth decile under no incarceration can experience downward mobility to as low as the fourth decile upon early incarceration (see Figure

[18] Figures 9b-9d repeat the same exercise for individuals in a given tertile of the cognitive skills distribution, for Q calculated using all individuals. Formally, $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], C \in \text{Tertile } t\right)$.

9b). Overall, relative rank persistence is more pronounced at the lower end than at the upper end of the distribution.

Nonetheless, lower cognitive skills increase mental health mobility at the top. For example, comparing the tenth row of Figure 9b (low cognition) to the tenth row of Figure 9d (high cognition), the share of individuals who remain in the top decile of the mental health distribution under early incarceration increases from 65.7% to 82.7% when we move from the first to the third tertile of the cognitive skills distribution. Focusing on the marginal distributions across Figures 9b-9d reveals that individuals with relatively high cognitive skills tend to have lower positions in the mental health distribution relative to individuals with low and medium cognition levels, especially in the no-incarceration potential mental health distribution.

Figure 10 presents the analogous mobility matrices comparing the joint distribution of potential outcomes of mental health in year 7 under no incarceration ($M_{7,0}$) and recent incarceration ($M_{7,2}$).[19] The patterns are broadly similar to those in Figure 9: effects of recent incarceration are more heterogeneous in the center of the distribution and more limited at the tail. Within the extremes, the lowest deciles exhibit greater immobility than the highest deciles, and this holds across all cognitive skill levels. Starting from the lowest deciles of the counterfactual distribution, upward mobility at the intensive margin remains constrained under recent incarceration, again across all cognitive skill groups.

A key difference relative to Figure 9 is that higher cognitive skills are associated with greater mobility at the top of the distribution. For example, comparing the tenth row of Figure 10b (low cognition) and Figure 10d (high cognition), the share of individuals remaining in the top decile after recent incarceration declines from 87.5% to 38.9% when moving from the lowest to the highest cognitive skill tertile.

The mobility matrices in Figures 9 and 10 are useful for visualizing mobility patterns across mental health deciles within each group. However, to compare overall levels of mobility across different groups (i.e., across the various mobility matrices in Figures 9 and 10), we require a summary

[19] Figures 10b-10d repeat the same exercise for individuals in a given tertile of the cognitive skills distribution, for Q calculated using all individuals. Formally, $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], C \in \text{Tertile } t\right)$. Appendix Figures A.3.4-A.3.6 replicate the mobility exercises on the potential mental health distributions conditional on incarceration status.

statistic. Table 8 presents three complementary measures of positional movement, each condensing the full mobility matrix into a single index. Column (1) reports the Chi-Square Mobility Index, which measures how much the observed mobility matrix deviates from random mobility. Column (2) presents the Mobility Ratio Index, which measures the share of individuals who do *not* remain on the same decile across any two distributions. Column (3) shows the Mean Deciles Move Index, which captures the average number of deciles each individual moves between any two distributions. In all cases, higher values of the index correspond to greater mobility.[20]

Each row of Table 8 corresponds to a different group defined by incarceration status (no, early, recent) and baseline cognitive skill level (low, medium, high). A consistent pattern emerges across all three metrics: mental health mobility follows a U-shaped relationship with cognitive skills—higher among individuals with either low or high cognitive skills, and lower among those in the middle of the cognitive distribution. This suggests that both ends of the cognitive spectrum may be more sensitive to the effects of early or recent incarceration on mental health rankings, while middle-skill individuals are relatively more anchored in their initial position.

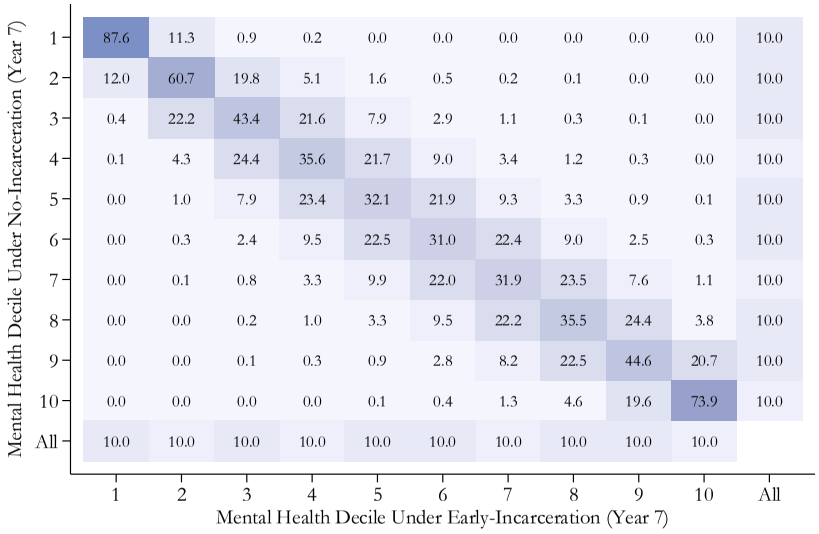
This U-shaped profile is an empirical regularity in our estimates rather than a mechanical implication of the model, so any interpretation we provide is necessarily speculative. A plausible explanation is that youths with median cognitive skills face relatively similar constraints and opportunities across counterfactual incarceration scenarios, so incarceration tends to shift their mental health within the broad middle of the distribution rather than pushing them toward the tails. In contrast, youths with either relatively low or relatively high cognitive skills may experience more heterogeneous responses to incarceration: for some, incarceration substantially worsens mental health, while for others it has more limited effects. This greater dispersion in responses at the lower and upper ends of the cognitive distribution naturally translates into higher positional mobility in mental health for those groups.

While positional mobility captures changes in relative standing in the mental health distribution, it does not reveal whether individuals experience improvements or declines in absolute mental health. A downward shift in decile may still coincide with a mental health improvement if the overall distribution shifts upward. To assess mental health *growth*, we fix the reference to the observed

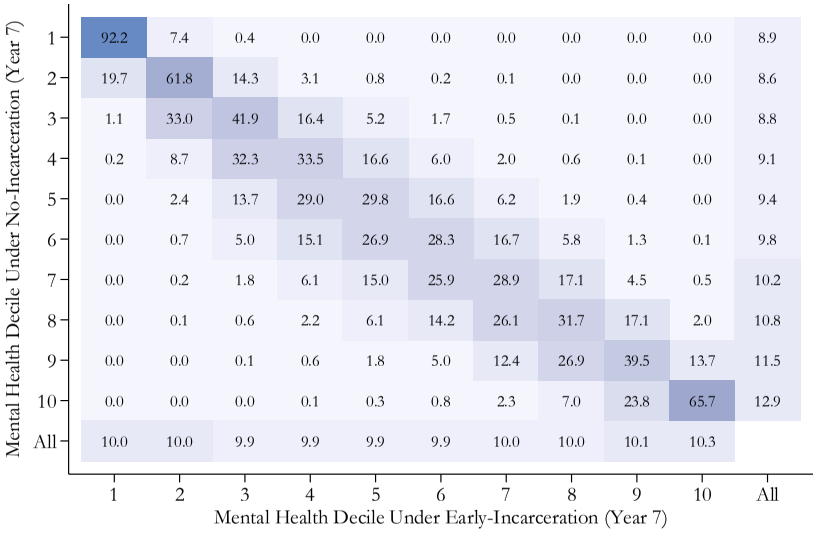
[20] Detailed formulas for each measure are provided in Appendix A.3.1.

Figure 9: Positional Mobility Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$

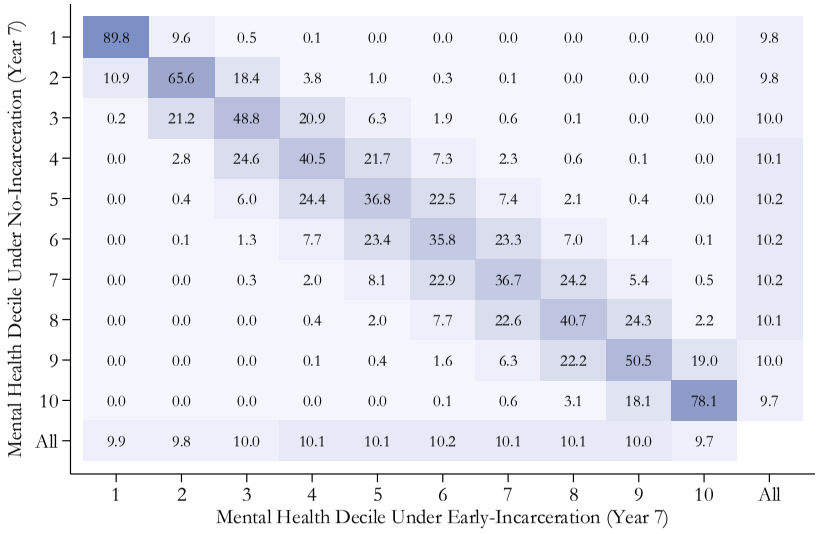
(a) Unconditional



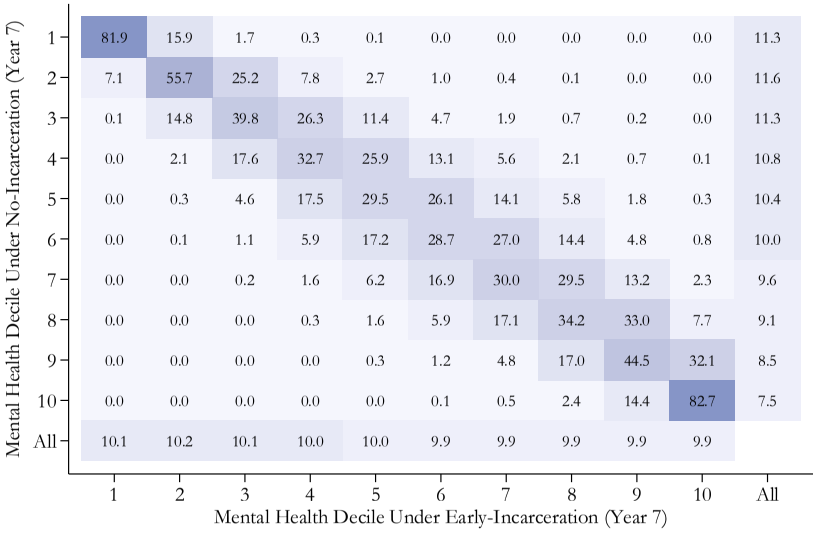
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



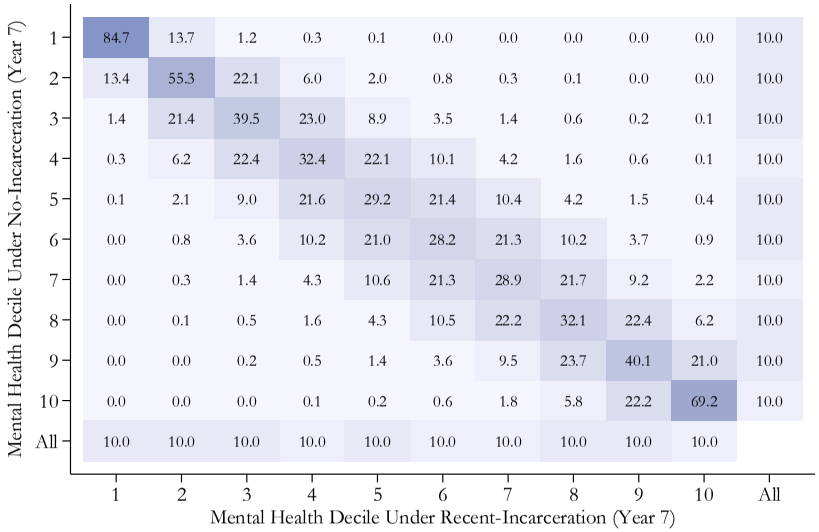
(d) For C (Cognitive) in Tertile 3



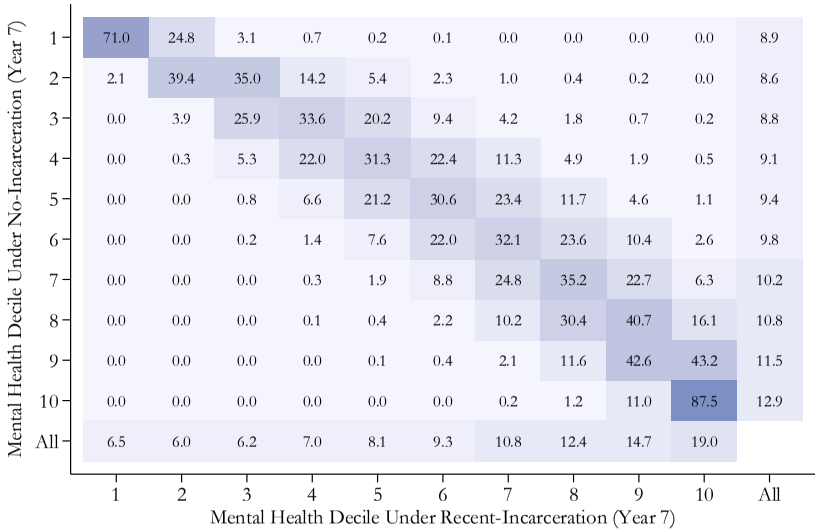
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,1}$ be the kth quantile of $M_{7,1}$. Cell (j,k) in Figure 9a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}]\right)$. Figures 9b-9d report $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure 10: Positional Mobility Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$

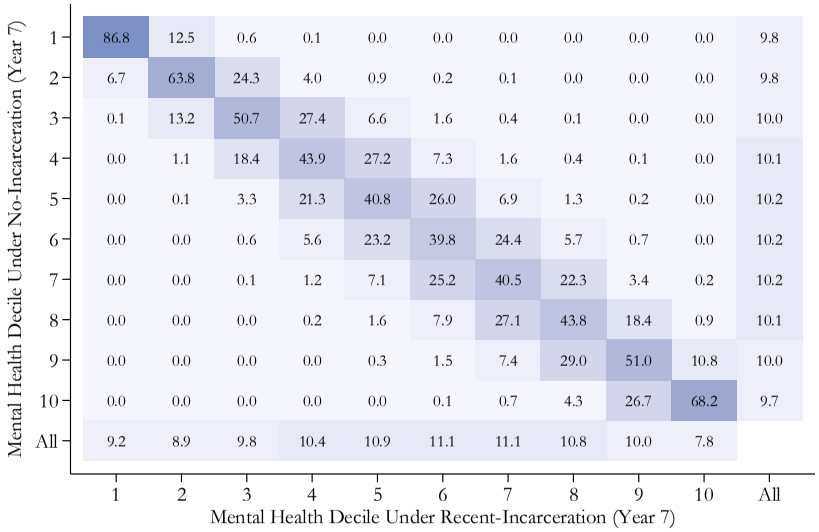
(a) Unconditional



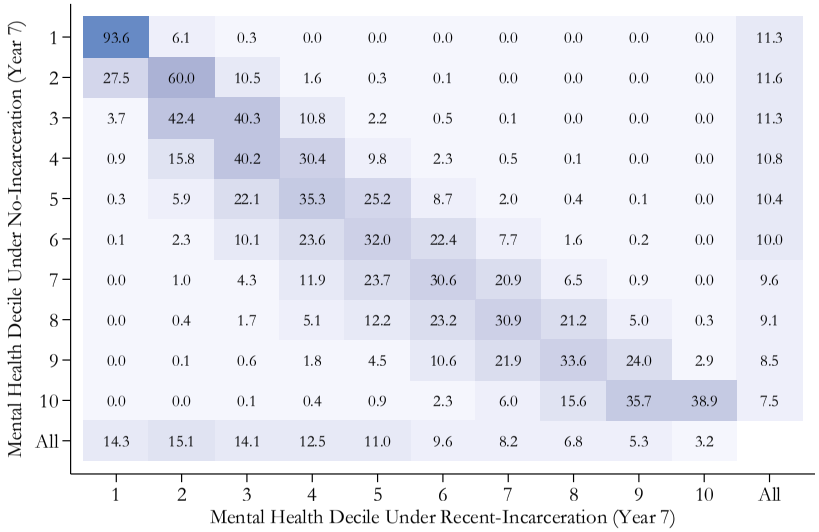
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



(d) For C (Cognitive) in Tertile 3



Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,2}$ be the kth quantile of $M_{7,2}$. Cell (j,k) in Figure 10a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}]\right)$. Figures 10b-10d report $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Table 8: Mobility Measures - Mental Health Year 7 (M_7)

	Chi-Square Mobility Index (1)	Mobility Ratio Index (2)	Mean Deciles Move Index (3)
Unconditional on Imprisonment:			
No Prison to Early-Prison	-2533.372	0.524	0.720
No Prison to Early-Prison | Low Cognitive	-2492.173	0.548	0.782
No Prison to Early-Prison | Med Cognitive	-2932.218	0.480	0.609
No Prison to Early-Prison | High Cognitive	-2527.854	0.544	0.769
No Prison to Recent-Prison	-2210.736	0.560	0.822
No Prison to Recent-Prison | Low Cognitive	-2458.539	0.599	0.938
No Prison to Recent-Prison | Med Cognitive	-2955.924	0.473	0.577
No Prison to Recent-Prison | High Cognitive	-2415.235	0.609	0.952
Conditional on D=0:			
No Prison to Early-Prison	-2551.872	0.523	0.717
No Prison to Early-Prison | Low Cognitive	-2532.457	0.547	0.778
No Prison to Early-Prison | Med Cognitive	-2951.538	0.479	0.606
No Prison to Early-Prison | High Cognitive	-2526.619	0.544	0.767
No Prison to Recent-Prison	-2216.266	0.562	0.825
No Prison to Recent-Prison | Low Cognitive	-2487.174	0.592	0.910
No Prison to Recent-Prison | Med Cognitive	-2964.565	0.473	0.576
No Prison to Recent-Prison | High Cognitive	-2405.000	0.617	0.979
Conditional on D=1:			
No Prison to Early-Prison	-2530.413	0.524	0.720
No Prison to Early-Prison | Low Cognitive	-2487.262	0.548	0.783
No Prison to Early-Prison | Med Cognitive	-2928.204	0.480	0.609
No Prison to Early-Prison | High Cognitive	-2533.311	0.543	0.769
No Prison to Recent-Prison	-2221.221	0.559	0.817
No Prison to Recent-Prison | Low Cognitive	-2454.973	0.600	0.941
No Prison to Recent-Prison | Med Cognitive	-2955.742	0.473	0.576
No Prison to Recent-Prison | High Cognitive	-2425.610	0.605	0.937
Conditional on D=2:			
No Prison to Early-Prison	-2495.288	0.526	0.728
No Prison to Early-Prison | Low Cognitive	-2433.322	0.550	0.790
No Prison to Early-Prison | Med Cognitive	-2904.005	0.481	0.614
No Prison to Early-Prison | High Cognitive	-2526.334	0.546	0.776
No Prison to Recent-Prison	-2210.733	0.561	0.827
No Prison to Recent-Prison | Low Cognitive	-2422.195	0.610	0.982
No Prison to Recent-Prison | Med Cognitive	-2942.052	0.473	0.578
No Prison to Recent-Prison | High Cognitive	-2435.827	0.595	0.904

Notes: This table reports three alternative measures of *positional* movement: (1) the Chi-Square Mobility Index, which measures how much the observed transition matrix deviates from random mobility; (2) the Mobility Ratio Index, which measures the share of individuals who do *not* remain on the same decile across the two distributions; and (3) the Mean Deciles Move Index, which captures the average number of deciles each individual moves between the two distributions. Detailed formulas for each measure are provided in Appendix A.3.1. In all three cases, higher values indicate greater mobility. Each row corresponds to a different subsample, defined by incarceration status and cognitive skill level.

distribution of mental health in year 7 (M_7). We then compare each individual's location in this distribution under the early, recent, and no-incarceration scenarios, i.e., we summarize ($f_{M_7}(M_{i,7,0})$, $f_{M_7}(M_{i,7,1})$) and ($f_{M_7}(M_{i,7,0})$, $f_{M_7}(M_{i,7,2})$).

Figure 11 presents a decile-by-decile mobility matrix summarizing this growth in individuals' potential mental health outcomes in year 7 under the joint distribution of no incarceration ($M_{7,0}$) and early incarceration ($M_{7,1}$).[21] Let Q_j denote the jth quantile of the distribution of M_7. The entry in cell (j, k) reports the probability

$$\Pr\left(M_{i,7,1} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j]\right), \tag{48}$$

that is, each cell reports the probability that an individual under the no-incarceration scenario ($M_{7,0}$) who falls into a given decile of the observed mental health distribution, would instead fall into a particular decile of the same distribution under the early incarceration scenario ($M_{7,1}$).[22] Figure 12 is analogous to Figure 11 and presents a decile-by-decile mobility matrix summarizing this growth in individuals' potential mental health outcomes in year 7 under the joint distribution of no incarceration ($M_{7,0}$) and recent incarceration ($M_{7,2}$).[23]

Unlike the positional mobility analysis, each matrix uses a single distribution to assign deciles and evaluates whether individuals move up or down in that common distribution based on their counterfactual outcome. The accompanying Table 9 quantifies growth using four complementary indices: (1) The share of individuals who move at least one decile downward; (2) The average number of deciles lost, assigning zero to non-losers; (3) The average mental health loss among those who move at least one decile downward ("losers"); and (4) The average mental health gain among individuals who move at least one decile upward ("winners").

Compared to the positional mobility matrices, Figures 11 and 12 show an even stronger pattern of persistence at the bottom of the distribution: individuals starting in the first decile almost always remain in the first decile. This persistence is particularly pronounced under recent incarceration. For

[21] Appendix Figures A.3.7-A.3.9 replicate this growth exercise on the potential mental health distributions conditional on incarceration status.

[22] Figures 11b-11d repeat the same exercise for individuals in a given tertile of the cognitive skills distribution, for Q calculated using all individuals. Formally, $\Pr\left(M_{i,7,1} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], C \in \text{Tertile } t\right)$.

[23] Appendix Figures A.3.10-A.3.12 replicate this growth exercise on the potential mental health distributions conditional on incarceration status. Figures 12b-12d repeat the same exercise for individuals in a given tertile of the cognitive skills distribution, for Q calculated using all individuals. Formally, $\Pr\left(M_{i,7,2} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], C \in \text{Tertile } t\right)$.

example, we read from Figure 12a, row 1, column 1, that an individual with potential mental health absent incarceration drawn from the first decile of the distribution of M_7 has a 97.0% probability of drawing a potential mental health under recent incarceration in the first decile of the distribution of M_7. This pattern holds across all levels of cognition, both for recent and early incarceration (see Figures 11b-11d and Figures 12b-12d).

Moreover, individuals who do not typically go to prison (i.e., conditional on $D = 0$) experience substantially larger declines in mental health if recently incarcerated than those who typically go to prison recently (i.e., conditional on $D = 2$), indicating that recent incarceration harms the former more severely (see Figures A.3.10 and A.3.12). These patterns are confirmed by our indices in Table 9. For instance, individuals who typically do not go to prison have a negative growth ratio for recent incarceration of 0.707 and an average downward move of 1.437 deciles, compared with 0.652 and 1.238, respectively, for individuals who typically go to prison recently. By contrast, for early incarceration, the differences between individuals who typically do not go to prison and those who go to prison early (i.e., conditional on $D = 1$) are considerably less pronounced (i.e., negative growth ratios for early incarceration of 0.486 and 0.485, and average downward moves of 0.726 and 0.727, respectively).

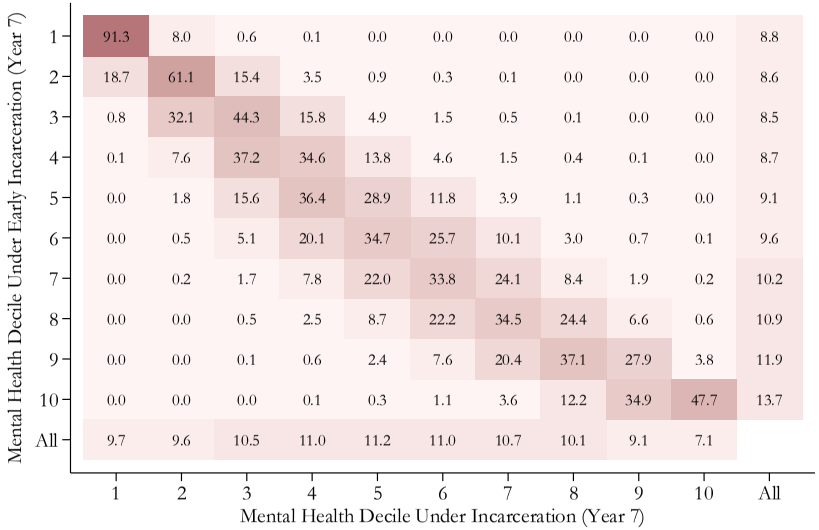
6 Conclusion: Towards a Richer Understanding of Policy Effects

This paper develops and estimates a dynamic factor model to analyze the distributional consequences of incarceration on mental health among justice-involved youth. By modeling the joint evolution of latent cognitive and mental health factors, we move beyond average effects to characterize heterogeneity and social mobility patterns in greater detail.

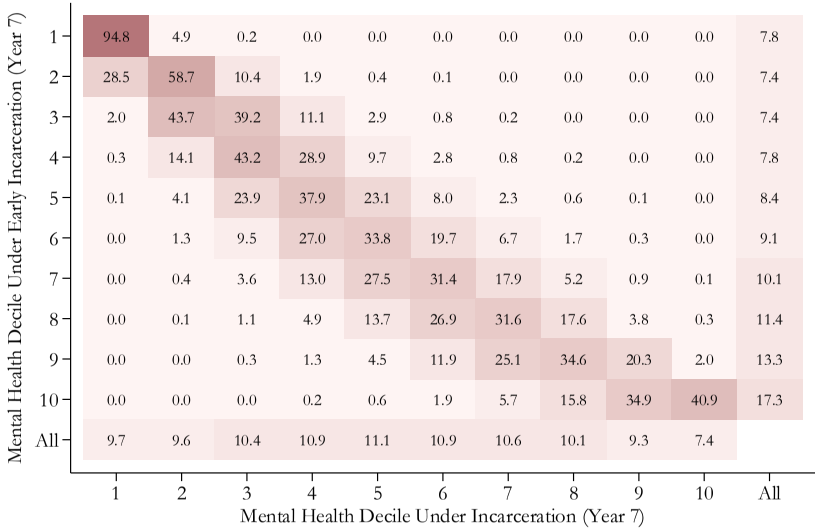
A central contribution is the nonparametric recovery of the joint distribution of unobserved skill factors without imposing restrictive parametric assumptions on their functional form. This feature enables a flexible and rich description of individual variation and policy impact heterogeneity. Nevertheless, this nonparametric identification and estimation rely on a set of carefully stated as-

Figure 11: Growth Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$

(a) Unconditional



(b) For C (Cognitive) in Tertile 1



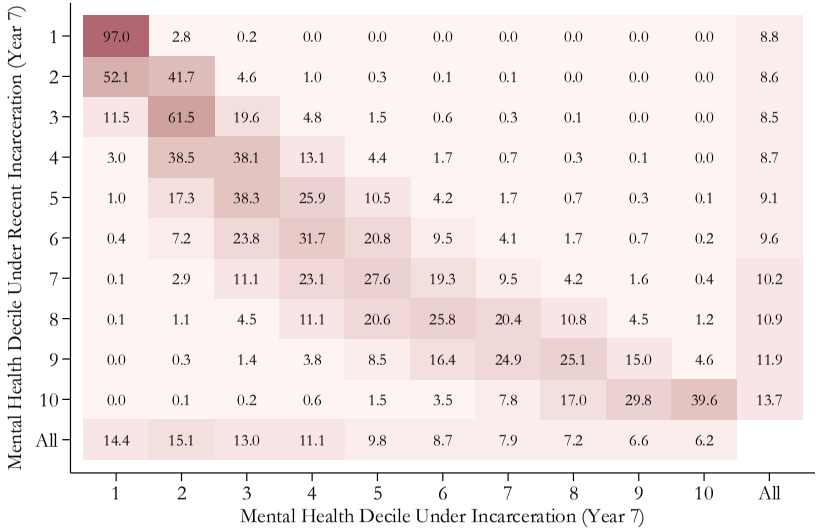
(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

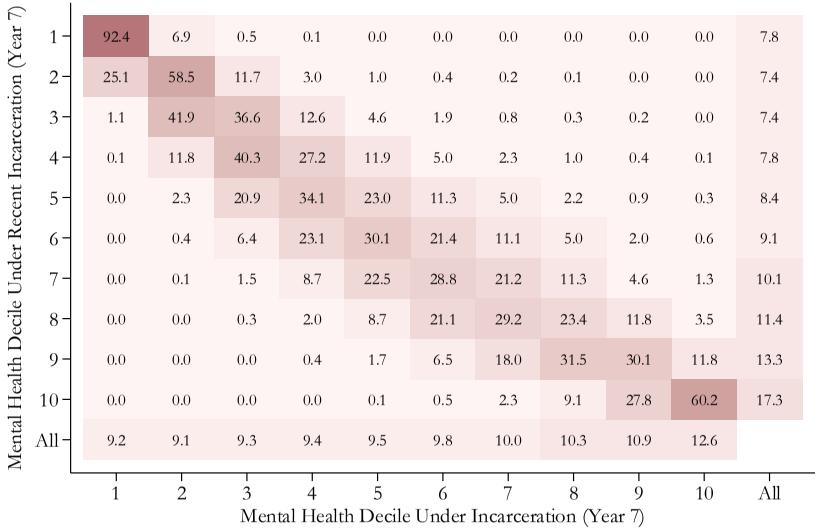
Notes: Let Q_j be the jth quantile of M_7. Cell (j,k) in Figure 11a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j]\right)$. Figures 11b-11d report $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure 12: Growth Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$

(a) Unconditional



(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j,k) in Figure 12a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j]\right)$. Figures 12b-12d report $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Table 9: Growth Measures - Mental Health Year 7 (M_7)

	Negative Growth Ratio (1)	Deciles Downward Move (2)	Mental Health Loss (Losers) (3)	Mental Health Gain (Winners) (4)
Unconditional on Imprisonment:				
No Prison to Early-Prison	0.485	0.726	0.393	0.328
No Prison to Early-Prison | Low Cognitive	0.592	0.984	0.458	0.295
No Prison to Early-Prison | Med Cognitive	0.495	0.690	0.353	0.293
No Prison to Early-Prison | High Cognitive	0.367	0.505	0.341	0.355
No Prison to Recent-Prison	0.687	1.362	0.591	0.365
No Prison to Recent-Prison | Low Cognitive	0.472	0.713	0.404	0.379
No Prison to Recent-Prison | Med Cognitive	0.754	1.351	0.512	0.277
No Prison to Recent-Prison | High Cognitive	0.836	2.022	0.768	0.290
Conditional on D=0:				
No Prison to Early-Prison	0.486	0.726	0.391	0.328
No Prison to Early-Prison | Low Cognitive	0.602	0.999	0.459	0.289
No Prison to Early-Prison | Med Cognitive	0.507	0.711	0.357	0.288
No Prison to Early-Prison | High Cognitive	0.375	0.519	0.345	0.353
No Prison to Recent-Prison	0.707	1.437	0.609	0.354
No Prison to Recent-Prison | Low Cognitive	0.484	0.733	0.406	0.368
No Prison to Recent-Prison | Med Cognitive	0.760	1.372	0.516	0.271
No Prison to Recent-Prison | High Cognitive	0.840	2.061	0.779	0.286
Conditional on D=1:				
No Prison to Early-Prison	0.485	0.727	0.394	0.327
No Prison to Early-Prison | Low Cognitive	0.590	0.982	0.458	0.295
No Prison to Early-Prison | Med Cognitive	0.492	0.683	0.352	0.294
No Prison to Early-Prison | High Cognitive	0.363	0.497	0.339	0.356
No Prison to Recent-Prison	0.681	1.336	0.583	0.366
No Prison to Recent-Prison | Low Cognitive	0.470	0.710	0.404	0.379
No Prison to Recent-Prison | Med Cognitive	0.753	1.346	0.510	0.279
No Prison to Recent-Prison | High Cognitive	0.833	2.001	0.762	0.291
Conditional on D=2:				
No Prison to Early-Prison	0.480	0.725	0.397	0.330
No Prison to Early-Prison | Low Cognitive	0.577	0.959	0.458	0.305
No Prison to Early-Prison | Med Cognitive	0.473	0.653	0.346	0.301
No Prison to Early-Prison | High Cognitive	0.350	0.475	0.333	0.361
No Prison to Recent-Prison	0.652	1.238	0.560	0.382
No Prison to Recent-Prison | Low Cognitive	0.454	0.682	0.401	0.394
No Prison to Recent-Prison | Med Cognitive	0.744	1.314	0.504	0.286
No Prison to Recent-Prison | High Cognitive	0.827	1.946	0.747	0.297

Notes: This table presents four measures of mental health growth. Column (1) reports the share of individuals who move down by at least one decile. Column (2) shows the average number of deciles lost among the full population, assigning zero to those who do not move down. Column (3) reports the average mental health loss for individuals who move at least one decile downward ("losers"), while Column (4) reports the average mental health gain for individuals who move at least one decile upward ("winners").

sumptions, such as the existence of multiple measurements (dedicated measures) and independence conditions among measurement errors and latent variables.

Methodologically, this work demonstrates the value and feasibility of combining latent factor models with dynamic structural equations to uncover distributional policy effects under a transparent set of identifying conditions. The introduction of MCMC methods for estimation dramatically reduces the computational complexity required for estimation.

Our empirical results illustrate substantial variation in mental health trajectories associated with incarceration status and baseline skill endowments. Counterfactual analyses reveal that policy effects differ across subpopulations defined by initial skill levels and mental health status, underscoring the limitations of mean effect summaries.

Our results also indicate that incarceration is especially damaging for youths who, under current patterns, are least likely to be incarcerated and have stronger baseline endowments, especially when incarceration happens later. In this sense, the mental health consequences of incarceration are most severe precisely for those with more to lose. A natural interpretation is that these youths start with greater stocks of cognitive and emotional resources, so incarceration represents a larger shock to their trajectories, and may also be less prepared to cope with the prison environment than peers who are more likely to experience it.

Our analysis is restricted to youth who have already engaged in relatively serious criminal behavior at an early age, a subgroup that accounts for a disproportionate share of overall youth crime. While this focus provides valuable insight into a high-impact segment of the population, it may not generalize to the broader population. In particular, the mental health consequences of incarceration documented here may not generalize to individuals without prior justice involvement. Whether similar patterns would arise in the broader youth population remains an open empirical question, and addressing this issue constitutes an important direction for future research.

References

F. Agostinelli and M. Wiswall. Estimating the technology of children's skill formation. *Journal of Political Economy*, 133(3):846–887, 2025.

M. Bhuller, G. B. Dahl, K. V. Løken, and M. Mogstad. Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324, 2020.

M. Bhuller, L. Khoury, and K. V. Løken. Mental health consequences of correctional sentencing. *American Economic Journal: Economic Policy*, 17(1):70–105, 2025.

B. Biasi, M. S. Dahl, and P. Moser. Career effects of mental health: Evidence from an innovation in treating bipolar disorder. *Journal of Political Economy: Microeconomics*, 2025. (forthcoming).

P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

I. A. Binswanger, M. F. Stern, R. A. Deyo, P. J. Heagerty, A. Cheadle, J. G. Elmore, and T. D. Koepsell. Release from prison - a high risk of death for former inmates. *New England Journal of Medicine*, 356(2):157–165, 2007.

I. A. Binswanger, P. M. Krueger, and J. F. Steiner. Prevalence of chronic medical conditions among jail and prison inmates in the usa compared with the general population. *Journal of Epidemiology & Community Health*, 63(11):912–919, 2009.

P. Carneiro, K. T. Hansen, and J. J. Heckman. 2001 lawrence r. klein lecture estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review*, 44(2):361–422, 2003.

F. Cunha, J. J. Heckman, and S. Navarro. Counterfactual analysis of inequality and social mobility. *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, pages 290–348, 2006.

F. Cunha, J. J. Heckman, and S. M. Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3):883–931, 2010.

J. Currie and M. Stabile. Child mental health and human capital accumulation: The case of adhd. *Journal of Health Economics*, 25(6):1094–1118, 2006.

J. Currie, M. Stabile, P. Manivong, and L. L. Roos. Child health and young adult outcomes. *Journal of Human Resources*, 45(3):517–548, 2010.

C. S. Diaz-Campo, B. H. Hamilton, M. Luccioni, and H. S. Suh. Mental health and the early career dynamics of young men. *Working Paper*, 2025.

S. N. Durlauf and D. S. Nagin. The deterrent effect of imprisonment. In *Controlling Crime:*

Strategies and Tradeoffs, pages 43–94. University of Chicago Press, 2010.

G. S. Fields. The many facets of economic mobility. In *Inequality, Poverty and Well-Being*, pages 123–142. Springer, 2006.

J. Freyberger. Normalizations and misspecification in skill formation models. *The Review of Economic Studies*, 2025. Forthcoming.

J. C. Fruehwirth, S. Navarro, and Y. Takahashi. How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, 34 (4):979–1021, 2016.

A. Garin, D. Koustas, C. McPherson, S. Norris, M. Pecenco, E. K. Rose, Y. Shem-Tov, and J. Weaver. The impact of incarceration on employment, earnings, and tax filing. *Econometrica*, 93(2):503–538, 2025.

G. Gordon, J. B. Jones, U. Neelakantan, and K. Athreya. Incarceration, employment, and earnings: Dynamics and differences. *Review of Economic Dynamics*, 51:677–697, 2023.

A. Haglund, D. Tidemalm, J. Jokinen, N. Långström, P. Lichtenstein, S. Fazel, and B. Runeson. Suicide after release from prison: A population-based cohort study from sweden. *Journal of Clinical Psychiatry*, 75(10):1047, 2014.

K. Hauck and T. Woutersen. Nonparametric identification. In *Teaching Econometrics: A Tribute to R. Carter Hill*, pages 197–206. Springer, 2026.

J. J. Heckman and S. Navarro. Dynamic discrete choice and dynamic treatment effects. *Journal of Econometrics*, 136(2):341–396, 2007.

J. J. Heckman and E. J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences*, 96 (8):4730–4734, 1999.

J. J. Heckman, S. Urzua, and E. Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

R. Hjalmarsson and M. J. Lindquist. The health effects of prison. *American Economic Journal: Applied Economics*, 14(4):234–270, 2022.

Y. Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008.

G. Jolivet and F. Postel-Vinay. A structural analysis of mental health and labour market trajectories. *The Review of Economic Studies*, 92(3):1920–1954, 2025.

K. G. Jöreskog. Structural equation models in the social sciences: Specification, estimation and testing. *Applications of Statistics*, pages 265–287, 1977.

K. G. Jöreskog and A. S. Goldberger. Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a): 631–639, 1975.

J. R. Kling. Incarceration length, employment, and earnings. *American Economic Review*, 96(3): 863–876, 2006.

I. Kotlarski. On characterizing the gamma and the normal distribution. *Pacific Journal of Mathematics*, 20(1):69–76, 1967.

I. Kuziemko. How should inmates be released from prison? an assessment of parole versus fixed-sentence regimes. *The Quarterly Journal of Economics*, 128(1):371–424, 2013.

M. A. Mancino. Rehabilitating futures: Assessing the effects of correctional employment-focused programs on recidivism and employment. *European Economic Review*, 173:104954, 2025.

C. F. Manski. Identification of binary response models. *Journal of the American Statistical Association*, 83(403):729–738, 1988.

R. L. Matzkin. Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica*, pages 239–270, 1992.

R. L. Matzkin. Nonparametric identification. *Handbook of Econometrics*, 6:5307–5368, 2007.

K. C. Monahan, L. Steinberg, E. Cauffman, and E. P. Mulvey. Trajectories of antisocial behavior and psychosocial maturity from adolescence to young adulthood. *Developmental Psychology*, 45 (6):1654, 2009.

M. Mueller-Smith. The criminal and labor market impacts of incarceration. *Working Paper*, 18, 2015.

S. Norris, M. Pecenco, and J. Weaver. The effect of incarceration on mortality. *The Review of Economics and Statistics*, 106(4):956–973, 2024.

B. P. Rao. *Identifiability in Stochastic Models*. Academic Press, 1992.

C. S. Roehrig. Conditions for identification in nonparametric and parametric models. *Econometrica*, 56(2):433–447, 1988.

E. K. Rose and Y. Shem-Tov. How does incarceration affect reoffending? estimating the dose-response function. *Journal of Political Economy*, 129(12):3302–3356, 2021.

D. F. Salazar. *Essays on Criminal Behaviour, Human Capital Formation, and Mental Health*. PhD thesis, The University of Western Ontario (Canada), 2020.

C. A. Schubert, E. P. Mulvey, L. Steinberg, E. Cauffman, S. H. Losoya, T. Hecker, L. Chassin, and G. P. Knight. Operational lessons from the pathways to desistance project. *Youth Violence and Juvenile Justice*, 2(3):237–255, 2004.

K. Turney, C. Wildeman, and J. Schnittker. As fathers and felons: Explaining the effects of current and recent incarceration on major depression. *Journal of Health and Social Behavior*, 53(4): 465–481, 2012.

B. Wang, R. Frank, and S. Glied. Lasting scars: The impact of depression in early adulthood on subsequent labor market outcomes. *Health Economics*, 32(12):2694–2708, 2023.

B. Williams. Identification of the linear factor model. *Econometric Reviews*, 39(1):92–109, 2020.

Appendix

A.1 Priors and Posterior Derivations Used in Estimation

In this appendix, we provide a detailed summary of the estimation steps for our Bayesian model. Because the derivations follow standard approaches in the Bayesian literature, we present the final forms of the derived posteriors and explicitly state the priors and updating rules used. For inference, we implement a Gibbs sampler that cycles through sampling from the conditional posterior distributions of parameters and latent variables, completing the likelihood at each step by conditioning on latent factors C, M, U.

Priors and Posteriors for Baseline Continuous Cognitive Measures

For any continuous baseline cognitive measure j, we assume the regression model:

$$\mathcal{C}_j = X\gamma_{\mathcal{C},j} + C\psi_j + \varepsilon_{\mathcal{C},j}, j \in 1, ..., N_C,\ \varepsilon_{i,\mathcal{C},j} \sim N\left(0, \sigma^2_{\varepsilon_{\mathcal{C},j}}\right). \tag{1}$$

- Priors
  - Coefficients on covariates, $\gamma_{\mathcal{C},j}$: non-informative prior.
  - Loading, $\psi_j \sim N\left(0, \sigma^2_{0,\psi,j}\right)$ with $\sigma^2_{0,\psi,j} = 10$ (diffuse).
  - Uniqueness variance: non-informative prior on $\frac{1}{\sigma^2_{\varepsilon_{\mathcal{C},j}}}$.
- Posteriors
  - Conditional posterior for $\gamma_{\mathcal{C},j}$:

$$\gamma_{\mathcal{C},j} \mid \centerdot \sim N\left(\widehat{\gamma}_{\mathcal{C},j}, \Sigma_{\gamma,\mathcal{C},j}\right),\ \widehat{\gamma}_{\mathcal{C},j} = \left(X'X\right)^{-1} X'\left(\mathcal{C}_j - C\psi_j\right),\ \Sigma_{\gamma,\mathcal{C},j} = \sigma^2_{\varepsilon_{\mathcal{C},j}}\left(X'X\right)^{-1}. \tag{2}$$

  - Conditional posterior for ψ_j:

$$\psi_j \sim N\left(\widehat{\psi}_j, \Sigma_{\psi,j}\right),\ \Sigma_{\psi,j} = \frac{1}{\sigma^2_{\varepsilon_{\mathcal{C},j}}} C'C + \frac{1}{\sigma^2_{0,\psi,j}},\ \widehat{\psi}_j = \Sigma^{-1}_{\psi,j} \frac{1}{\sigma^2_{\varepsilon_{\mathcal{C},j}}} C'\left(\mathcal{C}_j - X\gamma_{\mathcal{C},j}\right). \tag{3}$$

  - Conditional posterior for uniqueness precision:

$$\frac{1}{\sigma^2_{\varepsilon_{\mathcal{C},j}}} \sim Gamma\left(\frac{n_j}{2}, \frac{\varepsilon'_{\mathcal{C},j}\varepsilon_{\mathcal{C},j}}{2}\right), \tag{4}$$

where n_j is the number of observations for this equation.

The above procedure is repeated for each continuous cognitive measure.

Priors and Posteriors for Ordered Discrete Measures

For ordered discrete measures (e.g., BSI Depression), the model assumes latent indices

$$\mathcal{M}_{i,j}^{*} = X_i'\gamma_{\mathcal{M},j} + M_i\mu_j + \varepsilon_{i,\mathcal{M},j}, \quad \varepsilon_{i,\mathcal{M},j} \sim N(0,1). \tag{5}$$

We use $\mathcal{M}_{i,j}^{*}$ for data completion.

- Priors
  - Coefficients on covariates, $\gamma_{\mathcal{M},j}$: non-informative prior.
  - Loading, $\mu_j \sim N\left(0, \sigma_{0,\mu,j}^2\right)$ with $\sigma_{0,\mu,j}^2 = 10$ (diffuse).
  - Cutoffs, $o_{\mathcal{M},j,k} \sim \text{Unif}\left(\underline{o}, \overline{o}\right)$ with $\underline{o} = -1000$ and $\overline{o} = 1000$.
- Posteriors
  - Coefficients $\gamma_{\mathcal{M},j}, \mu_j$ sampled like in a linear regression (variance fixed to 1).
  - Cutoffs updated via uniform distributions constrained by observed latent categories:

$$o_{\mathcal{M},j,k} \sim \text{Unif}\left(\max\left(o_{\mathcal{M},j,k-1}, \max_{i:\mathcal{M}_{i,j}=k} \mathcal{M}_{i,j,}^{*}, \underline{o}\right), \min\left(o_{\mathcal{M},j,k+1}, \min_{i:\mathcal{M}_{i,j}=k+1} \mathcal{M}_{i,j,}^{*}, \overline{o}\right)\right). \tag{6}$$

  - Latent indices $\mathcal{M}_{i,j}^{*}$ sampled from truncated normal distributions:

$$\mathcal{M}_{i,j}^{*} \sim TN_{(o_{\mathcal{M},j,\mathcal{M}_{i,j}-1}, o_{\mathcal{M},j,\mathcal{M}_{i,j}}]}\left(X_i'\gamma_{\mathcal{M},j} + M_i\mu_j, 1\right). \tag{7}$$

Priors and Posteriors for Treatment Model

In the baseline specification, the treatment is modeled as an ordered outcome, so its estimation follows the same approach used for ordered discrete measures. The binary treatment case arises as a special instance of this framework. In particular, the treatment variable D_i can be represented with a single cutoff normalized to zero, implying that the associated latent index is drawn from truncated normal distributions, either above or below this threshold.

Multinomial Missingness Model

For the multinomial model governing missing data patterns seven years after baseline, let $R_i \in \{0, 1, 2\}$ denote the missingness category for individual i, where $R_i = 0$ denotes fully observed outcomes; $R_i = 1$ denotes missing mental health only; and $R_i = 2$ denotes missing both mental health and treatment.

The model assumes a latent index structure for each category $k = 1, 2$:

$$R_i^{*,k} = X_i \beta_R^k + \gamma_C^k C_i + \gamma_M^k M_i + \epsilon_{i,R}^k, \quad \epsilon_{i,R}^k \sim N(0, 1). \tag{8}$$

The observed missingness category is determined by:

$$R_i = \begin{cases} 0 & \text{if } R_i^{*,1} < 0 \text{ and } R_i^{*,2} < 0 \\ 1 & \text{if } R_i^{*,1} > 0 \text{ and } R_i^{*,1} > R_i^{*,2} \\ 2 & \text{if } R_i^{*,2} > 0 \text{ and } R_i^{*,2} > R_i^{*,1} \end{cases}. \tag{9}$$

Priors:

- Coefficients on covariates, β_R^k: non-informative prior.
- Loadings on latent factors, γ_C^k, γ_M^k: $N(0, \sigma_{0,\gamma}^2)$ with $\sigma_{0,\gamma}^2 = 10$ (diffuse).
- Error variances normalized to 1 for identification.

Posteriors:

Conditional on C_i and M_i, the coefficients β_R^k, γ_C^k, and γ_M^k are sampled using standard Bayesian probit regression methods.

The latent indices $R_i^{*,k}$ are sampled from truncated normal distributions conditional on the observed category:

$$R_i^{*,k} \sim \text{TN}\left(X_i \beta_R^k + \gamma_C^k C_i + \gamma_M^k M_i, 1, a_{R_i}^k, b_{R_i}^k\right) \tag{10}$$

where the truncation bounds $(a_{R_i}^k, b_{R_i}^k)$ depend on the observed missingness pattern R_i:

- If $R_i = 0$: $R_i^{*,1} \in (-\infty, 0)$, $R_i^{*,2} \in (-\infty, 0)$
- If $R_i = 1$: $R_i^{*,1} \in (0, \infty)$, $R_i^{*,2} \in (-\infty, R_i^{*,1})$

- If $R_i = 2$: $R_i^{*,2} \in (0, \infty)$, $R_i^{*,1} \in (-\infty, R_i^{*,2})$

This follows the standard data augmentation approach for multinomial probit models, where latent utilities are completed at each Gibbs iteration and regression parameters are updated conditional on these completed values.

Mental Health System in Period 7

Conditioned on (C_i, M_i, U_i) and all the other parameters, the latent factor $M_{i,7}$ is known from the law of motion equation. Thus, the measurement model at period 7 follows the same ordered discrete setup described above, and sampling proceeds analogously.

Posteriors for the Law of Motion Parameters

Incorporating the law of motion into the measurement model, define:

$$\mathcal{M}_{i,j,7}^{*} = X_i'\gamma_{\mathcal{M},j,7} + \left(\delta_T' D_i + \lambda_C C_i + \lambda_M M_i + \lambda_{CT}' C_i D_i + \lambda_{MT}' M_i D_i + U_i\right)\mu_{j,7} + \varepsilon_{i,\mathcal{M},j,7}, \tag{11}$$

where D_i is a vector that stacks the multiple treatment dummies, and $\delta_T, \lambda_{CT}, \lambda_{MT}$ are vectors of parameters stacking the effect for each treatment. Rearranging and defining:

$$\mathcal{M}_{i,j,7}^{**} = \delta_T D_{i,j} + \lambda_C C_{i,j} + \lambda_M M_{i,j} + \lambda_{CT} C D_{i,j} + \lambda_{MT} M D_{i,j} + U_{i,j} + \varepsilon_{i,\mathcal{M},j,7}, \tag{12}$$

with notation such as $D_{i,j} = D_i \mu_{j,7}$, $C_{i,j} = C_i \mu_{j,7}$, and so on. This is a standard linear regression model like the ones we have been describing, except that the same parameters δ_T, λ_C, etc, show up in all period 7 equations. As a consequence, the sample moves over both i and j. In other words, with a non-informative prior on δ_T, for example, the posterior would be

$$\delta_T \sim N\left(\mu_{\delta_T}, \Sigma_{\delta_T}\right) \tag{13}$$

with

$$Z_{ij}^{\delta_T} = \mathcal{M}_{i,j,7}^{**} - \lambda_C C_{i,j} - \lambda_M M_{i,j} - \lambda_{CT}' C D_{i,j} - \lambda_{MT}' M D_{i,j} - U_{i,j}, \tag{14}$$

$$\Sigma_{\delta_T} = \left(\sum_{i,j} D_{i,j} D_{i,j}'\right)^{-1} \tag{15}$$

and

$$\mu_{\delta_T} = \Sigma_{\delta_T} \left(\sum_{i,j} D_{i,j} Z_{i,j}^{\delta_T} \right). \tag{16}$$

An equivalent posterior follows for the remaining parameters, as in the case of the measurement systems we described before. The only difference being that the sum is done over both i and j.

Posteriors for the factors

Next is the derivation of the posterior for the parameters describing the distributions of the latent factors C, M, and U. We begin by describing the posterior for the parameters of the mixture for the univariate case of U. We complete the likelihood by defining h_{Ui} to be a variable that takes the value h if the element of the mixture U_i is drawn from is element h. Let $I_{U,h} = \{i : h_{Ui} = h\}, n_{U,h} = |I_{U,h}|$.

Given a normal prior for the h^{th} component mean such that

$$m_{U,h} \sim N\left(m_{U0}, s_{U0}\right), \tag{17}$$

the posterior for the h^{th} component mean is:

$$m_{U,h} \sim N\left(\bar{m}_{U,h}, V_{U,h}^{-1}\right), \quad V_{U,h} = s_{U0} + \frac{n_{U,h}}{\sigma_{U,h}^2}, \quad \bar{m}_{U,h} = \frac{s_{U0} m_{U0} + \frac{1}{\sigma_{U,h}^2} \sum_{i \in I_{U,h}} U_i}{V_{U,h}}. \tag{18}$$

In order to have a diffuse prior, we set $m_{U0} = 0$ and $s_{U0} = 10$.

Given a Gamma prior for the h^{th} component inverse of the variance such that

$$\frac{1}{\sigma_{U,h}^2} \sim Gamma\left(a_0, b_0\right) \tag{19}$$

The variances are updated according to the posterior:

$$\frac{1}{\sigma_{U,h}^2} \sim \text{Gamma}\left(a_0 + \frac{n_{U,h}}{2},\ b_0 + \frac{1}{2} \sum_{i \in I_{U,h}} \left(U_i - m_{U,h}\right)^2\right). \tag{20}$$

We set $a_0 = b_0 = 2$.

Given a Dirichlet prior for the mixture weights such that

$$\pi_U \sim Dirichlet\left(a_U, ..., a_U\right), \tag{21}$$

the posterior becomes,

$$\pi_U \sim \text{Dirichlet}(a_U + n_{U,1}, \ldots, a_U + n_{U,2}). \tag{22}$$

In order to update the mixture component indicators h_{Ui}, we notice that the conditional likelihood is given by

$$P(h_{Ui} = h \mid U_i, \{m_{U,h}, \sigma^2_{U,h}, \pi_{U,h}\}) \propto \pi_{U,h} \cdot \phi\left(U_i; m_{U,h}, \sigma^2_{U,h}\right). \tag{23}$$

This defines a discrete distribution over h, which can be sampled by computing log-weights:

$$\log w_{U,h} = \log \pi_{U,h} - \frac{1}{2}\left[\frac{(U_i - m_{U,h})^2}{\sigma^2_{U,h}} + \log \sigma^2_{U,h}\right]. \tag{24}$$

Then, defining $\rho_{i,U,h} \propto e^{\log w_{U,h}}$ and normalizing $\sum_h \rho_{i,U,h} = 1$ we can sample from

$$h_{Ui} \sim \text{Categorical}(\rho_{i,U,1}, \ldots, \rho_{i,U,3}). \tag{25}$$

Next, we notice that, conditional on the mixture component indicators, sampling the parameters of the distributions for C_i, M_i follows the exact same process as that for U_i. The only difference arises when sampling the mixture component indicators, as they need to consider both factors. In this case, the log-weights are computed as

$$\log w_{CM,h} = \log \pi_{CM,h} - \frac{1}{2}\left[\frac{(C_i - m_{C,h})^2}{\sigma^2_{C,h}} + \log \sigma^2_{C,h} + \frac{(M_i - m_{M,h})^2}{\sigma^2_{M,h}} + \log \sigma^2_{M,h}\right]. \tag{26}$$

Defining $\rho_{i,CM,h} \propto e^{\log w_{CM,h}}$ and normalizing $\sum_h \rho_{i,CM,h} = 1$ we can sample from the posterior

$$h_{CMi} \sim Categorical\left(\rho_{i,CM,h}, ..., \rho_{i,CM,6}\right). \tag{27}$$

The final step is to describe the Gibbs sampler steps to sample the latent factors that we use for the completion. Take any of the factors, say C_i. The relevant terms of the posterior are

$$\begin{aligned} f\left(C_i \mid \cdot\right) \propto & \left[\prod_{j=1}^{N_C} \phi\left(\mathcal{C}^*_{i,j}; X_i'\gamma_{\mathcal{C},j} + C_i\psi_j, \sigma^2_{\varepsilon_{\mathcal{C},j}}\right)\right] \times \phi\left(C_i; m_{C0}, \sigma^2_{C0}\right) \\ & \times \phi\left(T_i; X_i'\gamma_T + C_i\psi_T + M_i\mu_T, 1\right) \\ & \times \prod_{j=1}^{N_{M_7}} \phi\left(\mathcal{M}^{**}_{i,j,7}; \delta_T D_{i,j} + \lambda_C C_{i,j} + \lambda_M M_{ij} + \lambda_{CT} CD_{i,j} + \lambda_{MT} MD_{i,j} + U_{i,1j}, 1\right) \end{aligned} \tag{28}$$

This is the product of a normal likelihood and a normal prior, so the posterior is normal. Let us define:

$$\frac{1}{\sigma^2_{C|i}} = \sum_{j=1}^{N_C} \frac{\psi_j^2}{\sigma^2_{\varepsilon_{\mathcal{C},j}}} + \psi_T^2 + \sum_{j=1}^{N_{M_7}} \left(\lambda_C \mu_{j,7} + \lambda_{CT} D_i \mu_{j,7}^2\right) + \frac{1}{\sigma^2_{C0}} \tag{29}$$

$$\eta_{C|i} = \begin{pmatrix} \sum_{j=1}^{N_C} \frac{\psi_j}{\sigma^2_{\varepsilon_{\mathcal{C},j}}} \left(\mathcal{C}^*_{i,j} - X_i' \gamma_{\mathcal{C},j}\right) + \psi_T \left(T_i - X_i' \gamma_T - M_i \mu_T\right) + \frac{m_{c0}}{\sigma^2_{C0}} \\ \sum_{j=1}^{N_{M_7}} \left(\lambda_C \mu_{j,7} + \lambda_{CT} D_i \mu_{j,7}\right) \left(\mathcal{M}^{**}_{i,j,7} - \delta_T D_{i,j} - \lambda_M M_{i,j} - \lambda_{MT} M D_{i,j} - U_{i,j}\right) \end{pmatrix}$$

Then the posterior for C_i is:

$$C_i \mid \cdot \sim N\left(m_{C|i}, \sigma^2_{C|i}\right), \quad \text{where} \quad m_{C|i} = \sigma^2_{C|i} \eta_{C|i}. \tag{30}$$

The posteriors for M_i and U_i can be derived in a similar fashion.

A.2 Empirical Application - Additional Results

A.2.1 Additional Model Results

Table A.2.1: Estimated Parameters from Factor Model — Mixtures

	(1) Cognitive Factor C	(2) Mental Health Baseline M	(3) U
Mixture 1			
Mean	-0.005	-0.370	0.301
Variance	0.504	0.958	0.378
Probability	0.266	0.266	0.792
Mixture 2			
Mean	-0.028	-0.566	-1.412
Variance	0.586	1.072	1.396
Probability	0.217	0.217	0.208
Mixture 3			
Mean	-0.006	-0.433	
Variance	0.521	0.991	
Probability	0.279	0.279	
Mixture 4			
Mean	-0.030	-0.421	
Variance	0.575	1.001	
Probability	0.238	0.238	

Notes: This table reports the parameter estimates from the mixture model for the joint distribution of baseline latent skills (i.e., cognitive skills and mental health) and a subsequent shock (U_i) to the mental health production function. We report the mean of each parameter across 5,000 draws from the posterior distribution.

Table A.2.2: Estimated Parameters from Factor Model — Cognitive Skills Measures

	(1) WASI IQ	(2) Stroop Color	(3) Stroop Word	(4) Stroop Color/Word	(5) Trail Making Part A	(6) Trail Making Part B
Constant	-0.214	-0.325	-0.218	-0.502	0.818	1.119
	(0.200)	(0.216)	(0.210)	(0.210)	(0.258)	(0.275)
Age 15	-0.019	0.228	0.145	0.300	0.696	0.410
	(0.129)	(0.139)	(0.136)	(0.134)	(0.167)	(0.173)
Age 16	-0.085	0.311	0.266	0.402	0.805	0.421
	(0.120)	(0.126)	(0.124)	(0.125)	(0.152)	(0.160)
Age 17	0.071	0.389	0.201	0.468	0.751	0.531
	(0.125)	(0.133)	(0.129)	(0.129)	(0.158)	(0.167)
Age 18	0.097	0.094	0.234	0.150	0.899	0.511
	(0.173)	(0.182)	(0.178)	(0.176)	(0.214)	(0.229)
Female	-0.148	0.127	0.212	0.066	0.238	0.134
	(0.099)	(0.105)	(0.103)	(0.104)	(0.128)	(0.136)
White	0.484	0.155	0.117	0.280	0.452	0.290
	(0.189)	(0.204)	(0.200)	(0.198)	(0.242)	(0.254)
Hispanic	-0.231	-0.062	-0.150	-0.010	0.071	-0.134
	(0.187)	(0.202)	(0.196)	(0.192)	(0.235)	(0.247)
Black	-0.068	-0.006	-0.209	-0.038	-0.122	-0.192
	(0.188)	(0.203)	(0.197)	(0.195)	(0.237)	(0.254)
Phoenix	0.527	0.055	0.215	0.256	0.380	0.395
	(0.094)	(0.100)	(0.098)	(0.097)	(0.122)	(0.129)
Variance	0.682	0.382	0.530	0.579	1.000	1.000
	(0.041)	(0.038)	(0.037)	(0.039)	(0.000)	(0.000)
Cognitive Factor	1.000	1.649	1.391	1.277	0.927	1.344
	(0.000)	(0.126)	(0.112)	(0.109)	(0.132)	(0.153)
Cutoff 1					0.000	0.000
					(0.000)	(0.000)
Cutoff 2					0.789	1.199
					(0.083)	(0.074)
Cutoff 3					2.054	2.093
					(0.110)	(0.090)

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the cognitive measure system. Standard errors are shown in parentheses below the mean point estimates. IQ and the Stroop components are modeled using a linear-in-parameters specification, while the Trail-Making tests are estimated using an ordered threshold model. The cognitive skills factor is normalized to have a loading of one on the WASI IQ score.

Table A.2.3: Estimated Parameters from Factor Model — Mental Health Measures (Baseline)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	0.531	0.726	0.875	0.797	1.371	0.288	-0.289	1.707	0.934
	(0.350)	(0.368)	(0.392)	(0.328)	(0.417)	(0.329)	(0.336)	(0.359)	(0.368)
Age 15	0.171	0.124	-0.117	0.284	0.047	-0.001	-0.049	0.127	0.094
	(0.225)	(0.240)	(0.251)	(0.214)	(0.265)	(0.213)	(0.219)	(0.226)	(0.231)
Age 16	-0.090	0.154	-0.059	0.236	0.204	-0.063	-0.091	0.023	0.151
	(0.208)	(0.221)	(0.228)	(0.197)	(0.239)	(0.197)	(0.203)	(0.202)	(0.217)
Age 17	0.139	0.465	0.404	0.508	0.527	0.148	0.197	0.450	0.302
	(0.219)	(0.234)	(0.243)	(0.209)	(0.255)	(0.204)	(0.212)	(0.217)	(0.232)
Age 18	0.255	0.959	0.379	0.422	0.879	0.133	0.141	1.213	1.046
	(0.301)	(0.317)	(0.325)	(0.292)	(0.351)	(0.283)	(0.291)	(0.309)	(0.317)
Female	0.529	0.444	0.321	0.616	0.386	0.449	0.195	0.270	0.311
	(0.179)	(0.195)	(0.201)	(0.173)	(0.217)	(0.161)	(0.167)	(0.183)	(0.190)
White	0.133	-0.172	-0.242	0.061	-0.173	-0.219	-0.271	-0.432	-0.698
	(0.326)	(0.346)	(0.359)	(0.308)	(0.386)	(0.303)	(0.319)	(0.328)	(0.344)
Hispanic	0.063	0.053	0.028	-0.128	-0.186	-0.096	0.023	-0.234	-0.296
	(0.315)	(0.335)	(0.350)	(0.298)	(0.377)	(0.295)	(0.308)	(0.323)	(0.334)
Black	-0.471	-0.425	-0.649	-0.210	-0.595	-0.283	-0.210	-0.256	-0.597
	(0.325)	(0.348)	(0.360)	(0.306)	(0.386)	(0.304)	(0.314)	(0.327)	(0.339)
Phoenix	-0.216	-0.139	-0.058	-0.066	0.280	0.105	-0.042	-0.312	-0.070
	(0.165)	(0.179)	(0.188)	(0.158)	(0.200)	(0.158)	(0.159)	(0.169)	(0.177)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-1.072	-1.206	-1.278	-1.000	-1.442	-0.878	-0.844	-1.092	-1.173
	(0.092)	(0.103)	(0.109)	(0.000)	(0.123)	(0.079)	(0.080)	(0.095)	(0.101)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.641	0.727	0.725	0.569	0.563	0.716	0.612	0.618	0.675
	(0.058)	(0.061)	(0.066)	(0.060)	(0.063)	(0.057)	(0.057)	(0.068)	(0.067)
Cutoff 3	1.019	1.248	1.249	1.055	1.199	1.142	0.955	1.089	1.149
	(0.073)	(0.076)	(0.087)	(0.079)	(0.086)	(0.073)	(0.073)	(0.088)	(0.085)
Cutoff 4	1.462	1.635	1.706	1.457	1.631	1.657	1.268	1.582	1.715
	(0.090)	(0.088)	(0.103)	(0.090)	(0.101)	(0.092)	(0.088)	(0.102)	(0.104)
Cutoff 5	1.794	1.956	2.114	1.912	2.121		1.531	2.090	2.276
	(0.102)	(0.102)	(0.122)	(0.101)	(0.120)		(0.100)	(0.116)	(0.122)
Cutoff 6	2.094	2.243	2.451		2.543				
	(0.112)	(0.111)	(0.135)		(0.135)				
Cutoff 7	2.379								
	(0.125)								

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure.

Table A.2.4: Estimated Parameters from Factor Model — Mental Health Measures (Year 7)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	-0.124	0.669	0.149	0.078	1.219	-0.913	-0.985	0.867	-0.513
	(0.530)	(0.769)	(0.705)	(0.520)	(0.635)	(0.630)	(0.547)	(0.590)	(0.713)
Age 15	0.088	0.656	0.211	-0.115	0.247	0.320	0.466	0.588	0.731
	(0.288)	(0.403)	(0.351)	(0.278)	(0.320)	(0.336)	(0.322)	(0.310)	(0.381)
Age 16	-0.046	0.221	-0.466	-0.138	0.053	0.231	0.186	0.156	0.348
	(0.262)	(0.373)	(0.332)	(0.258)	(0.293)	(0.324)	(0.313)	(0.283)	(0.357)
Age 17	0.081	0.554	0.348	0.390	0.505	0.128	0.642	0.203	0.316
	(0.265)	(0.377)	(0.331)	(0.259)	(0.300)	(0.330)	(0.302)	(0.286)	(0.361)
Age 18	-0.023	0.012	-0.366	0.473	0.365	-0.433	0.447	0.451	0.504
	(0.377)	(0.535)	(0.489)	(0.360)	(0.429)	(0.508)	(0.419)	(0.400)	(0.485)
Female	0.410	0.118	0.173	0.489	0.519	0.422	-0.068	-0.097	-0.120
	(0.207)	(0.304)	(0.276)	(0.210)	(0.247)	(0.248)	(0.248)	(0.231)	(0.287)
White	0.027	-1.316	-0.043	0.674	-0.661	0.169	-0.032	-0.472	-0.700
	(0.520)	(0.761)	(0.712)	(0.517)	(0.623)	(0.609)	(0.540)	(0.584)	(0.693)
Hispanic	-0.322	-1.078	-0.241	0.354	-0.957	-0.415	0.258	-0.509	-0.079
	(0.517)	(0.752)	(0.703)	(0.515)	(0.615)	(0.611)	(0.530)	(0.578)	(0.691)
Black	-0.192	-1.121	0.041	0.483	-1.079	0.038	-0.116	-0.178	0.025
	(0.511)	(0.750)	(0.697)	(0.513)	(0.617)	(0.602)	(0.520)	(0.569)	(0.688)
Phoenix	-0.259	-0.397	0.074	-0.416	-0.017	-0.089	-0.569	-0.290	-0.247
	(0.224)	(0.296)	(0.274)	(0.210)	(0.242)	(0.262)	(0.239)	(0.234)	(0.290)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-0.902	-1.716	-1.561	-1.000	-1.322	-1.156	-0.898	-1.217	-1.516
	(0.114)	(0.233)	(0.183)	(0.000)	(0.166)	(0.148)	(0.120)	(0.155)	(0.192)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.518	0.438	0.672	0.808	0.605	0.546	0.499	0.666	0.640
	(0.073)	(0.081)	(0.087)	(0.090)	(0.081)	(0.089)	(0.086)	(0.084)	(0.100)
Cutoff 3	0.849	0.944	1.341	1.391	0.992	1.018	0.977	1.073	1.240
	(0.092)	(0.123)	(0.129)	(0.110)	(0.102)	(0.124)	(0.127)	(0.101)	(0.142)
Cutoff 4	1.231	1.563	1.873	1.790	1.553	1.322	1.509	1.447	1.677
	(0.117)	(0.172)	(0.163)	(0.126)	(0.127)	(0.147)	(0.169)	(0.116)	(0.168)
Cutoff 5	1.471	1.800	2.441	2.172	1.925		1.734	1.788	2.179
	(0.134)	(0.188)	(0.206)	(0.143)	(0.147)		(0.188)	(0.131)	(0.199)
Cutoff 6	1.604	2.163	2.711		2.121				
	(0.142)	(0.212)	(0.228)		(0.158)				
Cutoff 7	1.778								
	(0.152)								

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system in year 7. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure.

Table A.2.5: Estimated Parameters from Multinomial Model for Missing Data

	Treatment and MH Year 7 Measures Included	Treatment Included, MH Year 7 Measures Missing
Constant	0.140	-0.115
	(0.316)	(0.311)
Age 15	-0.112	0.136
	(0.196)	(0.202)
Age 16	-0.231	0.227
	(0.182)	(0.192)
Age 17	-0.085	-0.105
	(0.186)	(0.201)
Age 18	-0.137	0.075
	(0.256)	(0.275)
Female	0.146	-0.249
	(0.148)	(0.162)
White	0.352	-0.200
	(0.305)	(0.290)
Hispanic	0.278	-0.215
	(0.297)	(0.277)
Black	0.279	-0.118
	(0.298)	(0.282)
Phoenix	-0.271	0.461
	(0.143)	(0.149)
Cognitive Skills Factor C	0.134	0.165
	(0.130)	(0.139)
Mental Health Factor M	-0.209	0.135
	(0.058)	(0.060)

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the attrition equation. The reference category corresponds to observations with both treatment and mental health measures missing in year 7.

A.3 Counterfactuals - Additional Definitions and Figures

A.3.1 Measures of Positional Mobility

Chi-Square Mobility Index (CSMI): The Chi-Square Mobility Index measures how much the observed mobility matrix deviates from random mobility. It compares the observed mobility matrix to a benchmark of random mobility. That is, a hypothetical scenario in which an individual's position in the distribution under incarceration would be randomly assigned, conditional on his position absent incarceration, and we would expect a uniform distribution across deciles. Specifically, each cell in the mobility matrix would contain 10% of the mass. The CSMI quantifies deviations from this uniform benchmark by summing the squared differences between observed and predicted probabilities across all cells. Larger deviations imply more dependence between the two distributions and therefore *less* mobility. Formally, the index is computed as $\text{CSMI}(M_{7,0}, M_{7,1}) = (-1) \times \sum_{i=1}^{10} \sum_{j=1}^{10} \frac{(O_{ij} - 10)^2}{10}$, where O_{ij} is the probability of being in the ith decile of the distribution of $M_{7,0}$ and the jth decile of the distribution of $M_{7,1}$. By definition, the CSMI is negative, with higher values implying greater mobility.

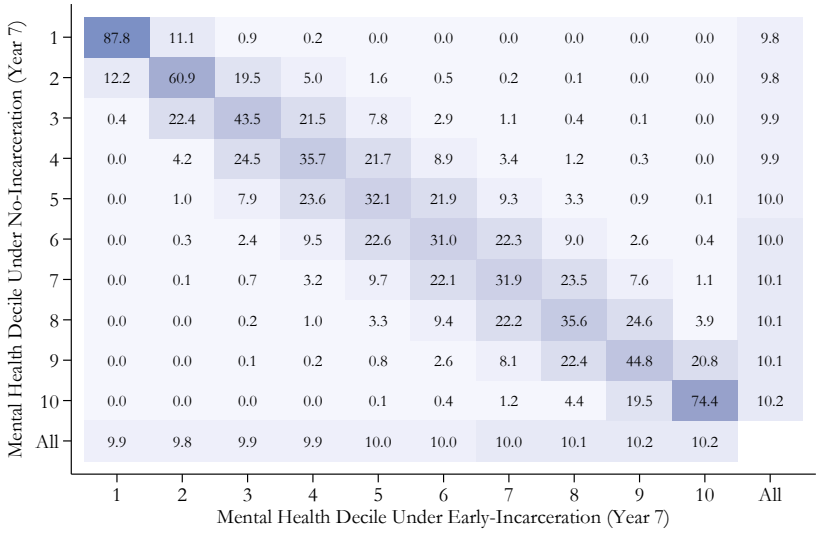
Mobility Ratio Index (MRI):, which measures the share of individuals who do *not* remain on the same decile across the two distributions. The Mobility Ratio Index is computed as $\text{MRI}(M_{7,0}, M_{7,1}) = 1 - \frac{\sum_{i=1}^{10} N_{ii}}{\sum_{i=1}^{10} \sum_{j=1}^{10} N_{ij}}$, where N_{ij} counts the number of individuals that are simultaneously in the ith decile of the distribution of $M_{7,0}$ and the jth decile of the distribution of $M_{7,1}$. The MRI ranges from 0 to 1, with higher values implying greater mobility.

Mean Deciles Move Index (MDMI):, which captures the average number of deciles each individual moves between the two distributions. The Mean Deciles Move Index is computed as $\text{MDMI}(M_{7,0}, M_{7,1}) = \frac{1}{N} \sum_{n=1}^{N} \left| Decile_n(M_{7,1}) - Decile_n(M_{7,0}) \right|$. Higher values of the MDMI imply greater mobility.

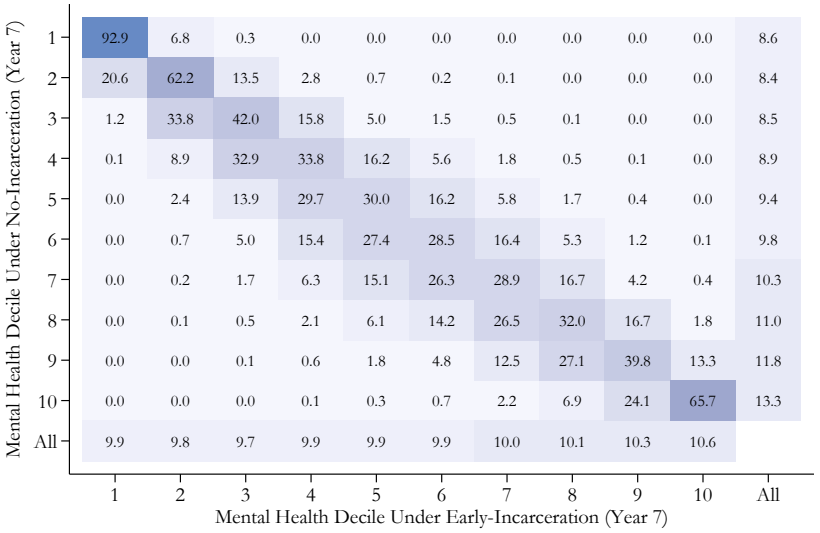
A.3.2 Additional Figures

Figure A.3.1: Positional Mobility Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who do not go to prison ($D = 0$)

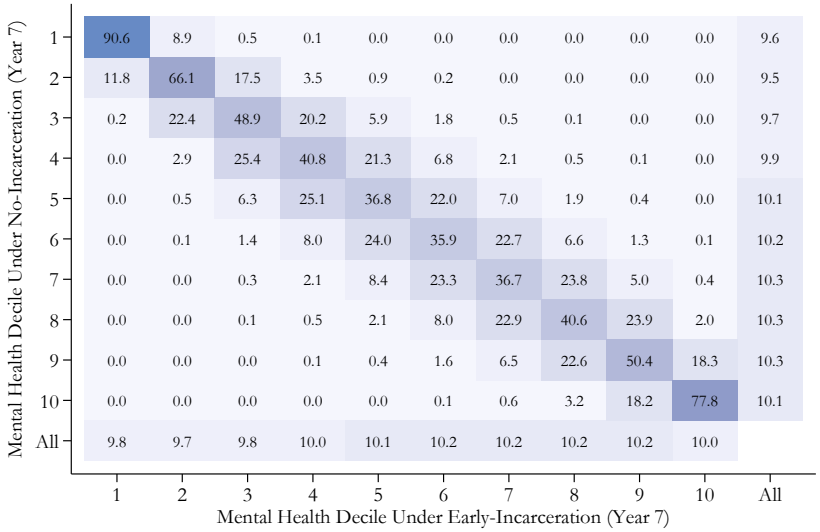
(a) Unconditional



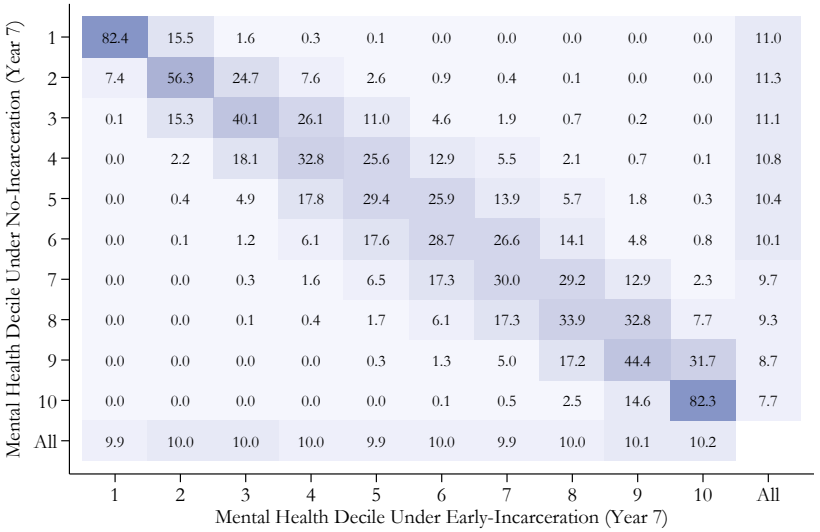
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



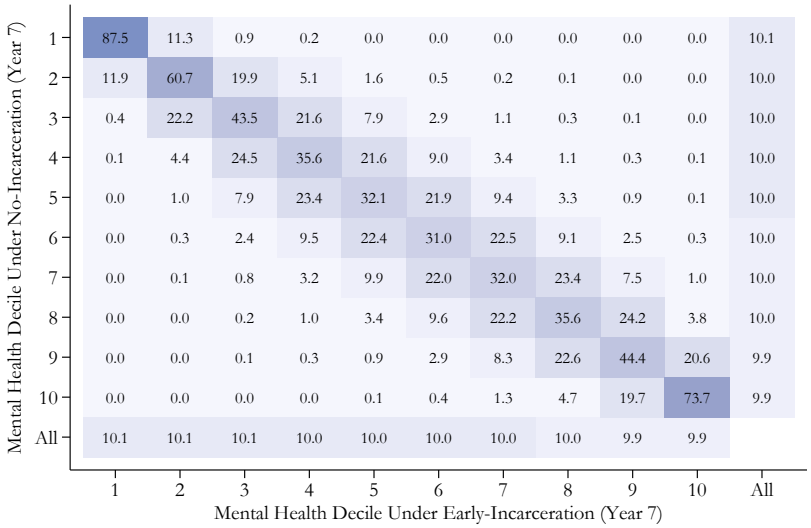
(d) For C (Cognitive) in Tertile 3



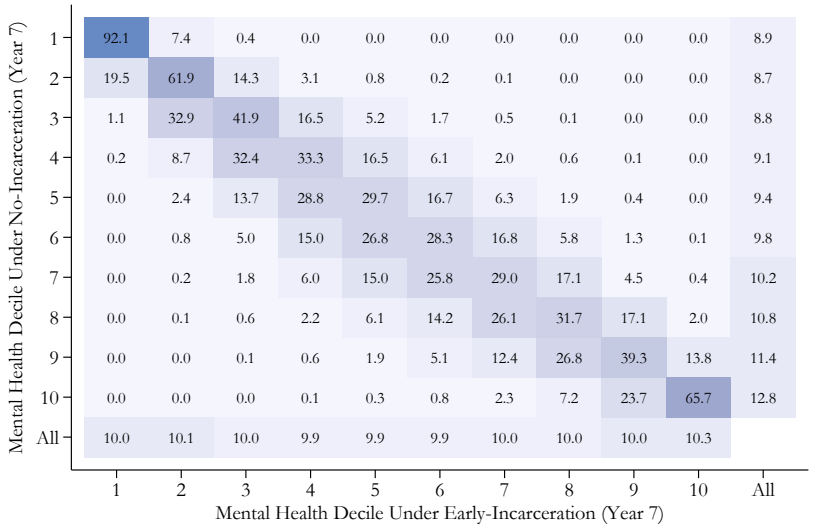
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,1}$ be the kth quantile of $M_{7,1}$. Cell (j,k) in Figure A.3.1a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 0\right)$. Figures A.3.1b-A.3.1d report $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 0, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.2: Positional Mobility Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who go to prison early ($D = 1$)

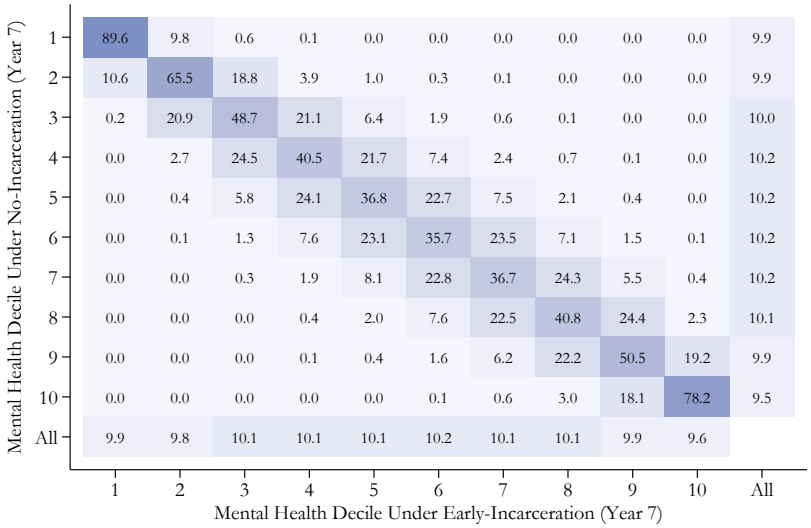
(a) Unconditional



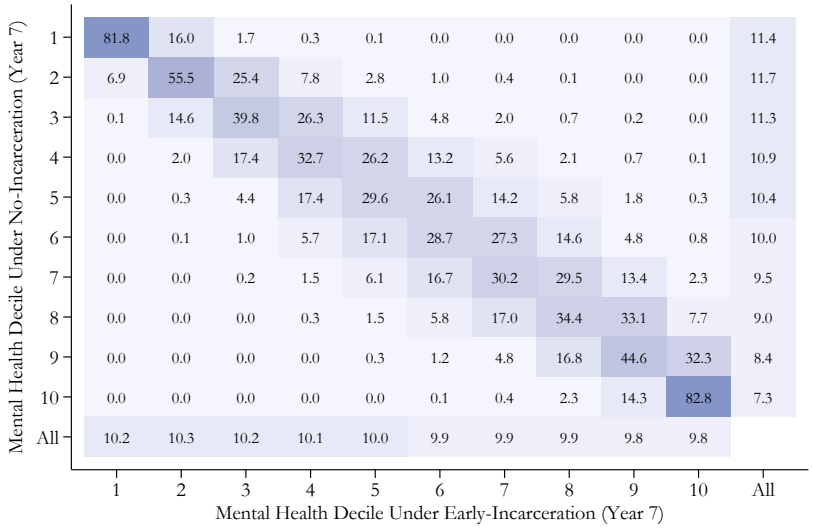
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



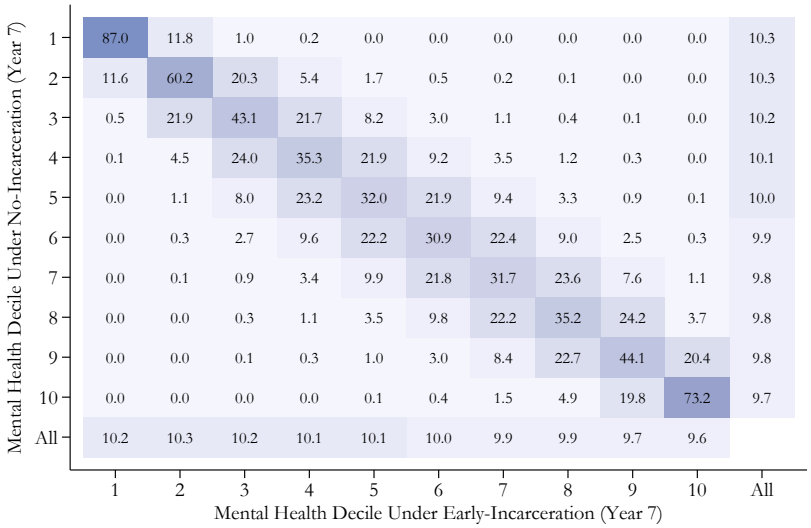
(d) For C (Cognitive) in Tertile 3



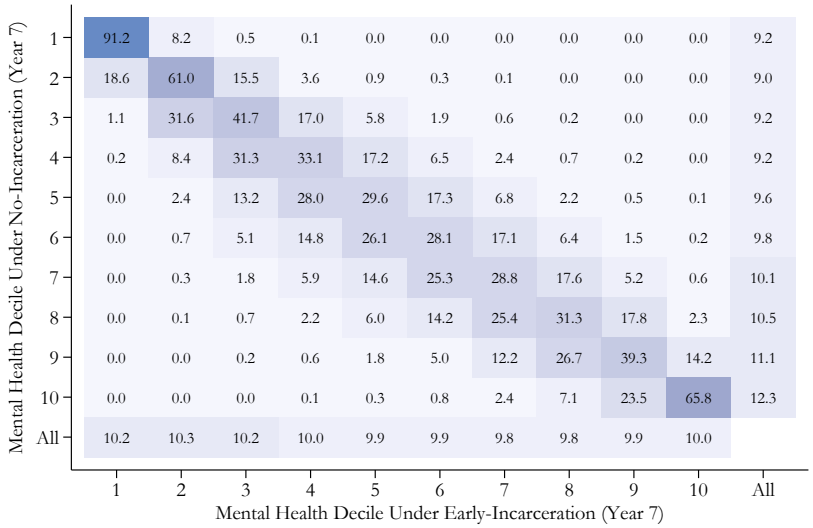
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,1}$ be the kth quantile of $M_{7,1}$. Cell (j,k) in Figure A.3.2a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 1\right)$. Figures A.3.2b-A.3.2d report $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 1, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.3: Positional Mobility Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who go to prison recent ($D = 2$)

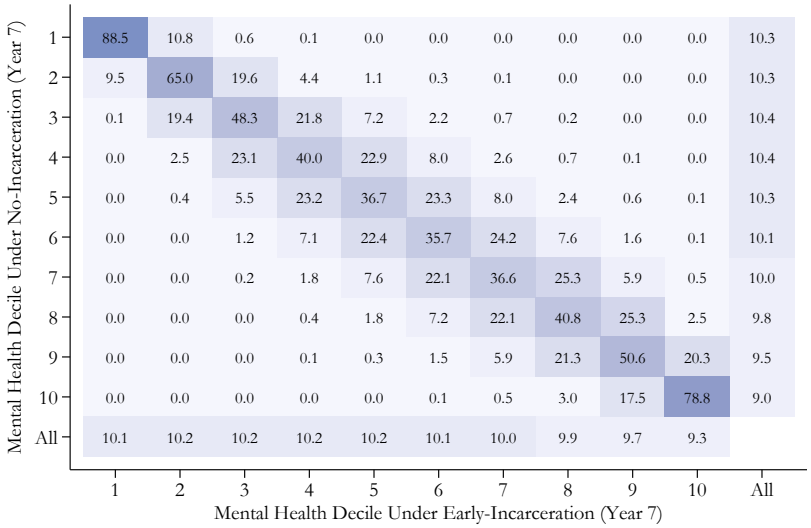
(a) Unconditional



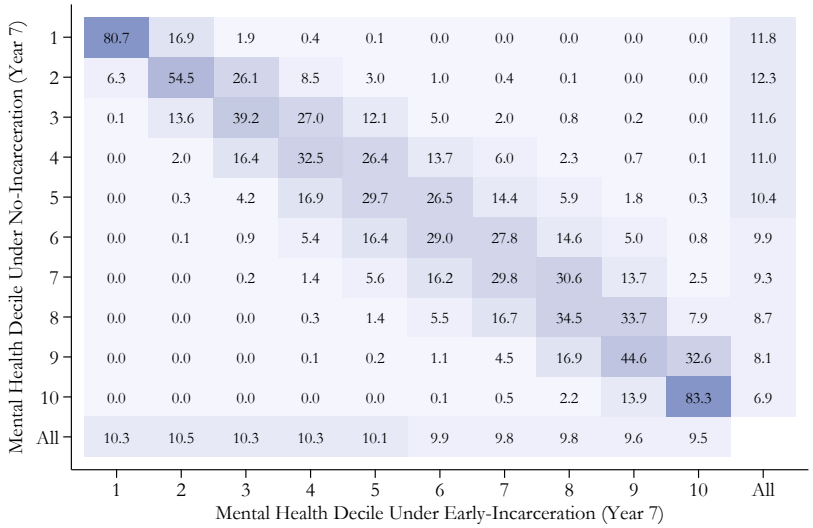
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



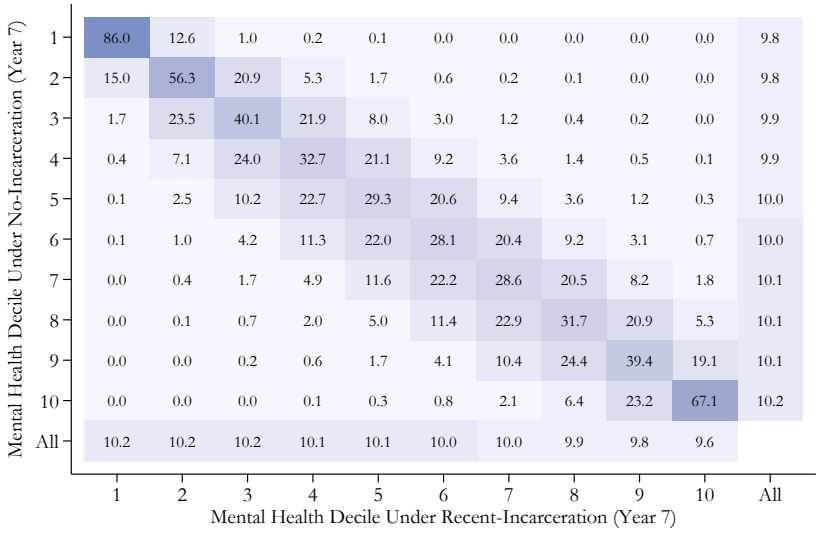
(d) For C (Cognitive) in Tertile 3



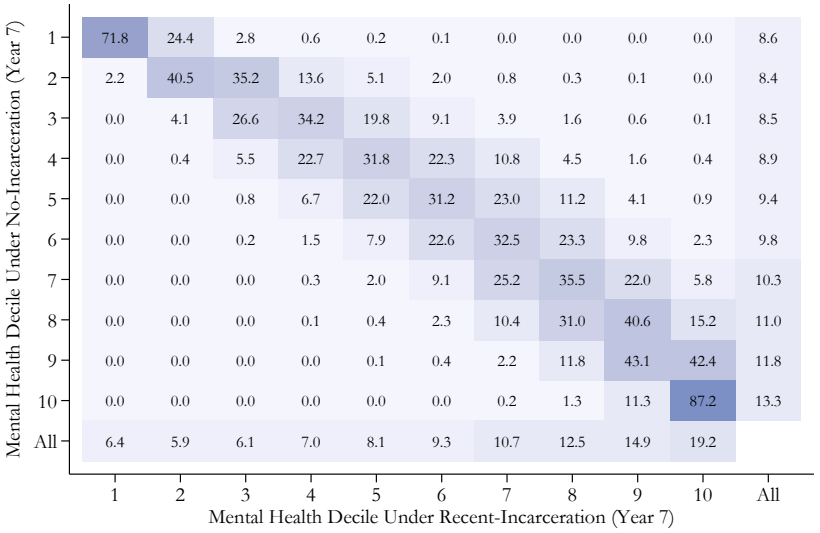
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,1}$ be the kth quantile of $M_{7,1}$. Cell (j,k) in Figure A.3.3a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 2\right)$. Figures A.3.3b-A.3.3d report $\Pr\left(M_{i,7,1} \in (Q_{k-1,1},\, Q_{k,1}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 2, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.4: Positional Mobility Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who do not go to prison ($D = 0$)

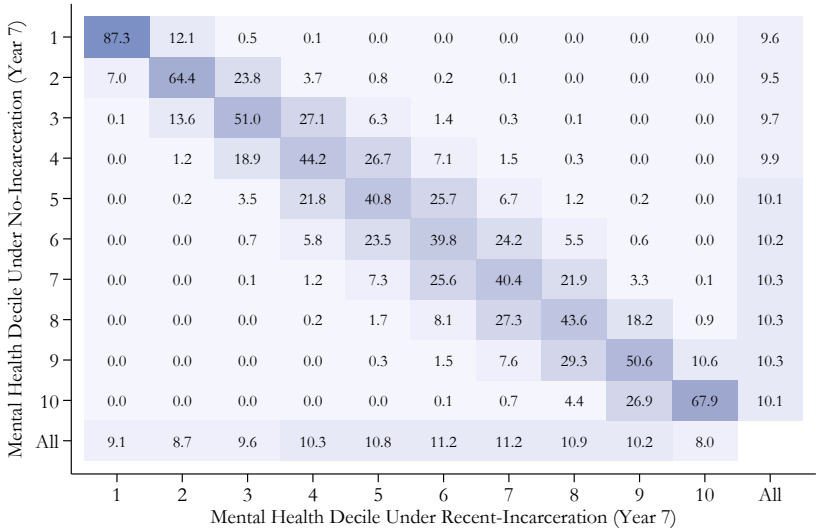
(a) Unconditional



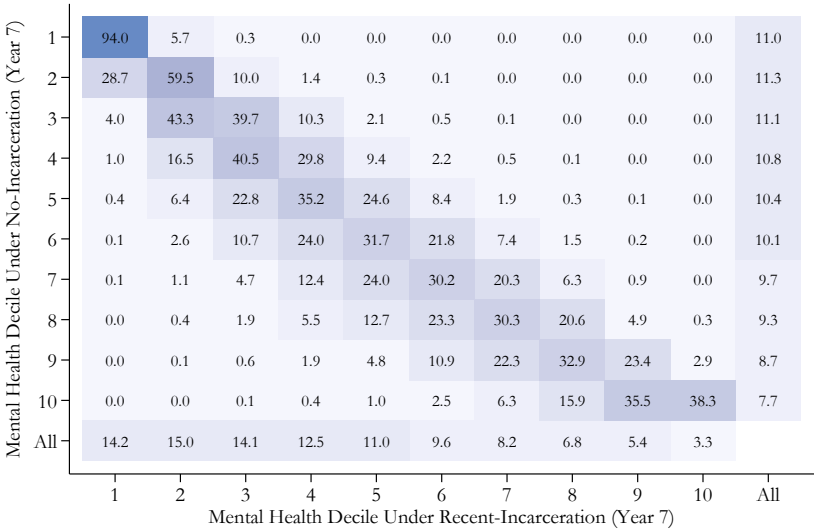
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



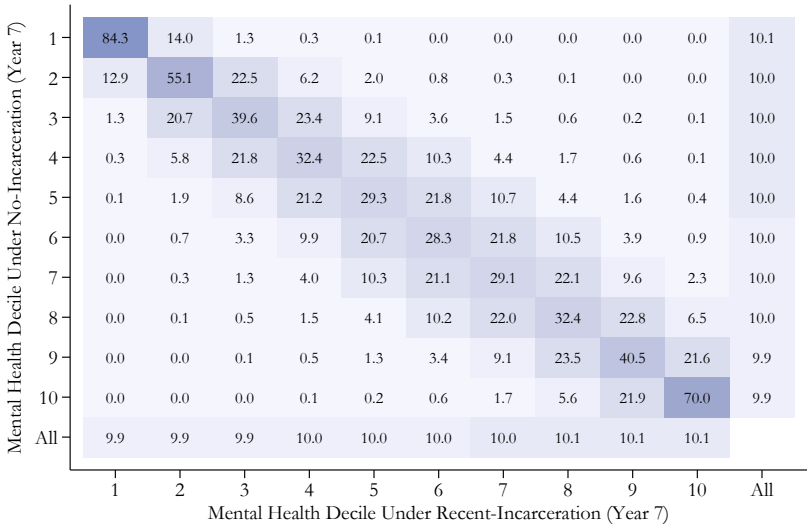
(d) For C (Cognitive) in Tertile 3



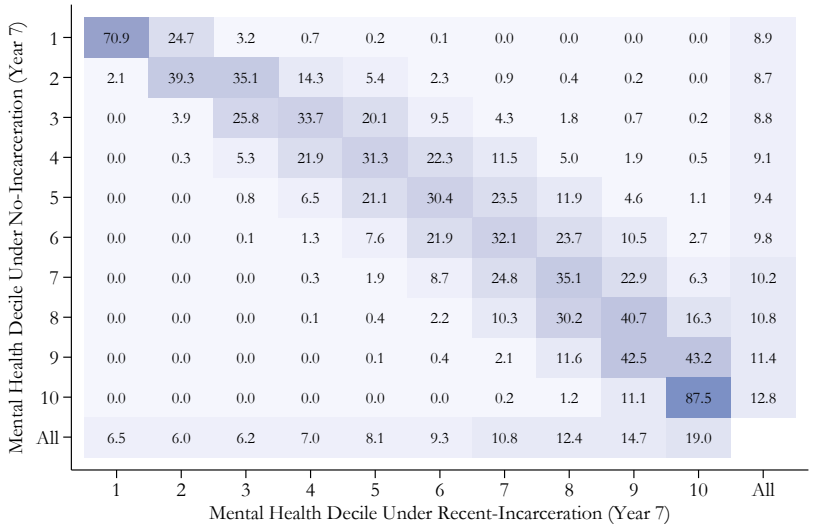
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,2}$ be the kth quantile of $M_{7,2}$. Cell (j,k) in Figure A.3.4a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 0\right)$. Figures A.3.4b-A.3.4d report $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 0, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.5: Positional Mobility Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who go to prison early ($D = 1$)

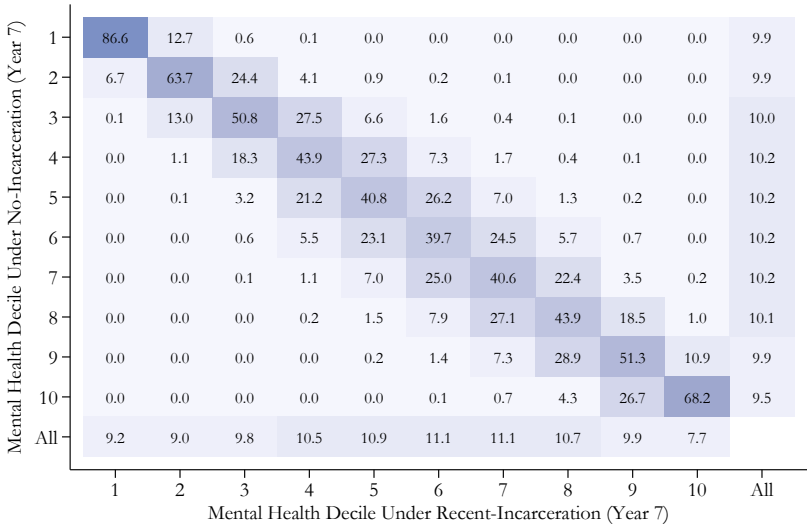
(a) Unconditional



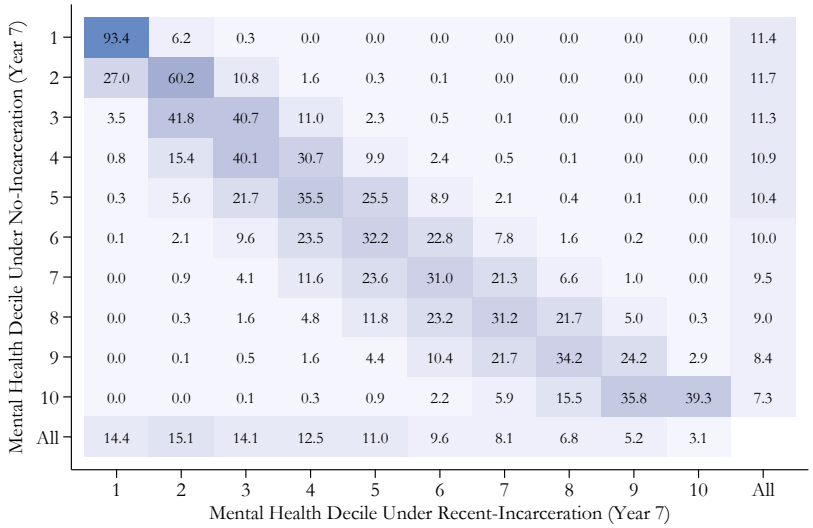
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



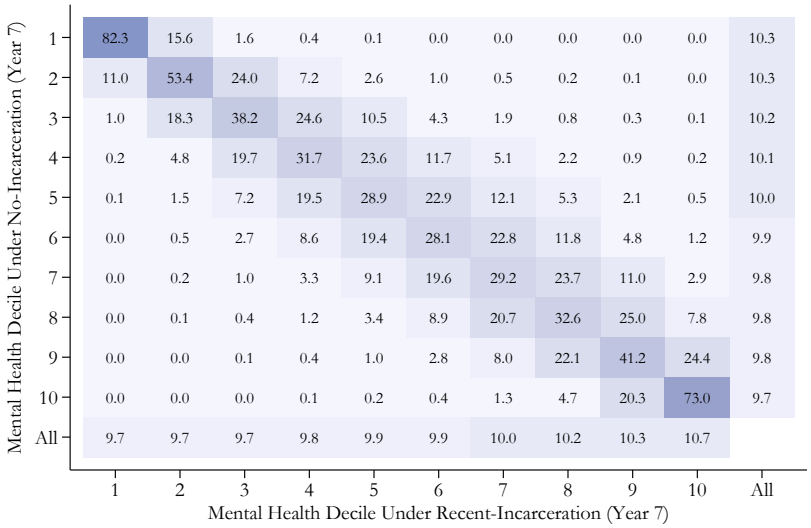
(d) For C (Cognitive) in Tertile 3



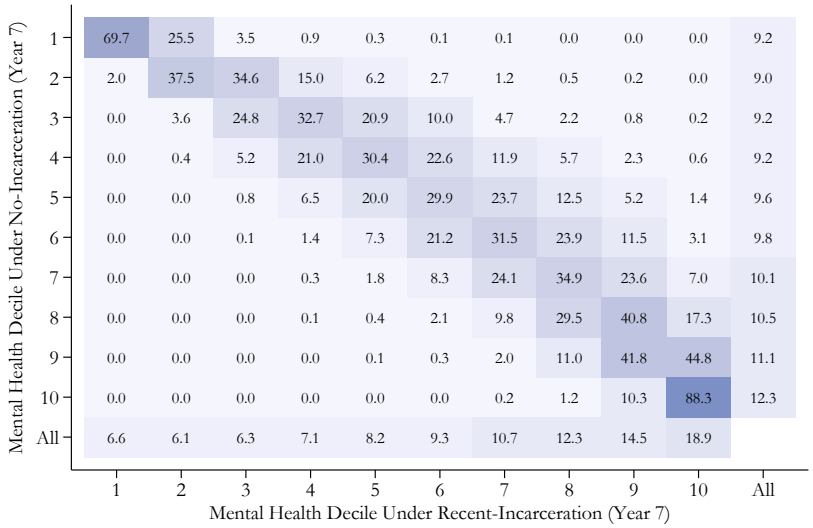
Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,2}$ be the kth quantile of $M_{7,2}$. Cell (j,k) in Figure A.3.5a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 1\right)$. Figures A.3.5b-A.3.5d report $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 1, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.6: Positional Mobility Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who go to prison recent ($D = 2$)

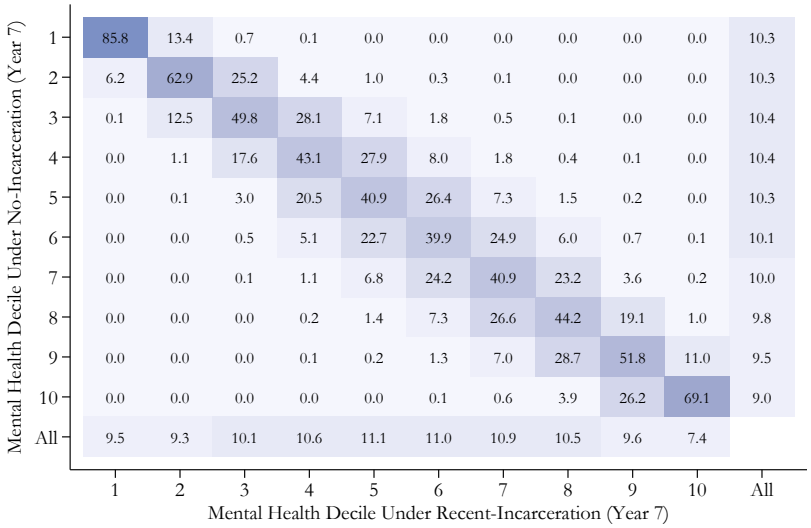
(a) Unconditional



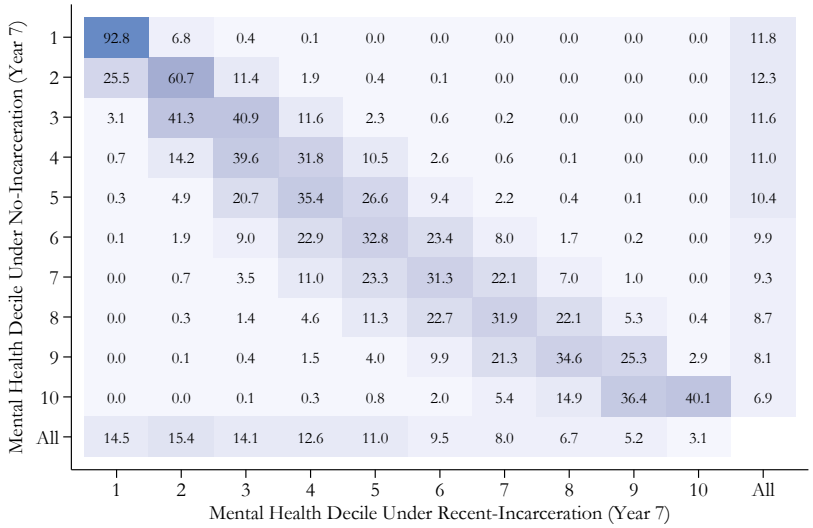
(b) For C (Cognitive) in Tertile 1



(c) For C (Cognitive) in Tertile 2



(d) For C (Cognitive) in Tertile 3



Notes: Let $Q_{j,0}$ be the jth quantile of $M_{7,0}$, and $Q_{k,2}$ be the kth quantile of $M_{7,2}$. Cell (j,k) in Figure A.3.6a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 2\right)$. Figures A.3.6b-A.3.6d report $\Pr\left(M_{i,7,2} \in (Q_{k-1,2},\, Q_{k,2}] \mid M_{i,7,0} \in (Q_{j-1,0},\, Q_{j,0}], D_i = 2, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark blue (100%).

Figure A.3.7: Growth Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who do not go to prison ($D = 0$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.7a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 0\right)$. Figures A.3.7b-A.3.7d report $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 0, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure A.3.8: Growth Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who go to prison early ($D = 1$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.8a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], D_i = 1\right)$. Figures A.3.8b-A.3.8d report $\Pr\left(M_{i,7,1} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], D_i = 1, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure A.3.9: Growth Analyses (Cell %) - $M_{7,1}$ versus $M_{7,0}$, for individuals who go to prison recent ($D = 2$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

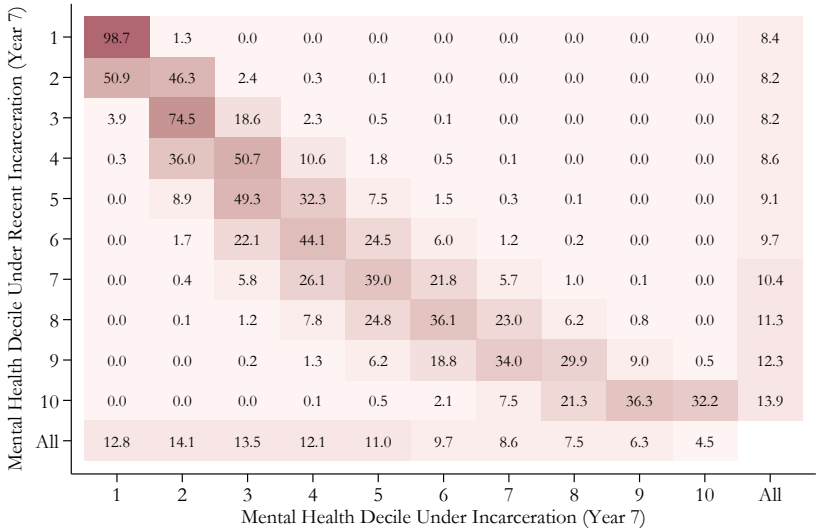
Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.9a reports $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 2\right)$. Figures A.3.9b-A.3.9d report $\Pr\left(M_{i,7,1} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 2, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure A.3.10: Growth Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who do not go to prison ($D = 0$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2



(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.10a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 0\right)$. Figures A.3.10b-A.3.10d report $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 0, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure A.3.11: Growth Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who go to prison early ($D = 1$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.11a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], D_i = 1\right)$. Figures A.3.11b-A.3.11d report $\Pr\left(M_{i,7,2} \in (Q_{k-1},\, Q_k] \mid M_{i,7,0} \in (Q_{j-1},\, Q_j], D_i = 1, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

Figure A.3.12: Growth Analyses (Cell %) - $M_{7,2}$ versus $M_{7,0}$, for individuals who go to prison recent ($D = 2$)

(a) Unconditional

(b) For C (Cognitive) in Tertile 1

(c) For C (Cognitive) in Tertile 2

(d) For C (Cognitive) in Tertile 3

Notes: Let Q_j be the jth quantile of M_7. Cell (j, k) in Figure A.3.12a reports $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 2\right)$. Figures A.3.12b-A.3.12d report $\Pr\left(M_{i,7,2} \in (Q_{k-1}, Q_k] \mid M_{i,7,0} \in (Q_{j-1}, Q_j], D_i = 2, C \in \text{Tertile } t\right)$. Shading ranges from white (0%) to dark red (100%).

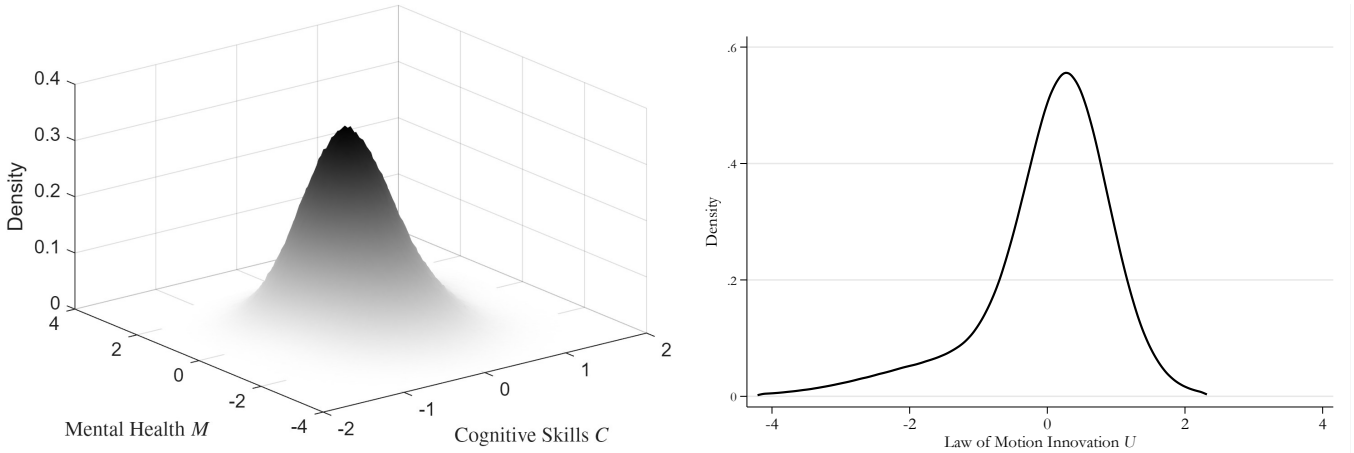
Online Appendix

OA.1 Empirical Application - Robustness Checks

OA.1.1 Binary Treatment Specification - Model Results

Figure OA.1.1: Estimated Posterior Distributions of the Factors

(a) Joint Distribution of C and M (b) Distribution of U



Notes: This figure reports the estimated posterior distributions of the factors. Figure OA.1.1a displays the estimated joint distribution of cognitive skills and mental health at baseline, while Figure OA.1.1b displays the estimated density of the law of motion innovation to mental health in year 7. Results correspond to the binary treatment model with mixtures 4-2.

Table OA.1.1: Estimated Parameters from Factor Model — Mixtures

	(1) Cognitive Factor C	(2) Mental Health Baseline M	(3) U
Mixture 1			
Mean	-0.035	-0.520	-1.367
Variance	0.570	1.072	1.382
Probability	0.238	0.238	0.212
Mixture 2			
Mean	-0.011	-0.415	0.296
Variance	0.527	1.018	0.372
Probability	0.258	0.258	0.788
Mixture 3			
Mean	-0.024	-0.396	
Variance	0.525	1.011	
Probability	0.251	0.251	
Mixture 4			
Mean	-0.020	-0.430	
Variance	0.532	0.991	
Probability	0.253	0.253	

Notes: This table reports the parameter estimates from the mixture model for the joint distribution of baseline latent skills (i.e., cognitive skills and mental health) and a subsequent shock (U_i) to the mental health production function. We report the mean of each parameter across 5,000 draws from the posterior distribution. Results correspond to the binary treatment model with mixtures 4-2.

Table OA.1.2: Estimated Parameters from Factor Model — Cognitive Skills Measures

	(1) WASI IQ	(2) Stroop Color	(3) Stroop Word	(4) Stroop Color/Word	(5) Trail Making Part A	(6) Trail Making Part B
Constant	-0.222	-0.328	-0.219	-0.503	0.823	1.116
	(0.203)	(0.215)	(0.210)	(0.214)	(0.251)	(0.275)
Age 15	-0.019	0.224	0.143	0.302	0.690	0.407
	(0.132)	(0.139)	(0.138)	(0.137)	(0.166)	(0.174)
Age 16	-0.082	0.310	0.267	0.404	0.797	0.420
	(0.122)	(0.127)	(0.125)	(0.123)	(0.149)	(0.159)
Age 17	0.072	0.385	0.198	0.471	0.745	0.530
	(0.125)	(0.130)	(0.131)	(0.127)	(0.154)	(0.165)
Age 18	0.101	0.099	0.238	0.151	0.894	0.515
	(0.173)	(0.180)	(0.180)	(0.179)	(0.213)	(0.228)
Female	-0.143	0.133	0.218	0.069	0.239	0.139
	(0.100)	(0.105)	(0.103)	(0.101)	(0.130)	(0.134)
White	0.493	0.164	0.121	0.279	0.442	0.293
	(0.192)	(0.202)	(0.201)	(0.201)	(0.237)	(0.254)
Hispanic	-0.221	-0.051	-0.144	-0.006	0.070	-0.132
	(0.188)	(0.198)	(0.196)	(0.198)	(0.229)	(0.249)
Black	-0.062	-0.005	-0.209	-0.041	-0.132	-0.193
	(0.190)	(0.198)	(0.197)	(0.199)	(0.231)	(0.254)
Phoenix	0.525	0.053	0.215	0.254	0.379	0.394
	(0.094)	(0.099)	(0.098)	(0.099)	(0.118)	(0.128)
Variance	0.682	0.378	0.531	0.583	1.000	1.000
	(0.040)	(0.038)	(0.037)	(0.039)	(0.000)	(0.000)
Cognitive Factor	1.000	1.652	1.389	1.269	0.919	1.333
	(0.000)	(0.127)	(0.114)	(0.109)	(0.129)	(0.151)
Cutoff 1					0.000	0.000
					(0.000)	(0.000)
Cutoff 2					0.783	1.195
					(0.071)	(0.077)
Cutoff 3					2.045	2.089
					(0.085)	(0.093)

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the cognitive measure system. Standard errors are shown in parentheses below the mean point estimates. IQ and the Stroop components are modeled using a linear-in-parameters specification, while the Trail-Making tests are estimated using an ordered threshold model. The cognitive skills factor is normalized to have a loading of one on the WASI IQ score. Results correspond to the binary treatment model with mixtures 4-2.

Table OA.1.3: Estimated Parameters from Factor Model — Mental Health Measures (Baseline)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	0.549	0.748	0.887	0.799	1.374	0.297	-0.282	1.680	0.926
	(0.350)	(0.375)	(0.386)	(0.332)	(0.423)	(0.324)	(0.335)	(0.360)	(0.367)
Age 15	0.178	0.136	-0.113	0.289	0.057	0.001	-0.035	0.137	0.110
	(0.234)	(0.243)	(0.256)	(0.222)	(0.269)	(0.215)	(0.224)	(0.223)	(0.239)
Age 16	-0.083	0.165	-0.051	0.240	0.215	-0.057	-0.072	0.039	0.168
	(0.214)	(0.222)	(0.230)	(0.199)	(0.246)	(0.194)	(0.204)	(0.205)	(0.216)
Age 17	0.143	0.472	0.406	0.505	0.532	0.149	0.206	0.449	0.307
	(0.225)	(0.235)	(0.243)	(0.214)	(0.266)	(0.205)	(0.215)	(0.219)	(0.229)
Age 18	0.267	0.964	0.390	0.421	0.888	0.138	0.151	1.205	1.045
	(0.302)	(0.315)	(0.330)	(0.287)	(0.353)	(0.278)	(0.293)	(0.314)	(0.308)
Female	0.532	0.438	0.322	0.610	0.381	0.446	0.192	0.265	0.308
	(0.178)	(0.196)	(0.200)	(0.173)	(0.216)	(0.162)	(0.168)	(0.179)	(0.188)
White	0.131	-0.185	-0.250	0.056	-0.177	-0.218	-0.284	-0.431	-0.697
	(0.325)	(0.345)	(0.355)	(0.302)	(0.383)	(0.304)	(0.320)	(0.326)	(0.339)
Hispanic	0.064	0.046	0.026	-0.135	-0.181	-0.092	0.015	-0.233	-0.292
	(0.317)	(0.339)	(0.349)	(0.298)	(0.375)	(0.297)	(0.309)	(0.320)	(0.328)
Black	-0.490	-0.450	-0.670	-0.223	-0.609	-0.292	-0.228	-0.266	-0.609
	(0.324)	(0.348)	(0.355)	(0.302)	(0.385)	(0.302)	(0.315)	(0.326)	(0.334)
Phoenix	-0.229	-0.154	-0.072	-0.079	0.266	0.094	-0.049	-0.320	-0.083
	(0.164)	(0.178)	(0.186)	(0.162)	(0.200)	(0.152)	(0.162)	(0.167)	(0.174)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-1.078	-1.209	-1.276	-1.000	-1.438	-0.878	-0.843	-1.075	-1.165
	(0.091)	(0.103)	(0.110)	(0.000)	(0.126)	(0.080)	(0.080)	(0.095)	(0.100)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.647	0.735	0.723	0.558	0.560	0.718	0.614	0.602	0.666
	(0.058)	(0.068)	(0.069)	(0.054)	(0.065)	(0.059)	(0.060)	(0.060)	(0.061)
Cutoff 3	1.031	1.258	1.246	1.037	1.193	1.147	0.956	1.061	1.135
	(0.072)	(0.089)	(0.092)	(0.071)	(0.090)	(0.073)	(0.076)	(0.074)	(0.078)
Cutoff 4	1.475	1.648	1.703	1.436	1.623	1.660	1.270	1.545	1.695
	(0.087)	(0.104)	(0.111)	(0.080)	(0.105)	(0.094)	(0.090)	(0.090)	(0.098)
Cutoff 5	1.811	1.970	2.108	1.892	2.110		1.530	2.045	2.254
	(0.100)	(0.116)	(0.128)	(0.091)	(0.122)		(0.103)	(0.102)	(0.118)
Cutoff 6	2.115	2.256	2.447		2.533				
	(0.109)	(0.126)	(0.143)		(0.139)				
Cutoff 7	2.400								
	(0.119)								

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the binary treatment model with mixtures 4-2.

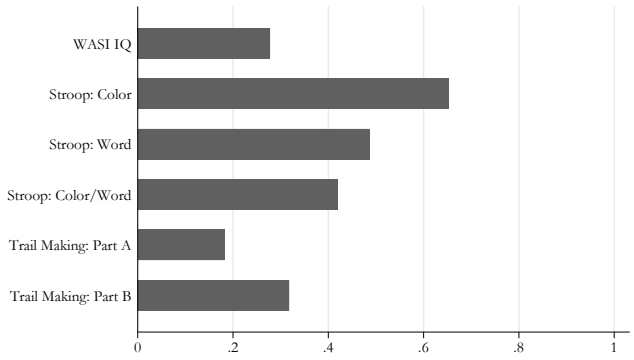
Table OA.1.4: Estimated Parameters from Factor Model — Mental Health Measures (Year 7)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	-0.075	0.738	0.213	0.133	1.280	-0.861	-0.945	0.928	-0.430
	(0.534)	(0.777)	(0.729)	(0.518)	(0.672)	(0.645)	(0.580)	(0.605)	(0.749)
Age 15	0.104	0.656	0.217	-0.102	0.254	0.325	0.463	0.596	0.747
	(0.285)	(0.403)	(0.356)	(0.278)	(0.320)	(0.352)	(0.329)	(0.311)	(0.384)
Age 16	-0.010	0.277	-0.414	-0.099	0.097	0.271	0.205	0.196	0.408
	(0.268)	(0.374)	(0.339)	(0.259)	(0.303)	(0.328)	(0.315)	(0.290)	(0.359)
Age 17	0.103	0.580	0.376	0.410	0.529	0.151	0.653	0.227	0.349
	(0.273)	(0.384)	(0.345)	(0.261)	(0.306)	(0.337)	(0.311)	(0.295)	(0.369)
Age 18	-0.021	0.006	-0.373	0.467	0.359	-0.450	0.434	0.451	0.497
	(0.382)	(0.533)	(0.487)	(0.364)	(0.426)	(0.520)	(0.421)	(0.408)	(0.489)
Female	0.400	0.083	0.152	0.468	0.504	0.401	-0.080	-0.119	-0.142
	(0.209)	(0.305)	(0.273)	(0.204)	(0.243)	(0.252)	(0.241)	(0.231)	(0.295)
White	-0.007	-1.343	-0.074	0.629	-0.686	0.147	-0.046	-0.505	-0.755
	(0.517)	(0.756)	(0.709)	(0.512)	(0.639)	(0.613)	(0.562)	(0.584)	(0.722)
Hispanic	-0.389	-1.176	-0.330	0.267	-1.029	-0.488	0.206	-0.585	-0.186
	(0.517)	(0.750)	(0.700)	(0.508)	(0.635)	(0.613)	(0.555)	(0.579)	(0.714)
Black	-0.227	-1.160	-0.004	0.429	-1.114	0.005	-0.143	-0.220	-0.034
	(0.512)	(0.756)	(0.705)	(0.509)	(0.636)	(0.609)	(0.549)	(0.585)	(0.721)
Phoenix	-0.304	-0.473	0.003	-0.455	-0.082	-0.147	-0.608	-0.349	-0.324
	(0.221)	(0.301)	(0.273)	(0.208)	(0.245)	(0.264)	(0.238)	(0.237)	(0.293)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-0.926	-1.745	-1.584	-1.000	-1.360	-1.195	-0.924	-1.251	-1.568
	(0.118)	(0.234)	(0.193)	(0.000)	(0.169)	(0.158)	(0.125)	(0.163)	(0.210)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.522	0.433	0.668	0.796	0.614	0.551	0.499	0.665	0.649
	(0.074)	(0.078)	(0.092)	(0.085)	(0.082)	(0.090)	(0.080)	(0.082)	(0.107)
Cutoff 3	0.854	0.935	1.325	1.371	1.001	1.030	0.979	1.075	1.256
	(0.093)	(0.117)	(0.135)	(0.109)	(0.103)	(0.130)	(0.121)	(0.103)	(0.151)
Cutoff 4	1.235	1.541	1.850	1.763	1.564	1.338	1.512	1.450	1.697
	(0.115)	(0.163)	(0.169)	(0.125)	(0.129)	(0.154)	(0.167)	(0.120)	(0.175)
Cutoff 5	1.473	1.775	2.415	2.142	1.936		1.736	1.790	2.198
	(0.133)	(0.178)	(0.210)	(0.140)	(0.150)		(0.188)	(0.132)	(0.205)
Cutoff 6	1.605	2.132	2.684		2.131				
	(0.141)	(0.203)	(0.231)		(0.161)				
Cutoff 7	1.779								
	(0.154)								

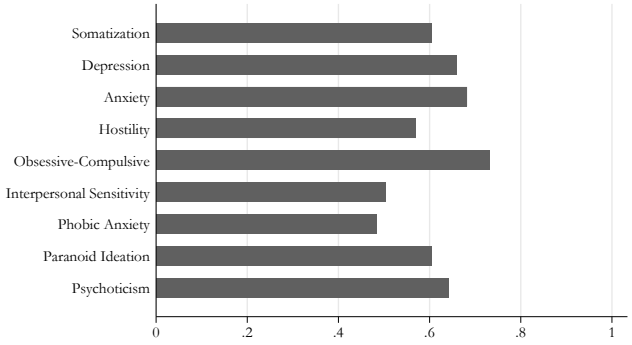
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system in year 7. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.2: Fraction of the Variance Explained by the Factor

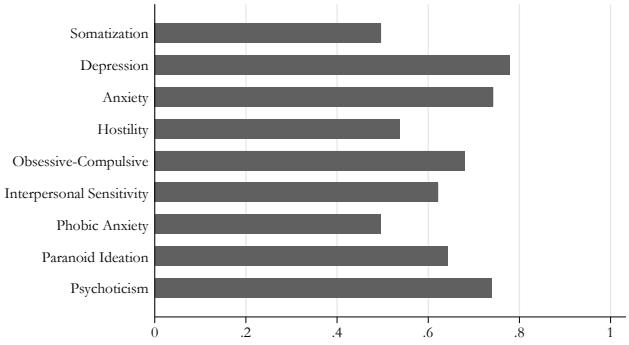
(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



(c) Mental Health (Year 7): M_7



Notes: These figures present the average fraction of the variance of each cognitive and mental health measure explained by the cognitive and mental health factors. For example, Figure OA.1.2a shows that 27.8% of the fraction of the variance of the residualized (against X) WASI IQ measure is explained by the cognitive skills C. Results correspond to the binary treatment model with mixtures 4-2.

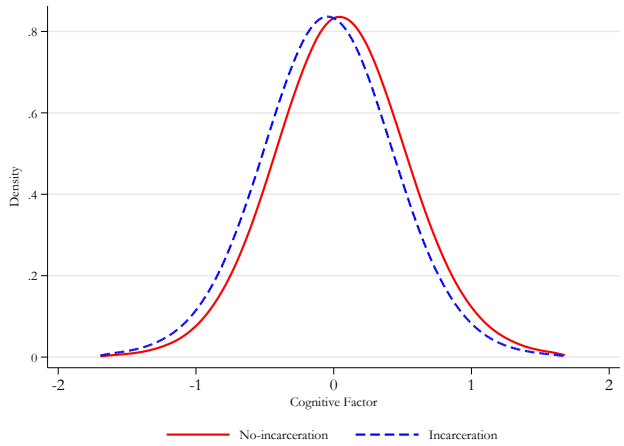
Table OA.1.5: Estimated Parameters from Model - Treatment Equation

	Mean	SD
Constant	-0.129	0.313
Age 15	0.259	0.202
Age 16	0.405	0.187
Age 17	0.182	0.194
Age 18	0.067	0.258
Female	-1.261	0.165
White	-0.120	0.298
Hispanic	0.083	0.292
Black	0.369	0.295
Phoenix	0.281	0.149
Cognitive Skills (ψ_T)	-0.295	0.139
Mental Health (μ_T)	-0.126	0.058
Cutoff 1 (κ_1)	0.000	0.000

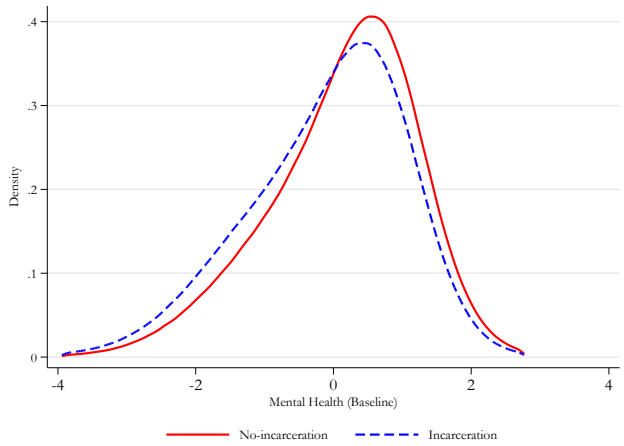
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the treatment equation. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.3: Distribution of Cognitive and Baseline Mental Health Factors By Treatment

(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



Notes: These figures show the estimated densities of the cognitive and baseline mental health factors conditional on incarceration status. For example, in Figure OA.1.3a, the red solid-line plots $f(C \mid D = 0)$ and the blue dashed-line plots $f(C \mid D = 1)$. Results correspond to the binary treatment model with mixtures 4-2.

Table OA.1.6: Estimated Parameters from Model - Law of Motion Equation

	Mean	SD
Incarceration (δ_T)	-0.246	0.115
Cognitive Skills (λ_C)	-0.225	0.159
Mental Health (λ_M)	0.377	0.084
Incarceration $\times$ Cognitive Skills (λ_{CT})	-0.023	0.239
Incarceration $\times$ Mental Health (λ_{MT})	-0.162	0.102

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the law of motion of mental health in the binary treatment model with mixtures 4-2.

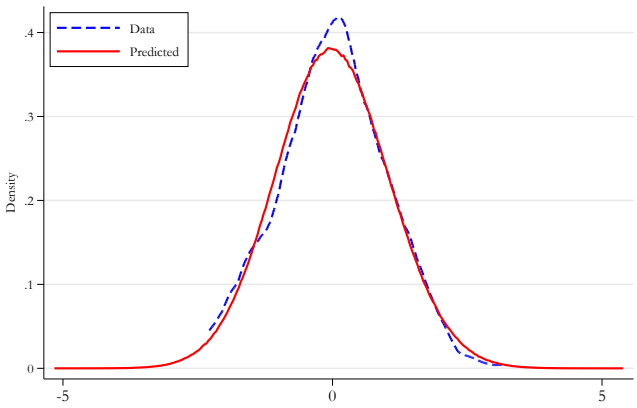
Table OA.1.7: Estimated Parameters from Multinomial Model for Missing Data

	Treatment and MH Year 7 Measures Included	Treatment Included, MH Year 7 Measures Missing
Constant	0.142	-0.105
	(0.307)	(0.306)
Age 15	-0.109	0.132
	(0.199)	(0.205)
Age 16	-0.232	0.220
	(0.181)	(0.189)
Age 17	-0.086	-0.109
	(0.186)	(0.199)
Age 18	-0.135	0.065
	(0.253)	(0.272)
Female	0.150	-0.255
	(0.149)	(0.162)
White	0.351	-0.200
	(0.299)	(0.289)
Hispanic	0.278	-0.215
	(0.292)	(0.278)
Black	0.279	-0.120
	(0.293)	(0.286)
Phoenix	-0.272	0.455
	(0.144)	(0.149)
Cognitive Skills Factor C	0.138	0.160
	(0.133)	(0.140)
Mental Health Factor M	-0.211	0.132
	(0.057)	(0.062)

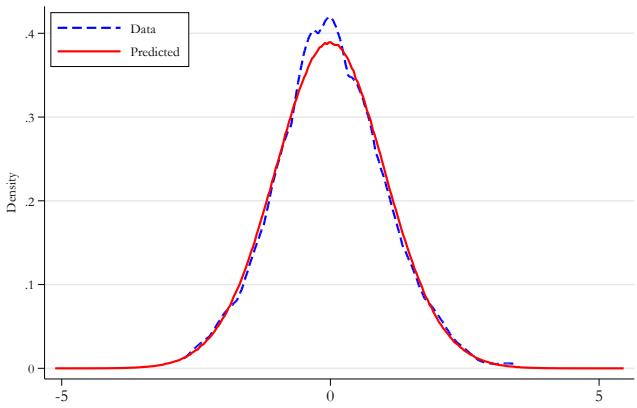
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the attrition equation. The reference category corresponds to observations with both treatment and mental health measures missing in year 7. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.4: Posterior Predictive Fit - Cognitive Measures

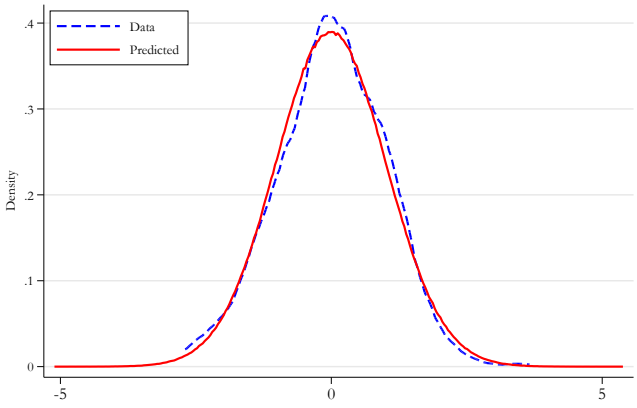
(a) WASI IQ



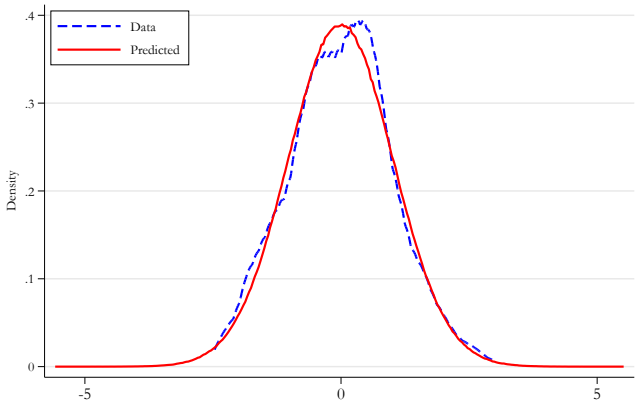
(b) Stroop Color



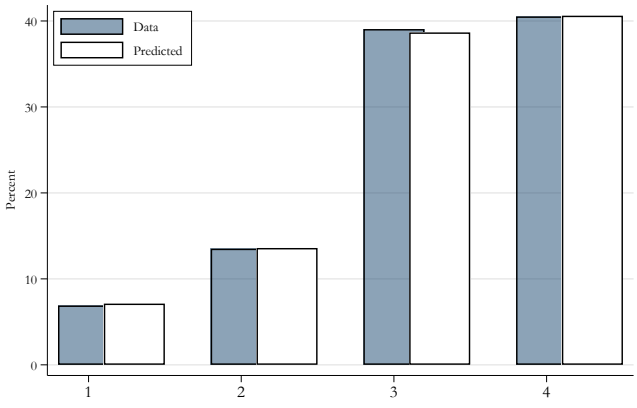
(c) Stroop Word



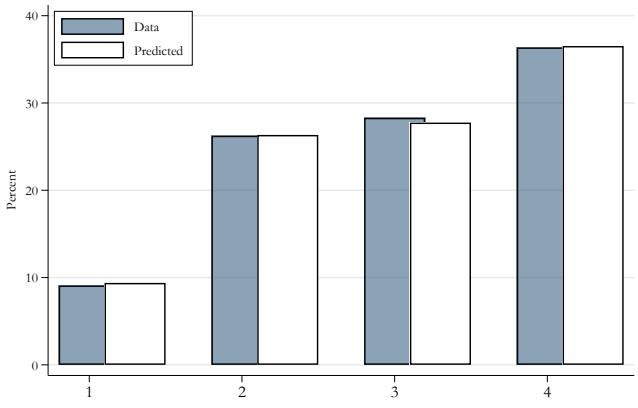
(d) Stroop Color/Word



(e) Trail Making A

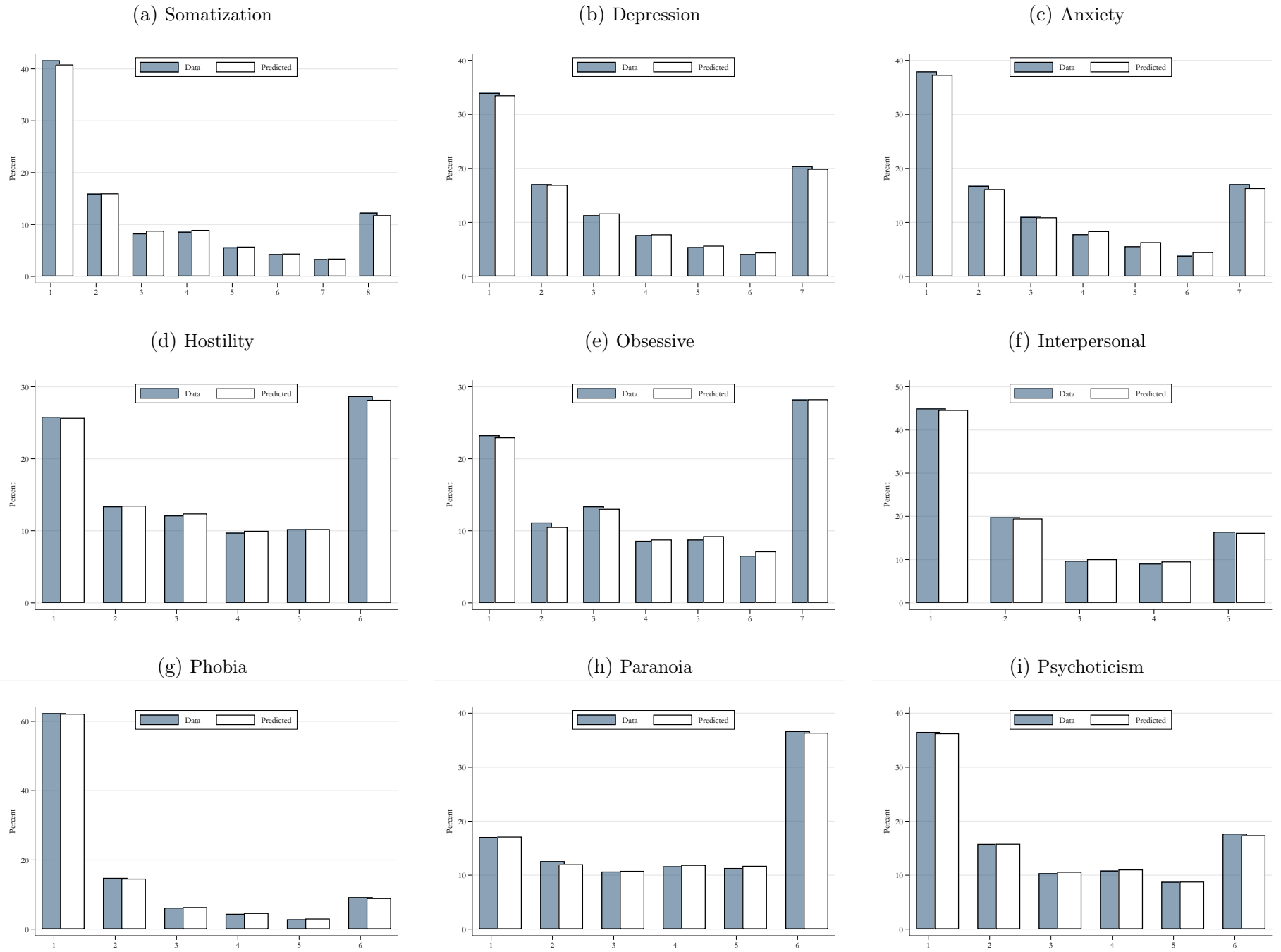


(f) Trail Making B



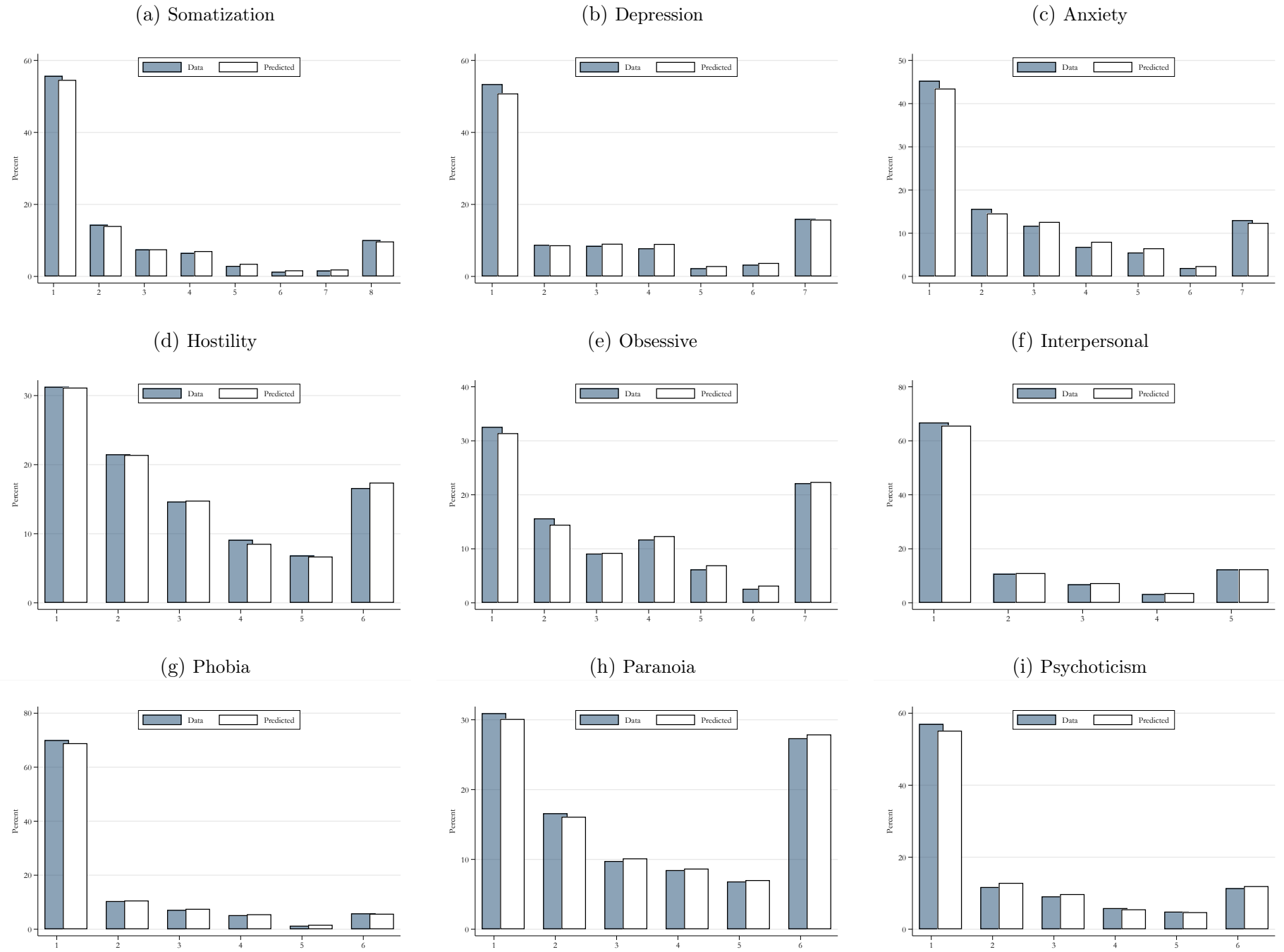
Notes: This figure compares observed and predicted distributions for each cognitive measure. For continuous variables (Figures OA.1.4a-OA.1.4d), observed distributions appear as blue dashed lines and predicted as solid red lines. For discrete variables (Figures OA.1.4e and OA.1.4f), observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.5: Posterior Predictive Fit - BSI Mental Health Measures (Baseline)



Notes: This figure compares observed and predicted distributions for each baseline mental health measure. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.6: Posterior Predictive Fit - BSI Mental Health Measures (Year 7)



Notes: This figure compares observed and predicted distributions for each mental health measure seven years after the baseline. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the binary treatment model with mixtures 4-2.

Table OA.1.8: Goodness Of Fit

	Chi-Square Classical p-value (1)	Mahalanobis Bayesian PPP (2)	Euclidean Bayesian PPP (3)
Cognitive Skills: C			
WASI IQ†	0.047	0.289	0.416
Stroop Color†	0.957	0.589	0.676
Stroop Word†	0.666	0.686	0.750
Stroop CW†	0.043	0.721	0.778
Trail Making A	0.997	0.909	0.923
Trail Making B	0.992	0.396	0.373
Mental Health: M			
BSI Somatization	1.000	0.108	0.331
BSI Depresion	0.999	0.771	0.882
BSI Anxiety	0.963	0.788	0.878
BSI Hostility	1.000	0.844	0.889
BSI Obsessive	0.995	0.363	0.562
BSI Interpersonal	0.993	0.063	0.043
BSI Phobia	0.999	0.136	0.009
BSI Paranoia	0.999	0.668	0.591
BSI Psychoticism	1.000	0.716	0.609
Mental Health: M_7			
BSI Somatization	0.999	0.704	0.758
BSI Depresion	0.977	0.888	0.889
BSI Anxiety	0.956	0.947	0.867
BSI Hostility	1.000	0.652	0.377
BSI Obsessive	0.989	0.416	0.390
BSI Interpersonal	0.994	0.594	0.467
BSI Phobia	0.997	0.772	0.683
BSI Paranoia	0.999	0.345	0.534
BSI Psychoticism	0.986	0.593	0.704

Notes: This table presents three measures of Goodness of Fit for our cognitive and mental health measures. Column (1) displays the *Classical* p-values from the chi-square distribution. Columns (2) and (3) report *Bayesian* Posterior Predictive P-values using the Mahalanobis distance and the joint Euclidean distance of means and standard deviations, respectively. For discrete variables, the total number of bins is the number of categories in the empirical distribution. For continuous variables (indexed by a dag), we set the number of bins equal to 4 and the cutoffs equal to the quartiles of the simulated distribution. Results correspond to the binary treatment model with mixtures 4-2.

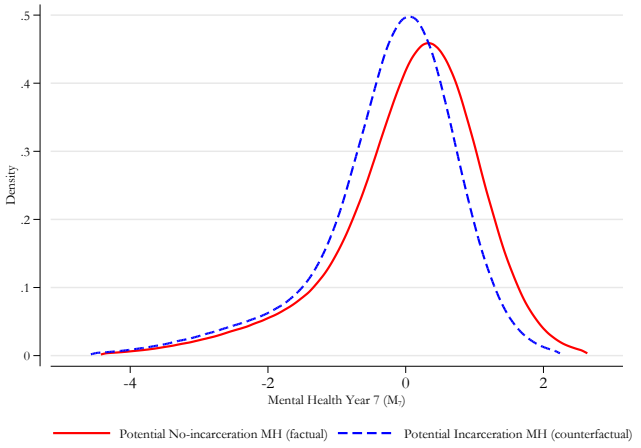
Table OA.1.9: Mean Treatment Effects

	Mean (1)	lb (2)	ub (3)	P($\cdot > 0$) (4)	P($\lvert \cdot \rvert < 0.01$) (5)	P($\lvert \cdot \rvert < 1\%$) (6)
ATE$(M_7, 1, 0)$	-0.246	-0.480	-0.032	0.011	0.006	0.006
ATT$(M_7, 1, 1, 0)$	-0.231	-0.456	-0.018	0.015	0.010	0.008
ATT$(M_7, 0, 1, 0)$	-0.268	-0.495	-0.045	0.009	0.004	0.004
MTE$(M_7, 1, 1, 0)$	-0.252	-0.496	-0.021	0.015	0.008	0.008

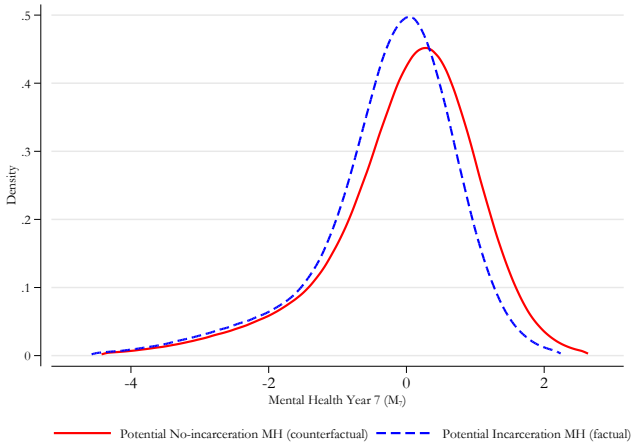
Notes: Let Θ_s be the collection of parameters evaluated using the s-th parameter draw of the estimated posterior of the model. Column (1) presents $\frac{1}{S}\sum_{s=1}^{S} ATE(M_7, 1, 0; \Theta_s)$. Columns (2) and (3) present the 2.5 and 97.5 percentiles of $\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S}$. Columns (4), (5), and (6) are $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[ATE(M_7, 1, 0; \Theta_s) > 0]$, $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01]$, and $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01 \times Range(\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S})]$, respectively. Similar definitions apply to the remaining parameters. Results correspond to the binary treatment model with mixtures 4-2.

Figure OA.1.7: Factual and Counterfactual Distributions of Mental Health and Mental Health Gain (Year 7) Conditional on Incarceration Status

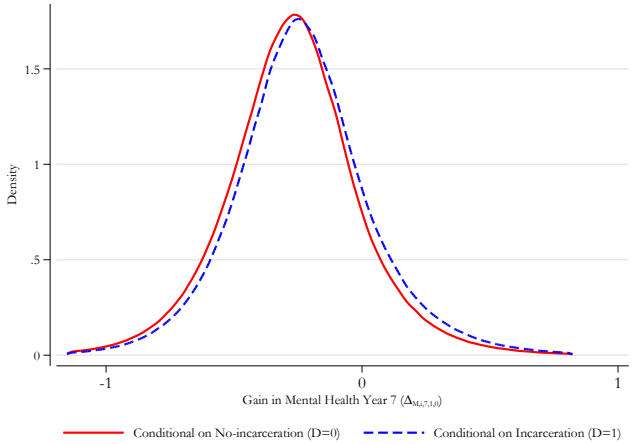
(a) M_7 Conditional on $D = 0$



(b) M_7 Conditional on $D = 1$



(c) $\Delta_{M_i,7,1,0}$ by Incarceration Status (D)

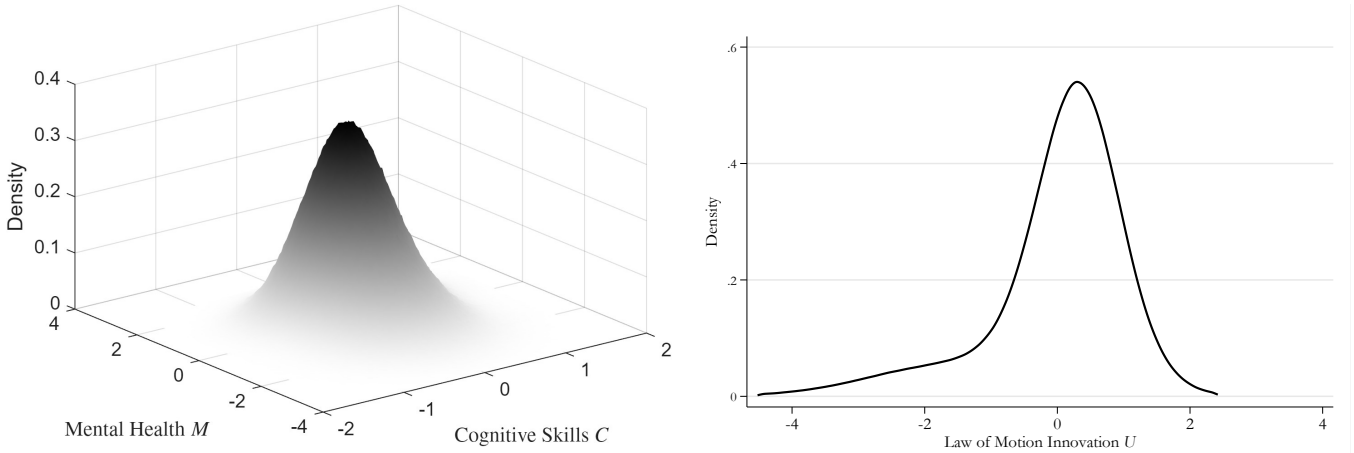


Notes: Figure OA.1.7a displays the factual density $f(M_{i,7,0}|D_i = 0)$ and counterfactual density $f(M_{i,7,1}|D_i = 0)$, for people who do not go to prison. Figure OA.1.7b displays the factual density $f(M_{i,7,1}|D_i = 1)$ and counterfactual density $f(M_{i,7,0}|D_i = 1)$ for people who go to prison. Figure OA.1.7c displays the distribution of mental health gain from imprisonment, $\Delta_{M_i,7,1,0}$, conditional on no imprisonment (red solid line) and imprisonment (blue dashed line). Results correspond to the binary treatment model with mixtures 4-2.

OA.1.2 Eight-Category Treatment - Model Results

Figure OA.1.8: Estimated Posterior Distributions of the Factors

(a) Joint Distribution of C and M (b) Distribution of U



Notes: This figure reports the estimated posterior distributions of the factors. Figure OA.1.8a displays the estimated joint distribution of cognitive skills and mental health at baseline, while Figure OA.1.8b displays the estimated density of the law of motion innovation to mental health in year 7. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Table OA.1.10: Estimated Parameters from Factor Model — Mixtures

	(1) Cognitive Factor C	(2) Mental Health Baseline M	(3) U
Mixture 1			
Mean	-0.043	-0.450	0.327
Variance	0.568	1.029	0.395
Probability	0.248	0.248	0.798
Mixture 2			
Mean	-0.028	-0.534	-1.565
Variance	0.510	1.006	1.531
Probability	0.260	0.260	0.202
Mixture 3			
Mean	-0.038	-0.520	
Variance	0.595	1.087	
Probability	0.218	0.218	
Mixture 4			
Mean	-0.008	-0.416	
Variance	0.518	0.955	
Probability	0.273	0.273	

Notes: This table reports the parameter estimates from the mixture model for the joint distribution of baseline latent skills (i.e., cognitive skills and mental health) and a subsequent shock (U_i) to the mental health production function. We report the mean of each parameter across 5,000 draws from the posterior distribution. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Table OA.1.11: Estimated Parameters from Factor Model — Cognitive Skills Measures

	(1) WASI IQ	(2) Stroop Color	(3) Stroop Word	(4) Stroop Color/Word	(5) Trail Making Part A	(6) Trail Making Part B
Constant	-0.207	-0.296	-0.193	-0.479	0.845	1.149
	(0.200)	(0.215)	(0.211)	(0.211)	(0.256)	(0.281)
Age 15	-0.023	0.211	0.132	0.292	0.694	0.399
	(0.130)	(0.136)	(0.137)	(0.134)	(0.163)	(0.173)
Age 16	-0.088	0.303	0.261	0.395	0.805	0.416
	(0.119)	(0.125)	(0.124)	(0.124)	(0.151)	(0.159)
Age 17	0.067	0.375	0.190	0.459	0.749	0.526
	(0.122)	(0.129)	(0.129)	(0.127)	(0.157)	(0.166)
Age 18	0.091	0.083	0.226	0.143	0.895	0.508
	(0.170)	(0.182)	(0.180)	(0.179)	(0.217)	(0.229)
Female	-0.144	0.130	0.217	0.068	0.244	0.136
	(0.098)	(0.105)	(0.103)	(0.102)	(0.128)	(0.135)
White	0.489	0.147	0.107	0.269	0.451	0.279
	(0.192)	(0.204)	(0.202)	(0.201)	(0.241)	(0.257)
Hispanic	-0.229	-0.074	-0.162	-0.023	0.069	-0.150
	(0.187)	(0.199)	(0.197)	(0.197)	(0.233)	(0.249)
Black	-0.076	-0.032	-0.232	-0.059	-0.139	-0.215
	(0.186)	(0.199)	(0.198)	(0.198)	(0.235)	(0.252)
Phoenix	0.520	0.047	0.210	0.251	0.375	0.391
	(0.093)	(0.101)	(0.098)	(0.097)	(0.122)	(0.131)
Variance	0.682	0.383	0.532	0.581	1.000	1.000
	(0.040)	(0.039)	(0.038)	(0.038)	(0.000)	(0.000)
Cognitive Factor	1.000	1.683	1.419	1.299	0.954	1.385
	(0.000)	(0.134)	(0.119)	(0.114)	(0.135)	(0.160)
Cutoff 1					0.000	0.000
					(0.000)	(0.000)
Cutoff 2					0.802	1.202
					(0.080)	(0.094)
Cutoff 3					2.075	2.102
					(0.100)	(0.114)

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the cognitive measure system. Standard errors are shown in parentheses below the mean point estimates. IQ and the Stroop components are modeled using a linear-in-parameters specification, while the Trail-Making tests are estimated using an ordered threshold model. The cognitive skills factor is normalized to have a loading of one on the WASI IQ score. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Table OA.1.12: Estimated Parameters from Factor Model — Mental Health Measures (Baseline)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	0.509	0.717	0.855	0.773	1.326	0.279	-0.302	1.695	0.906
	(0.351)	(0.371)	(0.391)	(0.327)	(0.422)	(0.326)	(0.342)	(0.358)	(0.369)
Age 15	0.168	0.123	-0.123	0.282	0.049	0.000	-0.043	0.127	0.100
	(0.229)	(0.240)	(0.255)	(0.215)	(0.267)	(0.214)	(0.225)	(0.225)	(0.239)
Age 16	-0.091	0.154	-0.063	0.237	0.209	-0.059	-0.084	0.027	0.158
	(0.213)	(0.220)	(0.233)	(0.198)	(0.243)	(0.197)	(0.207)	(0.208)	(0.218)
Age 17	0.160	0.492	0.429	0.527	0.559	0.168	0.221	0.469	0.330
	(0.221)	(0.233)	(0.244)	(0.207)	(0.255)	(0.206)	(0.213)	(0.220)	(0.231)
Age 18	0.238	0.943	0.361	0.403	0.858	0.117	0.134	1.189	1.028
	(0.300)	(0.321)	(0.331)	(0.285)	(0.356)	(0.285)	(0.298)	(0.312)	(0.312)
Female	0.532	0.447	0.328	0.613	0.389	0.449	0.196	0.273	0.312
	(0.182)	(0.197)	(0.205)	(0.176)	(0.221)	(0.165)	(0.169)	(0.184)	(0.190)
White	0.165	-0.138	-0.204	0.092	-0.138	-0.198	-0.253	-0.407	-0.664
	(0.328)	(0.352)	(0.364)	(0.307)	(0.389)	(0.302)	(0.324)	(0.331)	(0.346)
Hispanic	0.087	0.080	0.057	-0.106	-0.153	-0.082	0.036	-0.219	-0.270
	(0.320)	(0.344)	(0.356)	(0.301)	(0.382)	(0.294)	(0.314)	(0.324)	(0.337)
Black	-0.457	-0.419	-0.640	-0.198	-0.580	-0.281	-0.207	-0.252	-0.586
	(0.326)	(0.350)	(0.364)	(0.305)	(0.389)	(0.300)	(0.321)	(0.327)	(0.342)
Phoenix	-0.226	-0.152	-0.072	-0.078	0.264	0.093	-0.048	-0.322	-0.080
	(0.165)	(0.183)	(0.188)	(0.163)	(0.199)	(0.159)	(0.161)	(0.170)	(0.175)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-1.088	-1.236	-1.301	-1.000	-1.450	-0.893	-0.859	-1.110	-1.192
	(0.096)	(0.105)	(0.113)	(0.000)	(0.125)	(0.082)	(0.082)	(0.103)	(0.105)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.639	0.735	0.723	0.567	0.550	0.715	0.616	0.623	0.671
	(0.058)	(0.062)	(0.065)	(0.057)	(0.060)	(0.059)	(0.058)	(0.070)	(0.062)
Cutoff 3	1.015	1.262	1.246	1.049	1.180	1.142	0.959	1.091	1.145
	(0.071)	(0.077)	(0.083)	(0.070)	(0.082)	(0.074)	(0.074)	(0.090)	(0.081)
Cutoff 4	1.451	1.654	1.703	1.449	1.605	1.654	1.274	1.582	1.710
	(0.085)	(0.088)	(0.099)	(0.081)	(0.094)	(0.093)	(0.089)	(0.105)	(0.101)
Cutoff 5	1.783	1.975	2.111	1.899	2.088		1.534	2.087	2.267
	(0.097)	(0.098)	(0.118)	(0.092)	(0.110)		(0.102)	(0.116)	(0.121)
Cutoff 6	2.083	2.261	2.446		2.509				
	(0.106)	(0.110)	(0.135)		(0.125)				
Cutoff 7	2.366								
	(0.117)								

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

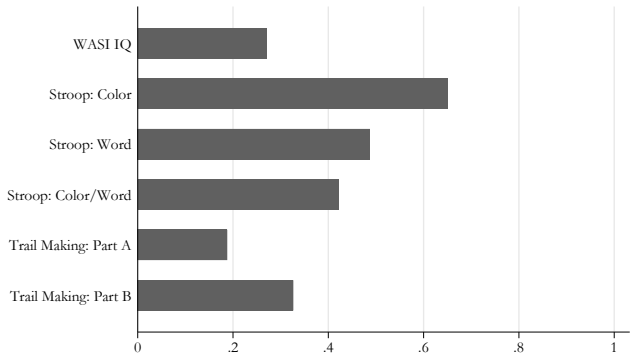
Table OA.1.13: Estimated Parameters from Factor Model — Mental Health Measures (Year 7)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	-0.173	0.575	0.043	-0.005	1.155	-1.002	-1.033	0.794	-0.613
	(0.541)	(0.813)	(0.743)	(0.570)	(0.674)	(0.654)	(0.575)	(0.618)	(0.756)
Age 15	0.103	0.659	0.214	-0.112	0.245	0.323	0.456	0.604	0.741
	(0.286)	(0.404)	(0.352)	(0.283)	(0.318)	(0.344)	(0.316)	(0.305)	(0.382)
Age 16	-0.028	0.256	-0.428	-0.119	0.074	0.254	0.191	0.182	0.383
	(0.263)	(0.378)	(0.331)	(0.267)	(0.294)	(0.323)	(0.307)	(0.281)	(0.353)
Age 17	0.081	0.542	0.341	0.400	0.501	0.122	0.624	0.201	0.304
	(0.264)	(0.381)	(0.337)	(0.270)	(0.302)	(0.331)	(0.304)	(0.286)	(0.360)
Age 18	-0.085	-0.132	-0.486	0.384	0.256	-0.523	0.363	0.352	0.381
	(0.382)	(0.544)	(0.495)	(0.380)	(0.432)	(0.509)	(0.421)	(0.405)	(0.486)
Female	0.384	0.068	0.128	0.463	0.476	0.386	-0.087	-0.143	-0.162
	(0.208)	(0.300)	(0.266)	(0.213)	(0.244)	(0.250)	(0.238)	(0.232)	(0.283)
White	0.090	-1.169	0.099	0.795	-0.553	0.278	0.047	-0.377	-0.567
	(0.523)	(0.791)	(0.730)	(0.552)	(0.651)	(0.626)	(0.555)	(0.604)	(0.737)
Hispanic	-0.283	-0.983	-0.143	0.436	-0.891	-0.334	0.312	-0.451	0.015
	(0.517)	(0.781)	(0.719)	(0.548)	(0.638)	(0.628)	(0.543)	(0.593)	(0.726)
Black	-0.099	-0.938	0.218	0.632	-0.948	0.180	-0.018	-0.053	0.198
	(0.511)	(0.770)	(0.720)	(0.548)	(0.640)	(0.619)	(0.539)	(0.589)	(0.724)
Phoenix	-0.264	-0.421	0.055	-0.441	-0.040	-0.089	-0.577	-0.307	-0.259
	(0.221)	(0.304)	(0.268)	(0.219)	(0.243)	(0.262)	(0.235)	(0.237)	(0.290)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-0.840	-1.600	-1.435	-1.000	-1.239	-1.074	-0.837	-1.129	-1.407
	(0.103)	(0.213)	(0.169)	(0.000)	(0.151)	(0.134)	(0.108)	(0.140)	(0.178)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.522	0.439	0.669	0.826	0.609	0.550	0.500	0.659	0.643
	(0.073)	(0.081)	(0.092)	(0.092)	(0.081)	(0.088)	(0.083)	(0.089)	(0.100)
Cutoff 3	0.856	0.949	1.332	1.424	0.994	1.026	0.981	1.064	1.245
	(0.094)	(0.127)	(0.141)	(0.114)	(0.102)	(0.123)	(0.121)	(0.110)	(0.146)
Cutoff 4	1.238	1.572	1.861	1.833	1.556	1.333	1.518	1.435	1.682
	(0.121)	(0.180)	(0.180)	(0.126)	(0.131)	(0.146)	(0.163)	(0.126)	(0.171)
Cutoff 5	1.478	1.812	2.429	2.223	1.928		1.742	1.774	2.175
	(0.135)	(0.198)	(0.223)	(0.141)	(0.153)		(0.184)	(0.137)	(0.205)
Cutoff 6	1.611	2.178	2.695		2.126				
	(0.146)	(0.229)	(0.244)		(0.165)				
Cutoff 7	1.786								
	(0.156)								

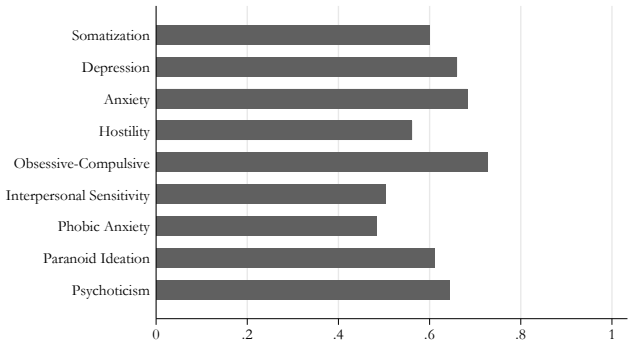
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system in year 7. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.9: Fraction of the Variance Explained by the Factor

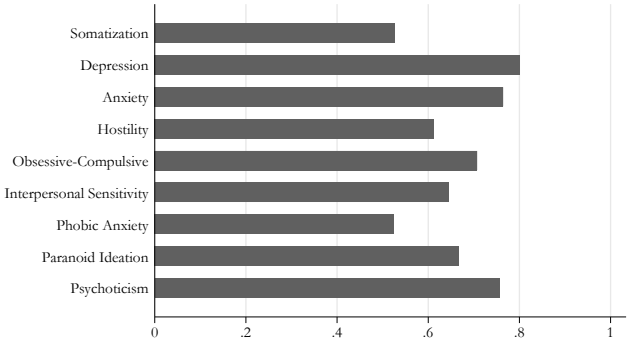
(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



(c) Mental Health (Year 7): M_7



Notes: These figures present the average fraction of the variance of each cognitive and mental health measure explained by the cognitive and mental health factors. For example, Figure OA.1.9a shows that 27.0% of the fraction of the variance of the residualized (against X) WASI IQ measure is explained by the cognitive skills C. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

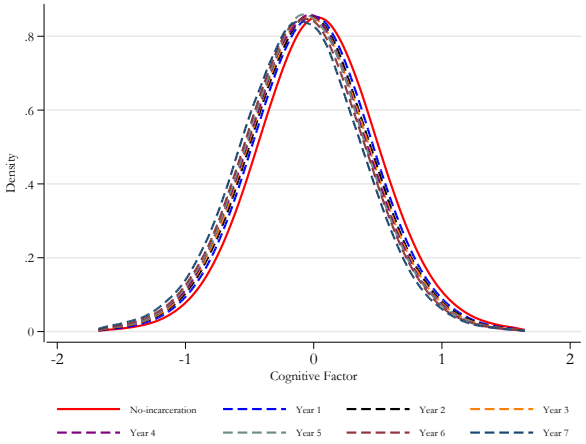
Table OA.1.14: Estimated Parameters from Model - Treatment Equation

	Mean	SD
Constant	-0.075	0.267
Age 15	0.255	0.170
Age 16	0.332	0.156
Age 17	0.129	0.164
Age 18	0.024	0.220
Female	-1.121	0.153
White	0.149	0.257
Hispanic	0.139	0.249
Black	0.413	0.251
Phoenix	-0.021	0.121
Cognitive Skills (ψ_T)	-0.199	0.118
Mental Health (μ_T)	-0.048	0.048
Cutoff 1 (κ_1)	0.000	0.000
Cutoff 2 (κ_2)	0.890	0.059
Cutoff 3 (κ_3)	1.209	0.069
Cutoff 4 (κ_4)	1.557	0.083
Cutoff 5 (κ_5)	1.755	0.092
Cutoff 6 (κ_6)	1.954	0.101
Cutoff 7 (κ_7)	2.401	0.137

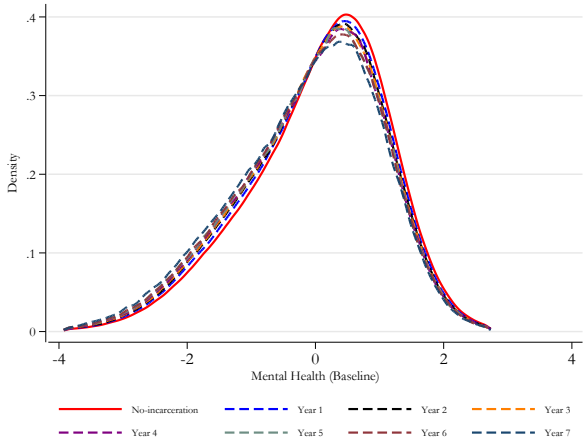
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the treatment equation. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.10: Distribution of Cognitive and Baseline Mental Health Factors By Treatment

(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



Notes: These figures show the estimated densities of the cognitive and baseline mental health factors conditional on incarceration status. For example, in Figure OA.1.10a, the red solid-line plots $f(C \mid D = 0)$ and the blue dashed-line plots $f(C \mid D = 1)$. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Table OA.1.15: Estimated Parameters from Model - Law of Motion Equation

	Mean	SD
Year 1 Incarceration (δ_{T1})	-0.234	0.135
Year 2 Incarceration (δ_{T2})	-0.111	0.245
Year 3 Incarceration (δ_{T3})	-0.480	0.247
Year 4 Incarceration (δ_{T4})	0.032	0.341
Year 5 Incarceration (δ_{T5})	-0.335	0.479
Year 6 Incarceration (δ_{T6})	-0.405	0.331
Year 7 Incarceration (δ_{T7})	-0.711	0.394
Cognitive Skills (λ_C)	-0.249	0.174
Mental Health (λ_M)	0.414	0.090
Year 1 Incarc. $\times$ Cognitive Skills (λ_{CT1})	0.185	0.308
Year 2 Incarc. $\times$ Cognitive Skills (λ_{CT2})	0.651	0.893
Year 3 Incarc. $\times$ Cognitive Skills (λ_{CT3})	-1.323	0.527
Year 4 Incarc. $\times$ Cognitive Skills (λ_{CT4})	0.912	0.957
Year 5 Incarc. $\times$ Cognitive Skills (λ_{CT5})	-2.166	1.494
Year 6 Incarc. $\times$ Cognitive Skills (λ_{CT6})	0.388	0.864
Year 7 Incarc. $\times$ Cognitive Skills (λ_{CT7})	0.228	1.059
Year 1 Incarc. $\times$ Mental Health (λ_{MT1})	-0.197	0.122
Year 2 Incarc. $\times$ Mental Health (λ_{MT2})	-0.312	0.302
Year 3 Incarc. $\times$ Mental Health (λ_{MT3})	0.155	0.257
Year 4 Incarc. $\times$ Mental Health (λ_{MT4})	-0.074	0.547
Year 5 Incarc. $\times$ Mental Health (λ_{MT5})	-0.587	0.505
Year 6 Incarc. $\times$ Mental Health (λ_{MT6})	0.226	0.347
Year 7 Incarc. $\times$ Mental Health (λ_{MT7})	-0.331	0.689

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the law of motion of mental health in the ordered model with eight treatment categories and mixtures 4-2.

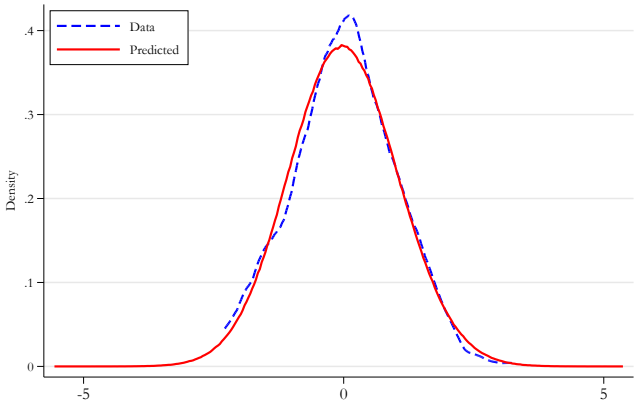
Table OA.1.16: Estimated Parameters from Multinomial Model for Missing Data

	Treatment and MH Year 7 Measures Included	Treatment Included, MH Year 7 Measures Missing
Constant	0.140	-0.102
	(0.310)	(0.303)
Age 15	-0.116	0.136
	(0.197)	(0.204)
Age 16	-0.237	0.224
	(0.184)	(0.191)
Age 17	-0.081	-0.108
	(0.187)	(0.200)
Age 18	-0.141	0.066
	(0.251)	(0.271)
Female	0.151	-0.252
	(0.148)	(0.164)
White	0.355	-0.210
	(0.300)	(0.286)
Hispanic	0.283	-0.224
	(0.294)	(0.277)
Black	0.281	-0.129
	(0.295)	(0.282)
Phoenix	-0.274	0.459
	(0.145)	(0.149)
Cognitive Skills Factor C	0.140	0.170
	(0.133)	(0.144)
Mental Health Factor M	-0.212	0.138
	(0.058)	(0.063)

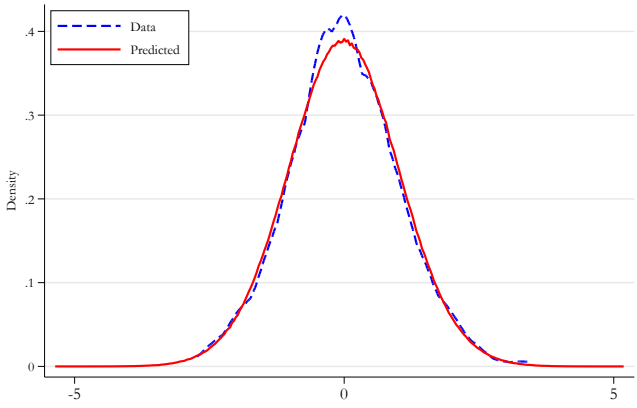
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the attrition equation. The reference category corresponds to observations with both treatment and mental health measures missing in year 7. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.11: Posterior Predictive Fit - Cognitive Measures

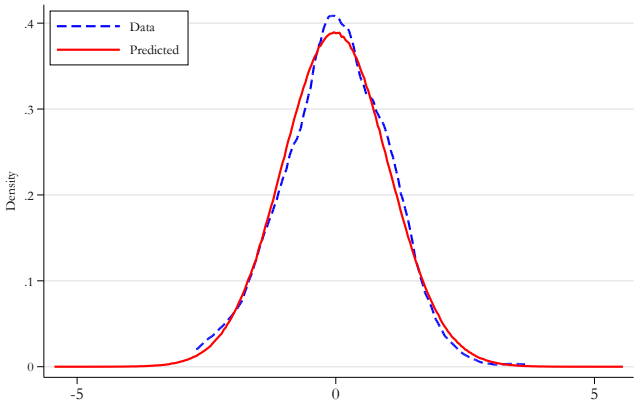
(a) WASI IQ



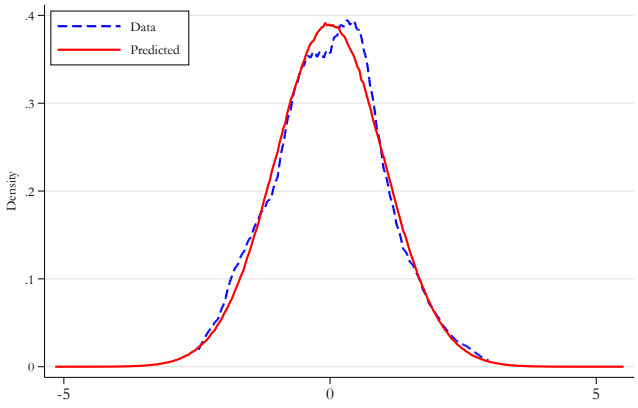
(b) Stroop Color



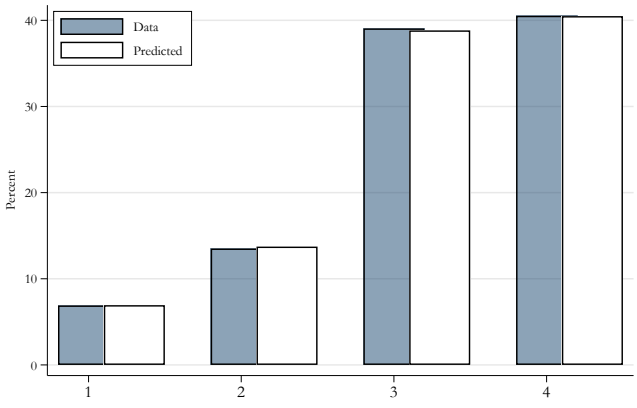
(c) Stroop Word



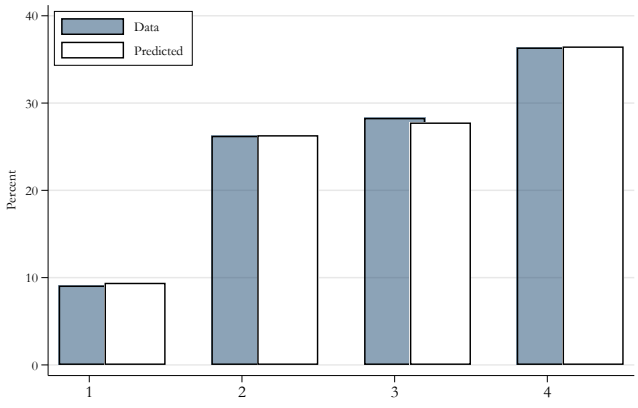
(d) Stroop Color/Word



(e) Trail Making A

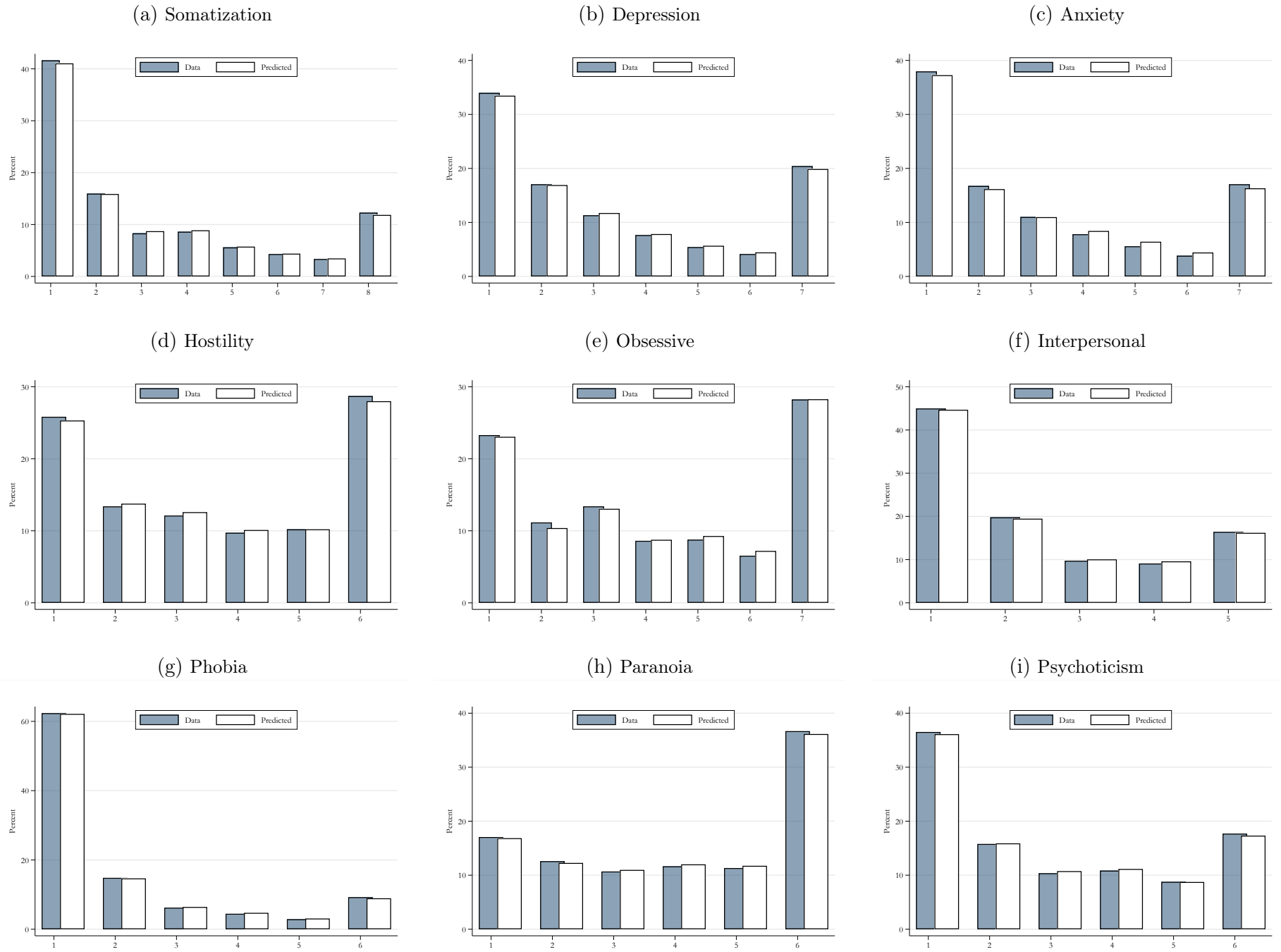


(f) Trail Making B



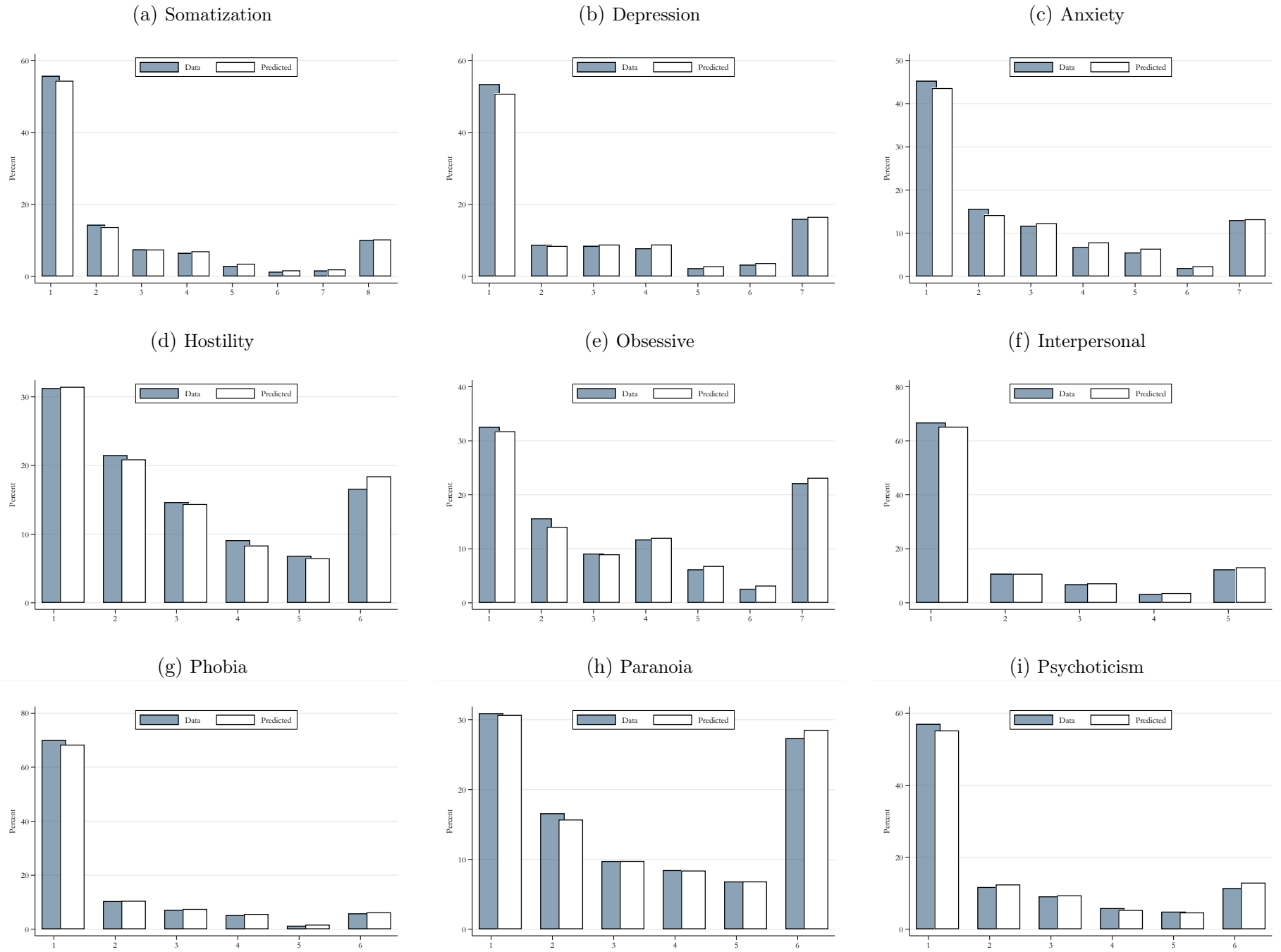
Notes: This figure compares observed and predicted distributions for each cognitive measure. For continuous variables (Figures OA.1.11a-OA.1.11d), observed distributions appear as blue dashed lines and predicted as solid red lines. For discrete variables (Figures OA.1.11e and OA.1.11f), observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.12: Posterior Predictive Fit - BSI Mental Health Measures (Baseline)



Notes: This figure compares observed and predicted distributions for each baseline mental health measure. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.13: Posterior Predictive Fit - BSI Mental Health Measures (Year 7)



Notes: This figure compares observed and predicted distributions for each mental health measure eight years after the baseline. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

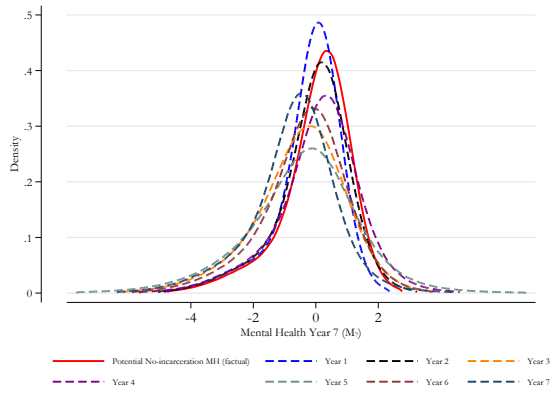
Table OA.1.17: Goodness Of Fit

	Chi-Square Classical p-value (1)	Mahalanobis Bayesian PPP (2)	Euclidean Bayesian PPP (3)
Cognitive Skills: C			
WASI IQ†	0.046	0.365	0.485
Stroop Color†	0.955	0.589	0.669
Stroop Word†	0.660	0.657	0.732
Stroop CW†	0.044	0.705	0.767
Trail Making A	0.998	0.349	0.396
Trail Making B	0.991	0.309	0.290
Mental Health: M			
BSI Somatization	1.000	0.089	0.181
BSI Depresion	0.999	0.433	0.305
BSI Anxiety	0.959	0.661	0.626
BSI Hostility	0.997	0.532	0.513
BSI Obsessive	0.991	0.229	0.680
BSI Interpersonal	0.996	0.351	0.569
BSI Phobia	0.999	0.343	0.110
BSI Paranoia	0.998	0.526	0.485
BSI Psychoticism	1.000	0.720	0.400
Mental Health: M_7			
BSI Somatization	0.997	0.987	0.938
BSI Depresion	0.981	0.335	0.352
BSI Anxiety	0.968	0.025	0.052
BSI Hostility	0.982	0.782	0.795
BSI Obsessive	0.988	0.610	0.849
BSI Interpersonal	0.989	0.550	0.427
BSI Phobia	0.993	0.481	0.355
BSI Paranoia	0.998	0.458	0.674
BSI Psychoticism	0.970	0.532	0.576

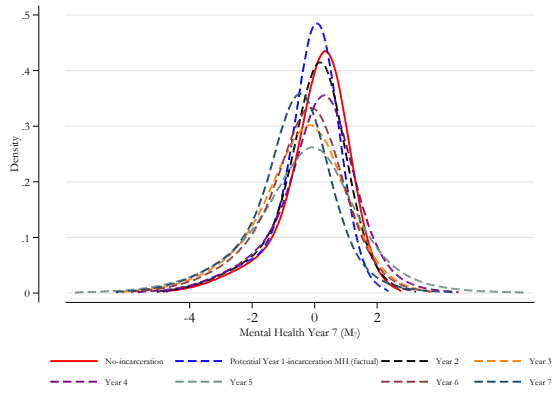
Notes: This table presents three measures of Goodness of Fit for our cognitive and mental health measures. Column (1) displays the *Classical* p-values from the chi-square distribution. Columns (2) and (3) report *Bayesian* Posterior Predictive P-values using the Mahalanobis distance and the joint Euclidean distance of means and standard deviations, respectively. For discrete variables, the total number of bins is the number of categories in the empirical distribution. For continuous variables (indexed by a dag), we set the number of bins equal to 4 and the cutoffs equal to the quartiles of the simulated distribution. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.14: Factual and Counterfactual Distributions of Mental Health (Year 7) Conditional on Incarceration Status

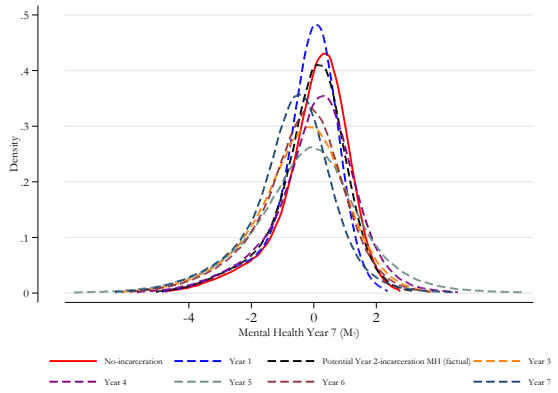
(a) M_7 Conditional on $D = 0$



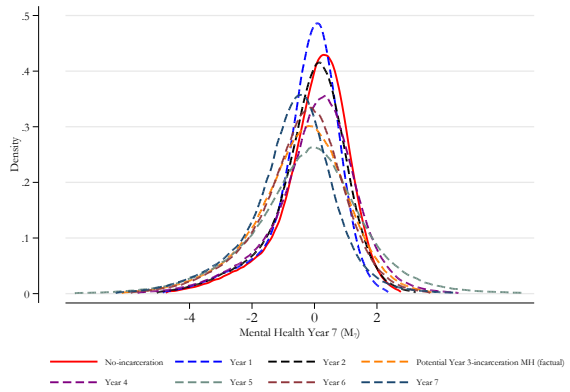
(b) M_7 Conditional on $D = 1$



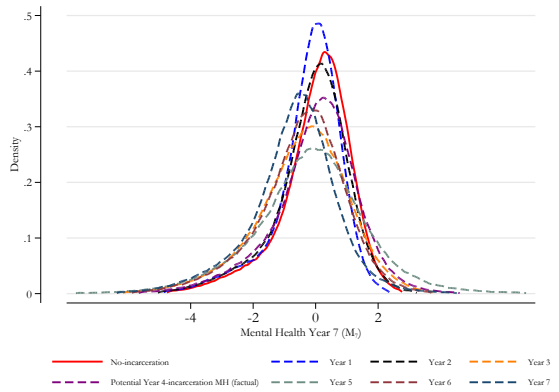
(c) M_7 Conditional on $D = 2$



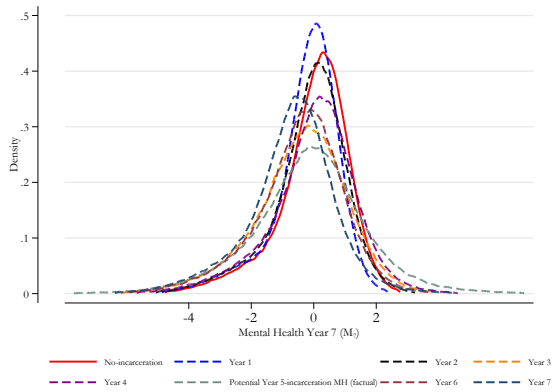
(d) M_7 Conditional on $D = 3$



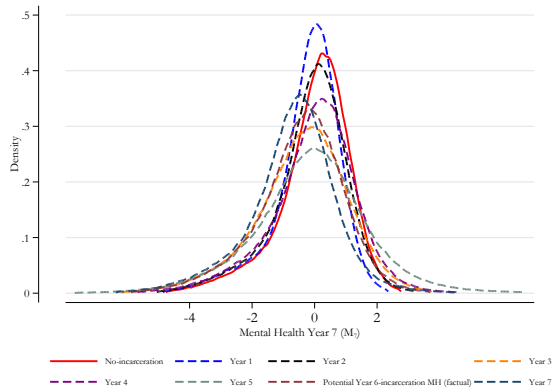
(e) M_7 Conditional on $D = 4$



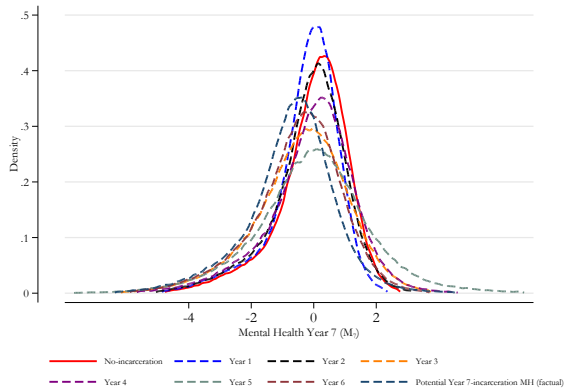
(f) M_7 Conditional on $D = 5$



(g) M_7 Conditional on $D = 6$



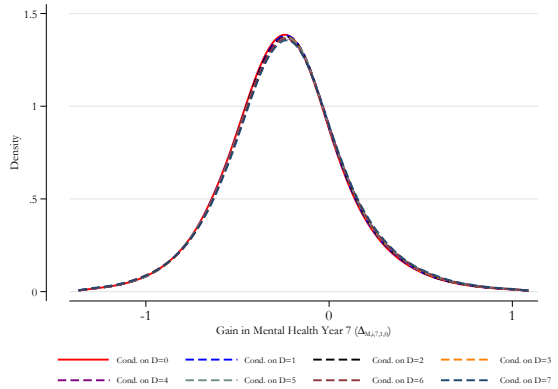
(h) M_7 Conditional on $D = 7$



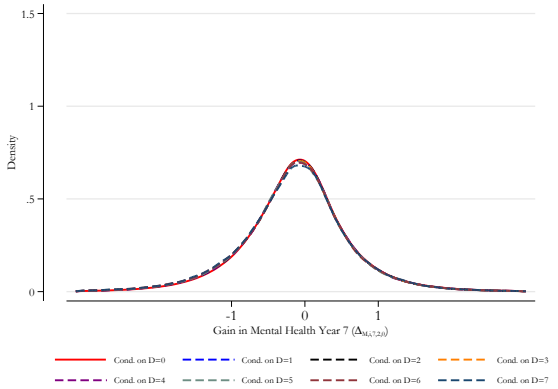
Notes: Figure OA.1.14a displays the factual density $f(M_{i,7,0}|D_i = 0)$ and counterfactual densities $f(M_{i,7,j}|D_i = 0)$, $j \in \{1, \ldots, 7\}$, for people who do not go to prison. Figure OA.1.14b displays the factual density $f(M_{i,7,1}|D_i = 1)$ and counterfactual densities $f(M_{i,7,j}|D_i = 1)$, $j \in \{0, 2, \ldots, 7\}$, for people who go to prison in Year 1. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

Figure OA.1.15: Counterfactual Distributions of Mental Health Gain (Year 7) by Incarceration Status

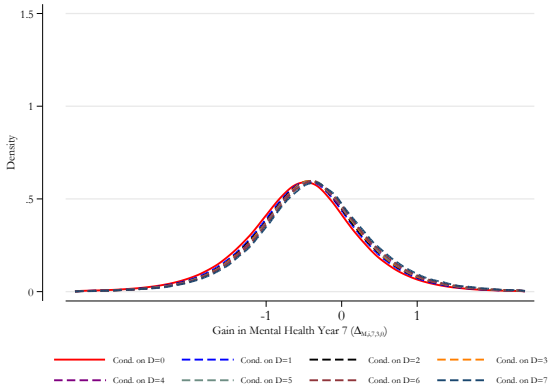
(a) $\Delta_{M_i,7,1,0}$ by Incarceration Status (D)



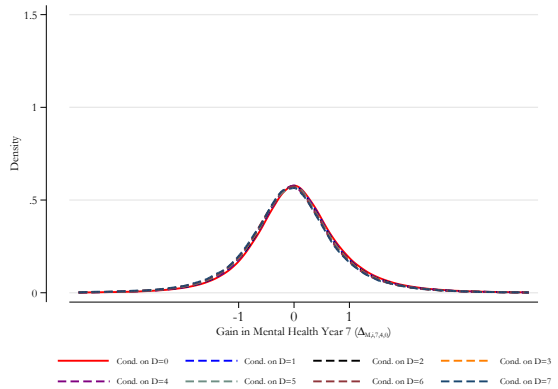
(b) $\Delta_{M_i,7,2,0}$ by Incarceration Status (D)



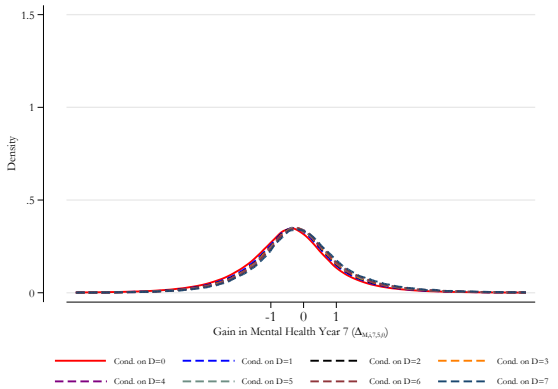
(c) $\Delta_{M_i,7,3,0}$ by Incarceration Status (D)



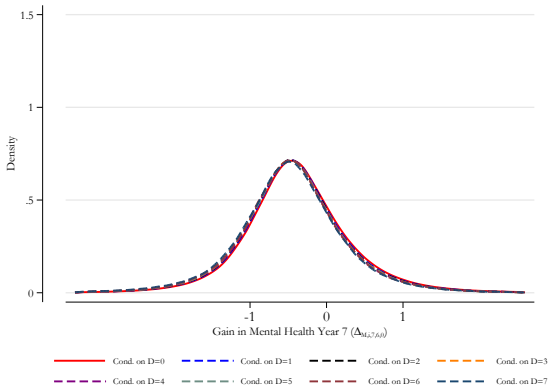
(d) $\Delta_{M_i,7,4,0}$ by Incarceration Status (D)



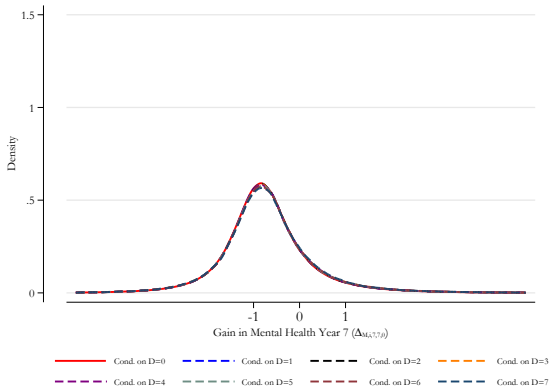
(e) $\Delta_{M_i,7,5,0}$ by Incarceration Status (D)



(f) $\Delta_{M_i,7,6,0}$ by Incarceration Status (D)



(g) $\Delta_{M_i,7,7,0}$ by Incarceration Status (D)

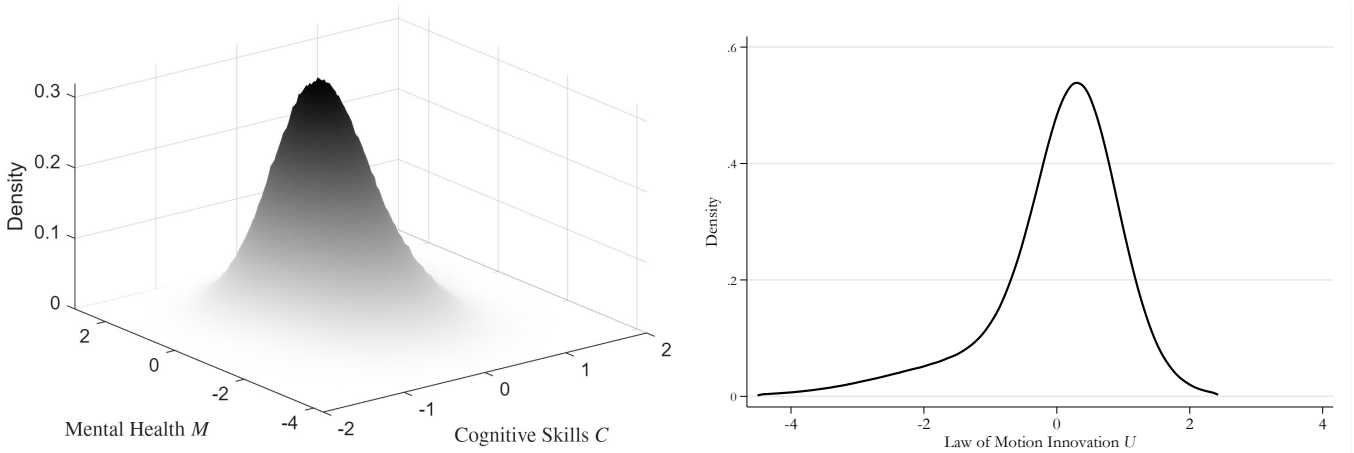


Notes: Figure OA.1.15a displays the distribution of mental health gain from imprisonment in Year 1, $\Delta_{M_i,7,1,0}$, by imprisonment status ($f(\Delta_{M_i,7,1,0}|D_i = j)$), for $j \in \{0, \ldots, 7\}$. Results correspond to the ordered model with eight treatment categories and mixtures 4-2.

OA.1.3 Increased Mixture Specification - Model Results

Figure OA.1.16: Estimated Posterior Distributions of the Factors

(a) Joint Distribution of C and M (b) Distribution of U



Notes: This figure reports the estimated posterior distributions of the factors. Figure OA.1.16a displays the estimated joint distribution of cognitive skills and mental health at baseline, while Figure OA.1.16b displays the estimated density of the law of motion innovation to mental health in year 7. Results correspond to the model with mixtures 6-3.

Table OA.1.18: Estimated Parameters from Factor Model — Mixtures

	(1) Cognitive Factor C	(2) Mental Health Baseline M	(3) U
Mixture 1			
Mean	-0.018	-0.464	-0.613
Variance	0.705	1.119	0.974
Probability	0.158	0.158	0.300
Mixture 2			
Mean	-0.017	-0.410	-0.594
Variance	0.618	1.080	0.949
Probability	0.187	0.187	0.319
Mixture 3			
Mean	-0.034	-0.502	-0.404
Variance	0.687	1.116	0.849
Probability	0.159	0.159	0.381
Mixture 4			
Mean	-0.030	-0.481	
Variance	0.712	1.131	
Probability	0.163	0.163	
Mixture 5			
Mean	-0.036	-0.511	
Variance	0.692	1.128	
Probability	0.154	0.154	
Mixture 6			
Mean	-0.004	-0.488	
Variance	0.637	1.113	
Probability	0.179	0.179	

Notes: This table reports the parameter estimates from the mixture model for the joint distribution of baseline latent skills (i.e., cognitive skills and mental health) and a subsequent shock (U_i) to the mental health production function. We report the mean of each parameter across 5,000 draws from the posterior distribution. Results correspond to the model with mixtures 6-3.

Table OA.1.19: Estimated Parameters from Factor Model — Cognitive Skills Measures

	(1) WASI IQ	(2) Stroop Color	(3) Stroop Word	(4) Stroop Color/Word	(5) Trail Making Part A	(6) Trail Making Part B
Constant	-0.220	-0.324	-0.219	-0.502	0.822	1.124
	(0.199)	(0.214)	(0.214)	(0.210)	(0.258)	(0.277)
Age 15	-0.013	0.229	0.149	0.306	0.700	0.416
	(0.130)	(0.139)	(0.137)	(0.134)	(0.166)	(0.177)
Age 16	-0.082	0.311	0.270	0.404	0.804	0.424
	(0.119)	(0.128)	(0.126)	(0.123)	(0.152)	(0.162)
Age 17	0.074	0.384	0.200	0.470	0.753	0.532
	(0.123)	(0.130)	(0.128)	(0.129)	(0.154)	(0.168)
Age 18	0.105	0.099	0.239	0.153	0.899	0.522
	(0.171)	(0.182)	(0.180)	(0.178)	(0.219)	(0.231)
Female	-0.147	0.127	0.212	0.063	0.238	0.137
	(0.100)	(0.106)	(0.105)	(0.100)	(0.126)	(0.135)
White	0.490	0.156	0.116	0.276	0.454	0.294
	(0.194)	(0.200)	(0.202)	(0.197)	(0.242)	(0.258)
Hispanic	-0.225	-0.060	-0.150	-0.011	0.076	-0.132
	(0.188)	(0.198)	(0.196)	(0.195)	(0.232)	(0.248)
Black	-0.064	-0.005	-0.207	-0.040	-0.122	-0.190
	(0.189)	(0.198)	(0.198)	(0.195)	(0.235)	(0.255)
Phoenix	0.522	0.053	0.215	0.254	0.377	0.397
	(0.093)	(0.098)	(0.100)	(0.096)	(0.124)	(0.131)
Variance	0.680	0.386	0.533	0.580	1.000	1.000
	(0.041)	(0.038)	(0.037)	(0.038)	(0.000)	(0.000)
Cognitive Factor	1.000	1.588	1.342	1.234	0.903	1.318
	(0.000)	(0.121)	(0.109)	(0.103)	(0.125)	(0.150)
Cutoff 1					0.000	0.000
					(0.000)	(0.000)
Cutoff 2					0.799	1.209
					(0.076)	(0.083)
Cutoff 3					2.058	2.109
					(0.096)	(0.097)

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the cognitive measure system. Standard errors are shown in parentheses below the mean point estimates. IQ and the Stroop components are modeled using a linear-in-parameters specification, while the Trail-Making tests are estimated using an ordered threshold model. The cognitive skills factor is normalized to have a loading of one on the WASI IQ score. Results correspond to the model with mixtures 6-3.

Table OA.1.20: Estimated Parameters from Factor Model — Mental Health Measures (Baseline)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	0.546	0.737	0.861	0.791	1.341	0.297	-0.291	1.692	0.932
	(0.342)	(0.365)	(0.393)	(0.329)	(0.414)	(0.321)	(0.331)	(0.353)	(0.375)
Age 15	0.173	0.123	-0.122	0.281	0.048	-0.003	-0.046	0.129	0.092
	(0.227)	(0.238)	(0.250)	(0.215)	(0.263)	(0.211)	(0.220)	(0.227)	(0.236)
Age 16	-0.091	0.158	-0.058	0.236	0.215	-0.060	-0.082	0.030	0.154
	(0.211)	(0.220)	(0.229)	(0.197)	(0.239)	(0.195)	(0.203)	(0.206)	(0.215)
Age 17	0.144	0.470	0.404	0.510	0.534	0.149	0.206	0.453	0.302
	(0.222)	(0.232)	(0.242)	(0.212)	(0.254)	(0.206)	(0.207)	(0.216)	(0.226)
Age 18	0.262	0.956	0.375	0.423	0.882	0.134	0.148	1.207	1.041
	(0.296)	(0.318)	(0.324)	(0.283)	(0.346)	(0.280)	(0.287)	(0.312)	(0.309)
Female	0.533	0.434	0.314	0.614	0.376	0.444	0.189	0.264	0.304
	(0.178)	(0.195)	(0.198)	(0.176)	(0.213)	(0.165)	(0.165)	(0.183)	(0.190)
White	0.136	-0.169	-0.230	0.076	-0.157	-0.220	-0.275	-0.416	-0.689
	(0.321)	(0.343)	(0.361)	(0.306)	(0.378)	(0.298)	(0.315)	(0.326)	(0.342)
Hispanic	0.066	0.058	0.042	-0.116	-0.166	-0.095	0.021	-0.219	-0.286
	(0.316)	(0.337)	(0.355)	(0.301)	(0.376)	(0.293)	(0.304)	(0.321)	(0.339)
Black	-0.471	-0.425	-0.639	-0.194	-0.577	-0.282	-0.207	-0.243	-0.589
	(0.318)	(0.339)	(0.362)	(0.303)	(0.378)	(0.298)	(0.313)	(0.324)	(0.344)
Phoenix	-0.213	-0.135	-0.056	-0.063	0.281	0.108	-0.034	-0.310	-0.065
	(0.169)	(0.183)	(0.187)	(0.164)	(0.201)	(0.160)	(0.160)	(0.172)	(0.177)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-1.066	-1.198	-1.254	-1.000	-1.410	-0.870	-0.829	-1.074	-1.159
	(0.089)	(0.105)	(0.106)	(0.000)	(0.120)	(0.078)	(0.078)	(0.092)	(0.102)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.658	0.738	0.718	0.569	0.552	0.721	0.613	0.618	0.676
	(0.059)	(0.067)	(0.068)	(0.056)	(0.058)	(0.060)	(0.060)	(0.067)	(0.063)
Cutoff 3	1.041	1.263	1.236	1.058	1.181	1.155	0.954	1.086	1.151
	(0.072)	(0.086)	(0.089)	(0.072)	(0.078)	(0.079)	(0.076)	(0.083)	(0.081)
Cutoff 4	1.484	1.655	1.691	1.463	1.609	1.670	1.267	1.579	1.721
	(0.090)	(0.102)	(0.110)	(0.083)	(0.087)	(0.098)	(0.094)	(0.094)	(0.103)
Cutoff 5	1.821	1.978	2.095	1.922	2.093		1.528	2.083	2.279
	(0.103)	(0.117)	(0.126)	(0.096)	(0.102)		(0.107)	(0.104)	(0.118)
Cutoff 6	2.124	2.263	2.431		2.517				
	(0.115)	(0.127)	(0.139)		(0.118)				
Cutoff 7	2.409								
	(0.126)								

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the model with mixtures 6-3.

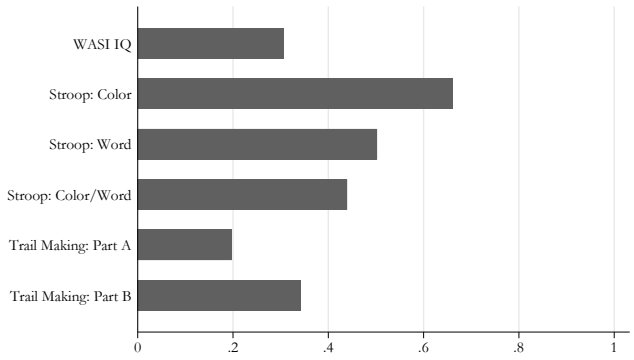
Table OA.1.21: Estimated Parameters from Factor Model — Mental Health Measures (Year 7)

	(1) Somatization	(2) Depression	(3) Anxiety	(4) Hostility	(5) Obsessive Compulsive	(6) Interpersonal Sensitivity	(7) Phobic Anxiety	(8) Paranoid Ideation	(9) Psychoticism
Constant	-0.107	0.672	0.131	0.070	1.214	-0.923	-0.979	0.863	-0.513
	(0.525)	(0.793)	(0.727)	(0.541)	(0.656)	(0.644)	(0.564)	(0.604)	(0.725)
Age 15	0.095	0.653	0.213	-0.114	0.243	0.319	0.459	0.595	0.736
	(0.286)	(0.409)	(0.359)	(0.282)	(0.324)	(0.348)	(0.324)	(0.308)	(0.388)
Age 16	-0.042	0.223	-0.472	-0.136	0.054	0.230	0.176	0.156	0.347
	(0.261)	(0.378)	(0.335)	(0.260)	(0.298)	(0.323)	(0.312)	(0.289)	(0.359)
Age 17	0.083	0.562	0.360	0.402	0.507	0.135	0.643	0.210	0.330
	(0.271)	(0.378)	(0.335)	(0.263)	(0.306)	(0.329)	(0.305)	(0.288)	(0.363)
Age 18	-0.018	0.031	-0.348	0.480	0.367	-0.429	0.445	0.468	0.515
	(0.388)	(0.539)	(0.491)	(0.366)	(0.430)	(0.514)	(0.420)	(0.408)	(0.493)
Female	0.418	0.123	0.174	0.485	0.520	0.418	-0.057	-0.095	-0.111
	(0.209)	(0.304)	(0.279)	(0.207)	(0.239)	(0.251)	(0.246)	(0.235)	(0.289)
White	0.006	-1.310	-0.019	0.679	-0.653	0.183	-0.030	-0.467	-0.698
	(0.516)	(0.780)	(0.719)	(0.532)	(0.631)	(0.618)	(0.551)	(0.591)	(0.703)
Hispanic	-0.343	-1.076	-0.222	0.357	-0.951	-0.404	0.258	-0.501	-0.081
	(0.512)	(0.759)	(0.716)	(0.528)	(0.626)	(0.612)	(0.542)	(0.581)	(0.694)
Black	-0.198	-1.117	0.065	0.486	-1.070	0.057	-0.113	-0.169	0.030
	(0.506)	(0.764)	(0.713)	(0.526)	(0.627)	(0.611)	(0.537)	(0.582)	(0.695)
Phoenix	-0.246	-0.381	0.093	-0.406	-0.007	-0.074	-0.560	-0.275	-0.235
	(0.220)	(0.305)	(0.272)	(0.211)	(0.243)	(0.261)	(0.242)	(0.234)	(0.289)
Variance	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Mental Health Factor	-0.885	-1.692	-1.539	-1.000	-1.295	-1.142	-0.885	-1.198	-1.502
	(0.111)	(0.225)	(0.188)	(0.000)	(0.162)	(0.147)	(0.116)	(0.151)	(0.193)
Cutoff 1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Cutoff 2	0.522	0.435	0.677	0.802	0.600	0.551	0.502	0.669	0.642
	(0.073)	(0.082)	(0.094)	(0.090)	(0.079)	(0.087)	(0.080)	(0.086)	(0.102)
Cutoff 3	0.854	0.942	1.345	1.385	0.983	1.025	0.982	1.077	1.245
	(0.092)	(0.124)	(0.137)	(0.114)	(0.102)	(0.124)	(0.120)	(0.105)	(0.150)
Cutoff 4	1.236	1.559	1.871	1.785	1.537	1.332	1.523	1.448	1.687
	(0.114)	(0.174)	(0.171)	(0.127)	(0.133)	(0.147)	(0.167)	(0.119)	(0.174)
Cutoff 5	1.475	1.796	2.440	2.165	1.906		1.749	1.784	2.185
	(0.129)	(0.191)	(0.212)	(0.141)	(0.154)		(0.191)	(0.134)	(0.208)
Cutoff 6	1.608	2.158	2.709		2.102				
	(0.138)	(0.216)	(0.232)		(0.164)				
Cutoff 7	1.782								
	(0.151)								

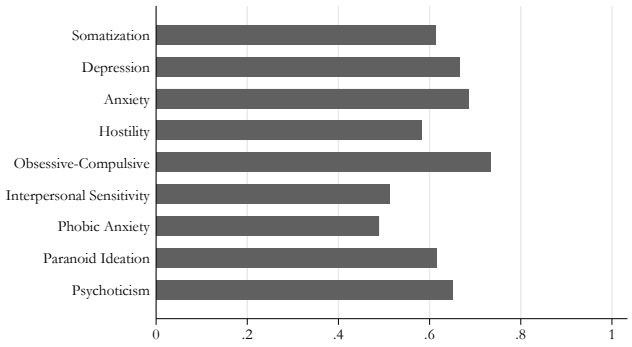
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the mental health measure system in year 7. Standard errors are shown in parentheses below the mean point estimates. Each component of the BSI is modeled using an ordered threshold model. For each measure, the number of values (K) corresponds to the number of distinct values between zero and one, including zero and one. Thus, the number of cutoffs varies across measures. The mental health factor is normalized to have a loading of negative one on the BSI hostility measure. Results correspond to the model with mixtures 6-3.

Figure OA.1.17: Fraction of the Variance Explained by the Factor

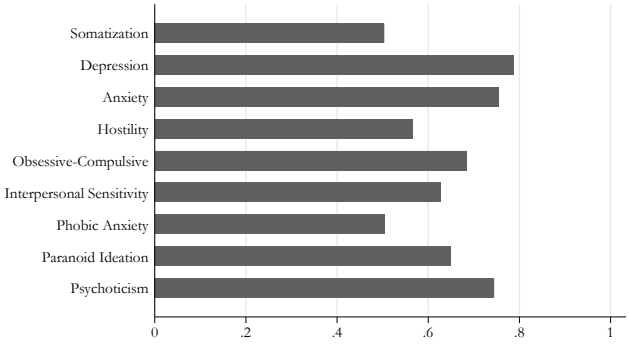
(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



(c) Mental Health (Year 7): M_7



Notes: These figures present the average fraction of the variance of each cognitive and mental health measure explained by the cognitive and mental health factors. For example, Figure OA.1.17a shows that 30.6% of the fraction of the variance of the residualized (against X) WASI IQ measure is explained by the cognitive skills C. Results correspond to the model with mixtures 6-3.

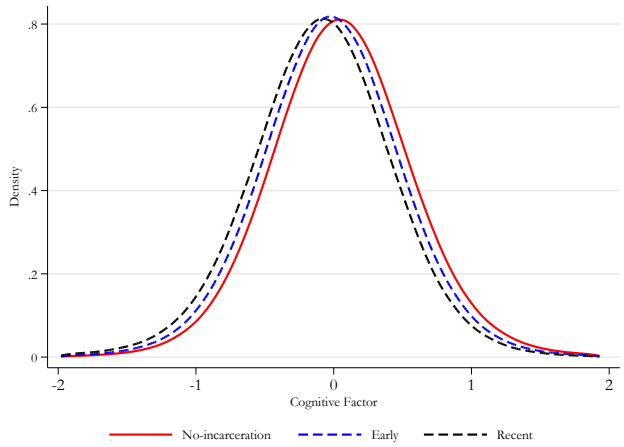
Table OA.1.22: Estimated Parameters from Model - Treatment Equation

	Mean	SD
Constant	-0.157	0.281
Age 15	0.328	0.179
Age 16	0.381	0.165
Age 17	0.166	0.172
Age 18	0.023	0.233
Female	-1.130	0.157
White	0.142	0.264
Hispanic	0.130	0.255
Black	0.434	0.260
Phoenix	0.057	0.126
Cognitive Skills (ψ_T)	-0.247	0.117
Mental Health (μ_T)	-0.058	0.048
Cutoff 1 (κ_1)	0.000	0.000
Cutoff 2 (κ_2)	1.207	0.072

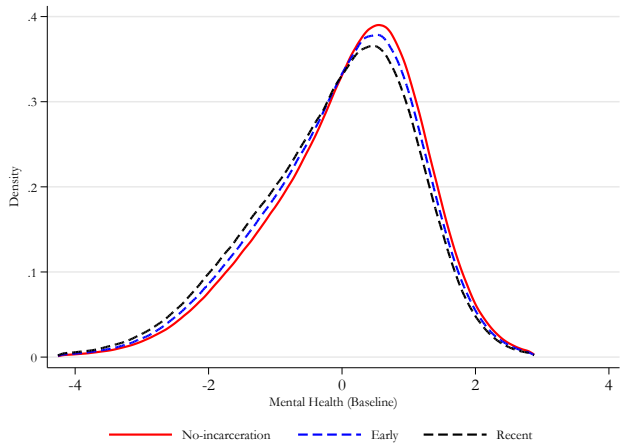
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the treatment equation. Results correspond to the model with mixtures 6-3.

Figure OA.1.18: Distribution of Cognitive and Baseline Mental Health Factors By Treatment

(a) Cognitive Skills: C



(b) Mental Health (Baseline): M



Notes: These figures show the estimated densities of the cognitive and baseline mental health factors conditional on incarceration status. For example, in Figure OA.1.18a, the red solid-line plots $f(C \mid D = 0)$, the blue dashed-line plots $f(C \mid D = 1)$, and the black dashed-line plots $f(C \mid D = 2)$. Results correspond to the model with mixtures 6-3.

Table OA.1.23: Estimated Parameters from Model - Law of Motion Equation

	Mean	SD
Early Incarceration (δ_{T1})	-0.187	0.125
Recent Incarceration (δ_{T2})	-0.451	0.160
Cognitive Skills (λ_C)	-0.230	0.162
Mental Health (λ_M)	0.373	0.082
Early Incarc. $\times$ Cognitive Skills (λ_{CT1})	0.233	0.275
Recent Incarc. $\times$ Cognitive Skills (λ_{CT2})	-0.506	0.336
Early Incarc. $\times$ Mental Health (λ_{MT1})	-0.186	0.110
Recent Incarc. $\times$ Mental Health (λ_{MT2})	-0.080	0.148

Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the law of motion of mental health. Results correspond to the model with mixtures 6-3.

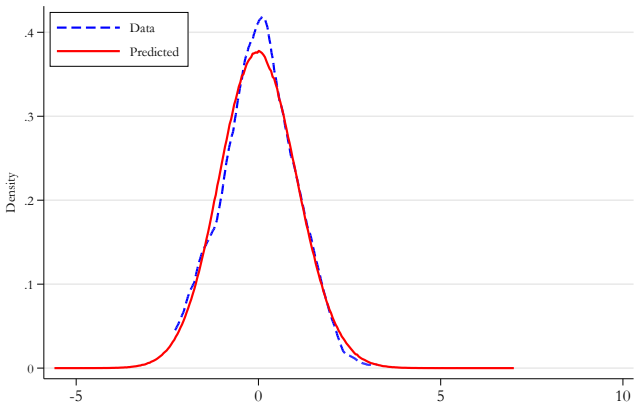
Table OA.1.24: Estimated Parameters from Multinomial Model for Missing Data

	Treatment and MH Year 7 Measures Included	Treatment Included, MH Year 7 Measures Missing
Constant	0.144	-0.108
	(0.318)	(0.310)
Age 15	-0.110	0.134
	(0.198)	(0.210)
Age 16	-0.230	0.224
	(0.184)	(0.190)
Age 17	-0.084	-0.112
	(0.185)	(0.202)
Age 18	-0.133	0.061
	(0.254)	(0.271)
Female	0.149	-0.252
	(0.149)	(0.162)
White	0.346	-0.199
	(0.304)	(0.292)
Hispanic	0.272	-0.216
	(0.297)	(0.284)
Black	0.275	-0.121
	(0.301)	(0.284)
Phoenix	-0.273	0.454
	(0.146)	(0.153)
Cognitive Skills Factor C	0.134	0.154
	(0.128)	(0.136)
Mental Health Factor M	-0.206	0.134
	(0.057)	(0.061)

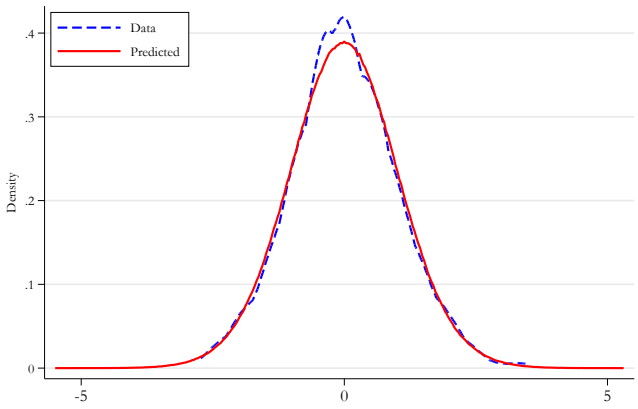
Notes: This table reports the mean and standard deviation of each parameter across 5,000 draws from the posterior distribution for the attrition equation. The reference category corresponds to observations with both treatment and mental health measures missing in year 7. Results correspond to the model with mixtures 6-3.

Figure OA.1.19: Posterior Predictive Fit - Cognitive Measures

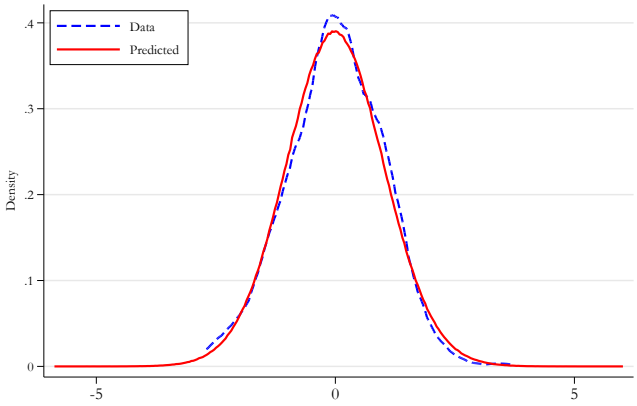
(a) WASI IQ



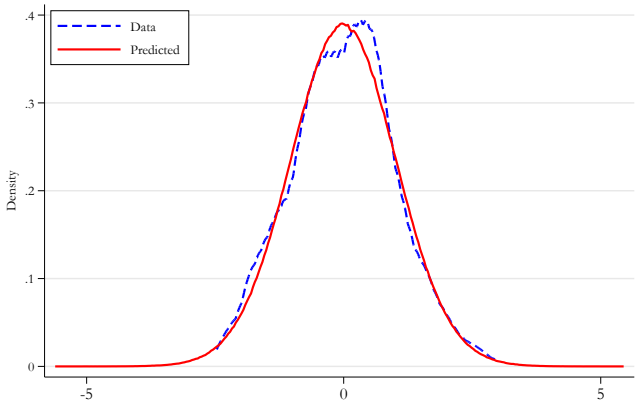
(b) Stroop Color



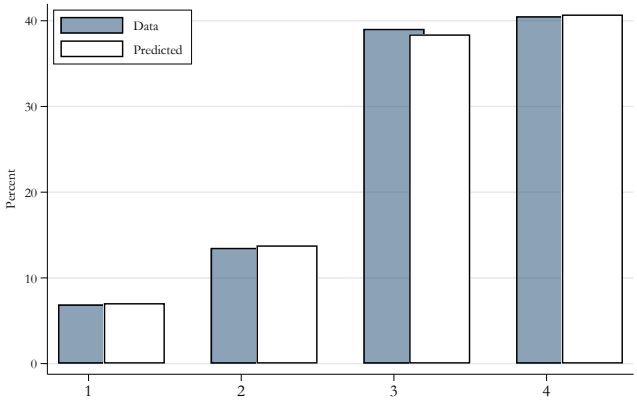
(c) Stroop Word



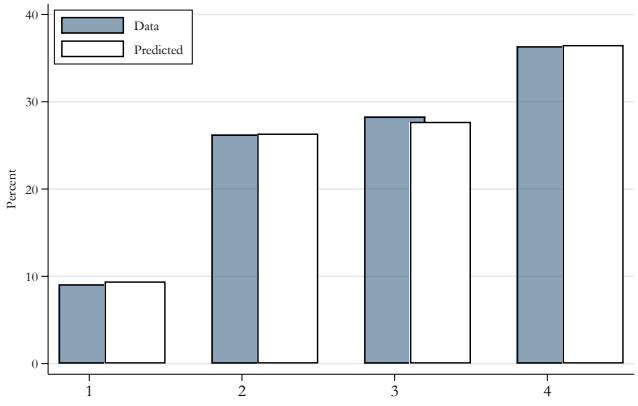
(d) Stroop Color/Word



(e) Trail Making A



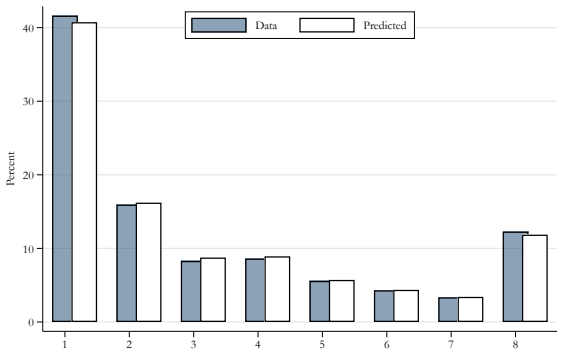
(f) Trail Making B



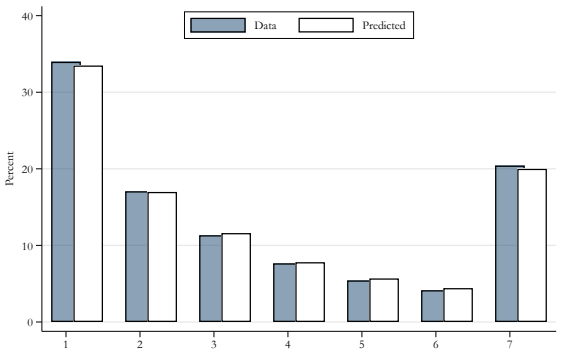
Notes: This figure compares observed and predicted distributions for each cognitive measure. For continuous variables (Figures OA.1.19a-OA.1.19d), observed distributions appear as blue dashed lines and predicted as solid red lines. For discrete variables (Figures OA.1.19e and OA.1.19f), observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the model with mixtures 6-3.

Figure OA.1.20: Posterior Predictive Fit - BSI Mental Health Measures (Baseline)

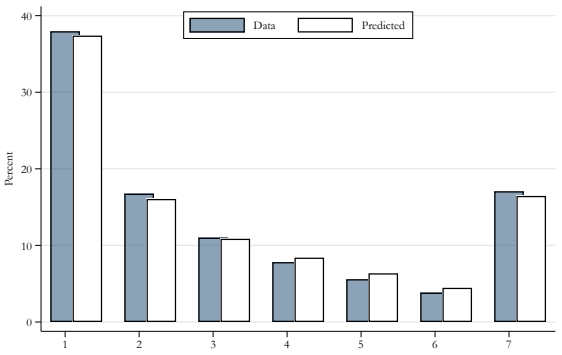
(a) Somatization



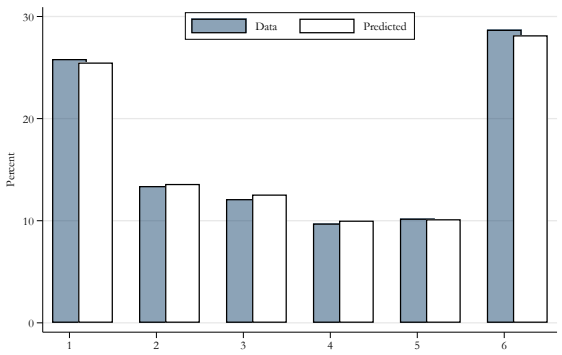
(b) Depression



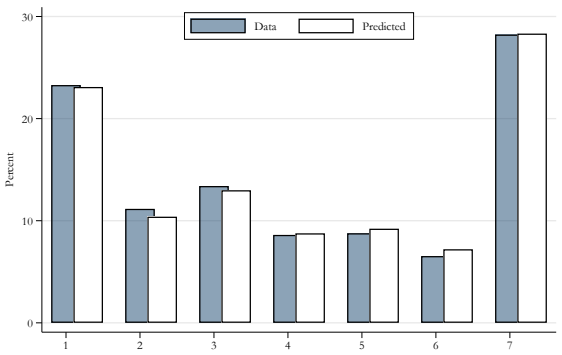
(c) Anxiety



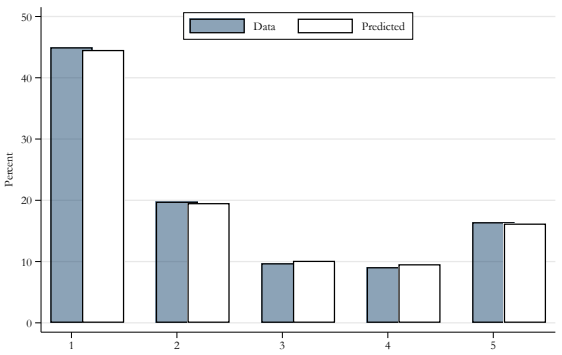
(d) Hostility



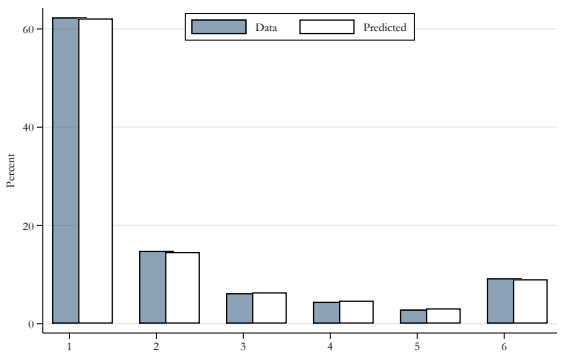
(e) Obsessive



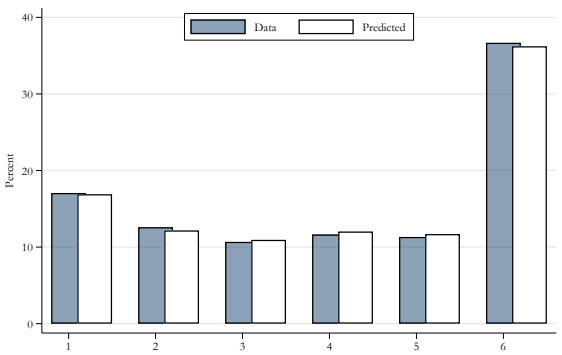
(f) Interpersonal



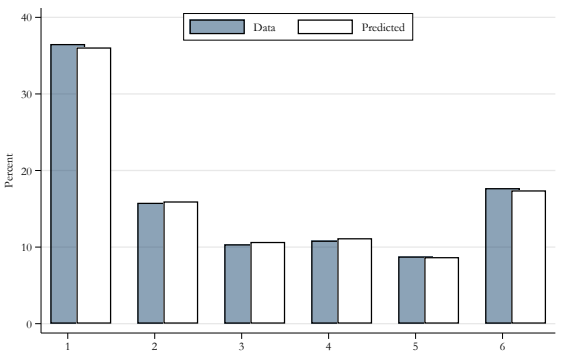
(g) Phobia



(h) Paranoia

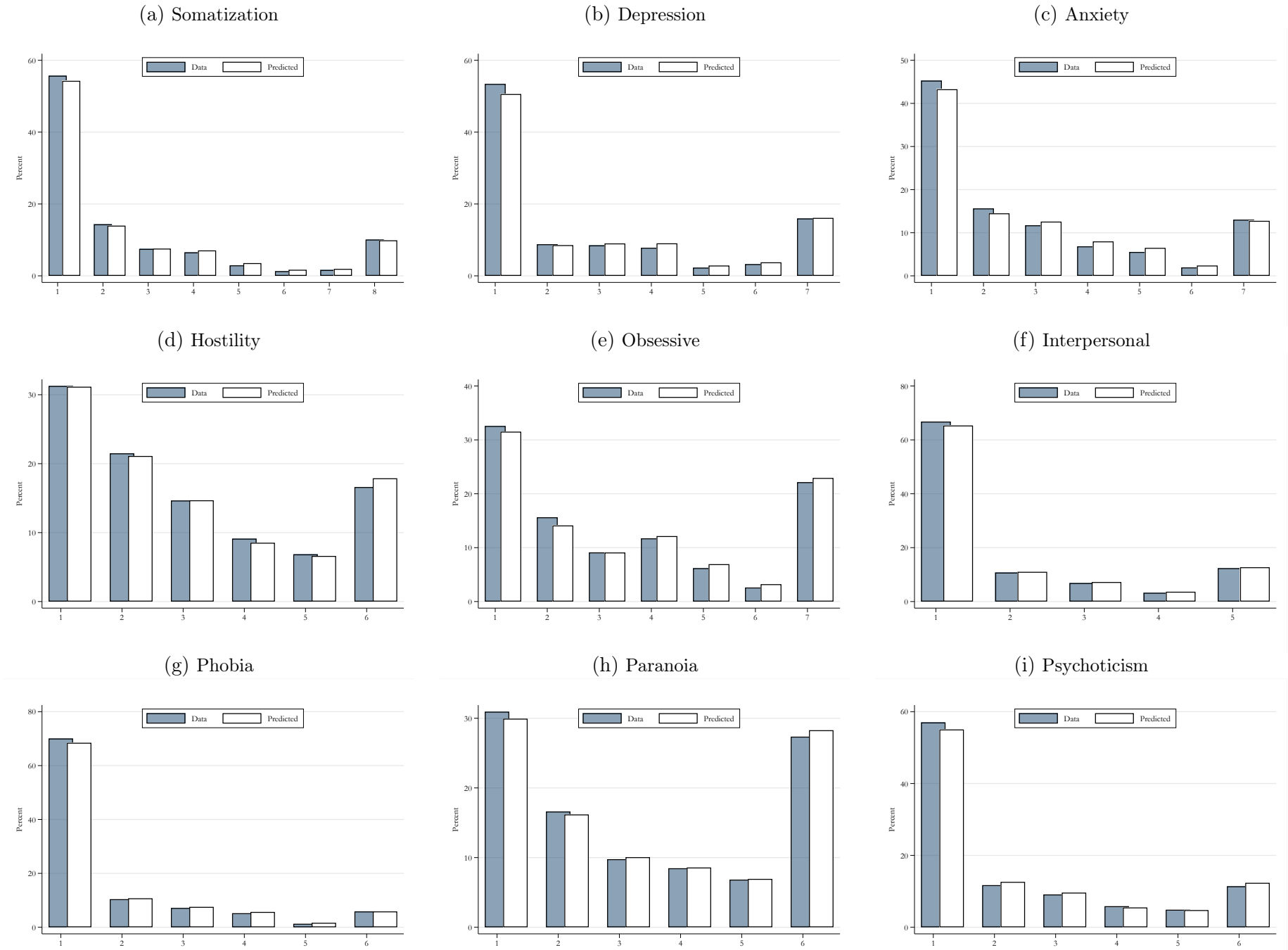


(i) Psychoticism



Notes: This figure compares observed and predicted distributions for each baseline mental health measure. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the model with mixtures 6-3.

Figure OA.1.21: Posterior Predictive Fit - BSI Mental Health Measures (Year 7)



Notes: This figure compares observed and predicted distributions for each mental health measure seven years after the baseline. Observed data are shown as shaded blue bars, with predicted bars lightly overlaid in white. Predicted values are based on 10 million simulation draws from the posterior distributions. Results correspond to the model with mixtures 6-3.

Table OA.1.25: Goodness Of Fit

	Chi-Square Classical p-value (1)	Mahalanobis Bayesian PPP (2)	Euclidean Bayesian PPP (3)
Cognitive Skills: C			
WASI IQ†	0.043	0.173	0.282
Stroop Color†	0.955	0.422	0.516
Stroop Word†	0.675	0.550	0.640
Stroop CW†	0.040	0.594	0.676
Trail Making A	0.990	0.633	0.365
Trail Making B	0.990	0.777	0.842
Mental Health: M			
BSI Somatization	1.000	0.770	0.299
BSI Depresion	1.000	0.534	0.146
BSI Anxiety	0.966	0.718	0.804
BSI Hostility	0.999	0.909	0.888
BSI Obsessive	0.992	0.859	0.764
BSI Interpersonal	0.993	0.024	0.027
BSI Phobia	0.999	0.918	0.963
BSI Paranoia	0.997	0.634	0.633
BSI Psychoticism	0.999	0.930	0.858
Mental Health: M_7			
BSI Somatization	0.997	0.072	0.392
BSI Depresion	0.971	0.214	0.212
BSI Anxiety	0.959	0.289	0.301
BSI Hostility	0.997	0.224	0.508
BSI Obsessive	0.982	0.559	0.825
BSI Interpersonal	0.993	0.583	0.592
BSI Phobia	0.995	0.106	0.315
BSI Paranoia	1.000	0.554	0.378
BSI Psychoticism	0.986	0.598	0.249

Notes: This table presents three measures of Goodness of Fit for our cognitive and mental health measures. Column (1) displays the *Classical* p-values from the chi-square distribution. Columns (2) and (3) report *Bayesian* Posterior Predictive P-values using the Mahalanobis distance and the joint Euclidean distance of means and standard deviations, respectively. For discrete variables, the total number of bins is the number of categories in the empirical distribution. For continuous variables (indexed by a dag), we set the number of bins equal to 4 and the cutoffs equal to the quartiles of the simulated distribution. Results correspond to the model with mixtures 6-3.

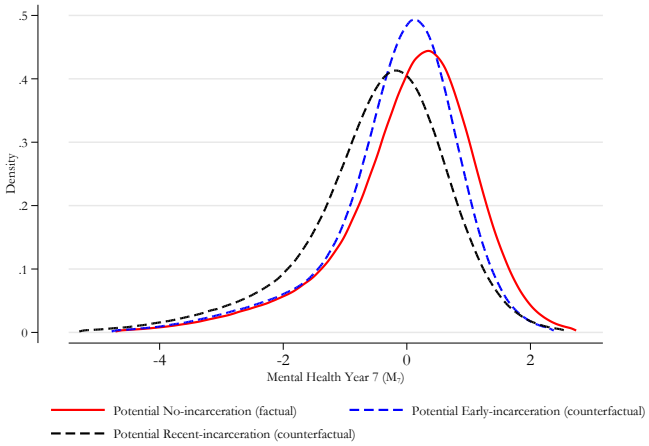
Table OA.1.26: Mean Treatment Effects

	Mean (1)	lb (2)	ub (3)	P($\cdot > 0$) (4)	P($\lvert\cdot\rvert < 0.01$) (5)	P($\lvert\cdot\rvert < 1\%$) (6)
ATE($M_7, 1, 0$)	-0.187	-0.436	0.057	0.067	0.021	0.019
ATE($M_7, 2, 0$)	-0.451	-0.773	-0.136	0.003	0.002	0.002
ATT($M_7, 1, 1, 0$)	-0.187	-0.438	0.055	0.068	0.022	0.021
ATT($M_7, 1, 2, 0$)	-0.439	-0.758	-0.124	0.004	0.002	0.003
ATT($M_7, 2, 1, 0$)	-0.188	-0.453	0.062	0.079	0.027	0.028
ATT($M_7, 2, 2, 0$)	-0.391	-0.722	-0.057	0.012	0.003	0.004
ATT($M_7, 0, 1, 0$)	-0.186	-0.439	0.065	0.072	0.019	0.018
ATT($M_7, 0, 2, 0$)	-0.486	-0.819	-0.162	0.001	0.001	0.001
MTE($M_7, 1, 1, 0$)	-0.189	-0.455	0.072	0.078	0.023	0.023
MTE($M_7, 1, 2, 0$)	-0.457	-0.801	-0.116	0.005	0.002	0.002
MTE($M_7, 2, 1, 0$)	-0.188	-0.473	0.081	0.090	0.022	0.024
MTE($M_7, 2, 2, 0$)	-0.419	-0.776	-0.055	0.013	0.004	0.005

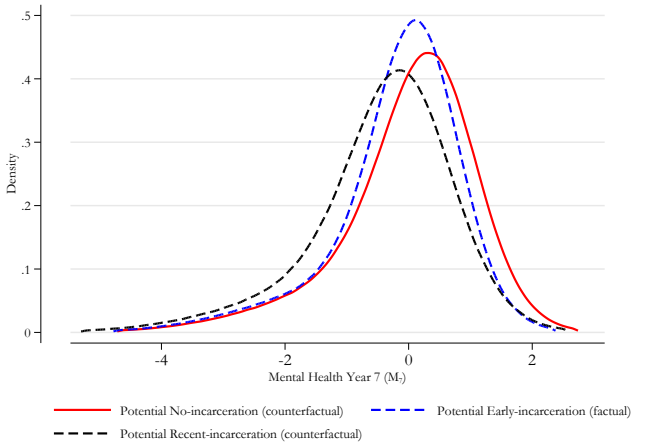
Notes: Let Θ_s be the collection of parameters evaluated using the s-th parameter draw of the estimated posterior of the model. Column (1) presents $\frac{1}{S}\sum_{s=1}^{S} ATE(M_7, 1, 0; \Theta_s)$. Columns (2) and (3) present the 2.5 and 97.5 percentiles of $\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S}$. Columns (4), (5), and (6) are $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[ATE(M_7, 1, 0; \Theta_s) > 0]$, $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01]$, and $\frac{1}{S}\sum_{s=1}^{S} \mathbb{1}[|ATE(M_7, 1, 0; \Theta_s)| < 0.01 \times Range(\{ATE(M_7, 1, 0; \Theta_s)\}_{s=1}^{S})]$, respectively. Similar definitions apply to the remaining parameters. Results correspond to the model with mixtures 6-3.

Figure OA.1.22: Factual and Counterfactual Distributions of Mental Health and Mental Health Gain (Year 7) Conditional on Incarceration Status

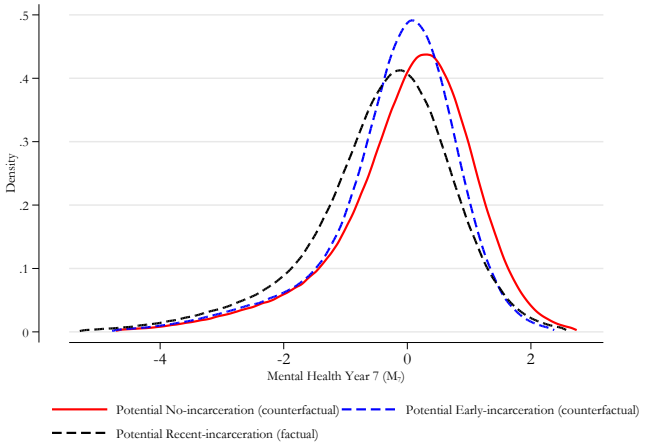
(a) M_7 Conditional on $D = 0$



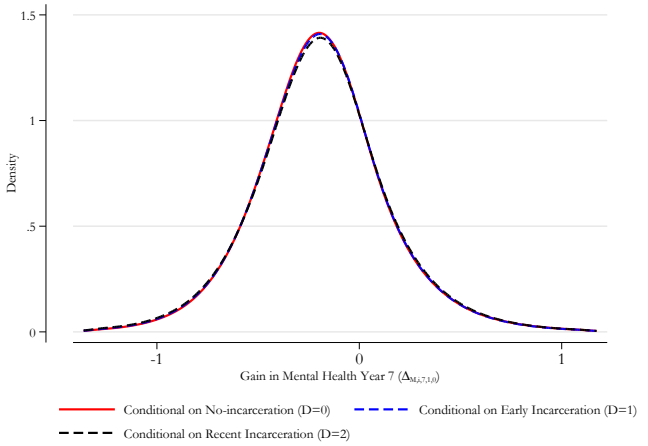
(b) M_7 Conditional on $D = 1$



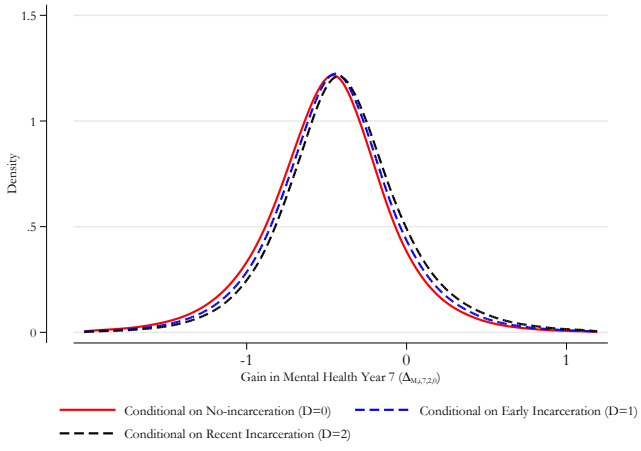
(c) M_7 Conditional on $D = 2$



(d) $\Delta_{M_i,7,1,0}$ by Incarceration Status (D)



(e) $\Delta_{M_i,7,2,0}$ by Incarceration Status (D)

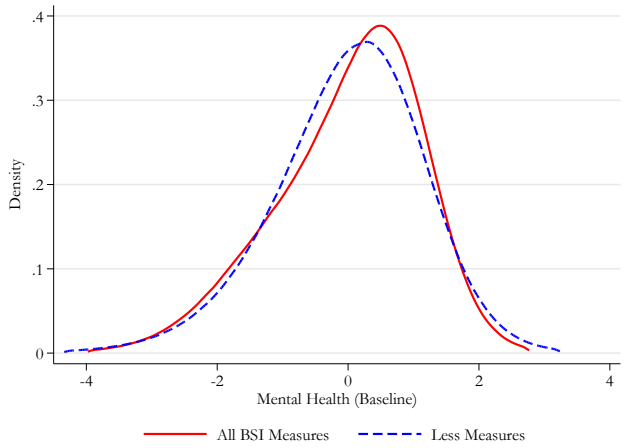


Notes: Figure OA.1.22a displays the factual density $f(M_{i,7,0}|D_i = 0)$ and counterfactual densities $f(M_{i,7,1}|D_i = 0)$ and $f(M_{i,7,2}|D_i = 0)$, for people who do not go to prison. Similar definitions apply to Figures OA.1.22b and OA.1.22c. Figure OA.1.22d displays the distribution of mental health gain from early imprisonment, $\Delta_{M_i,7,1,0}$, conditional on no imprisonment (red solid line), early imprisonment (blue dashed line), and recent imprisonment (black dashed line). Similarly, Figure OA.1.22e displays the distribution of mental health gain from recent imprisonment, $\Delta_{M_i,7,2,0}$, by incarceration status. Results correspond to the model with mixtures 6-3.

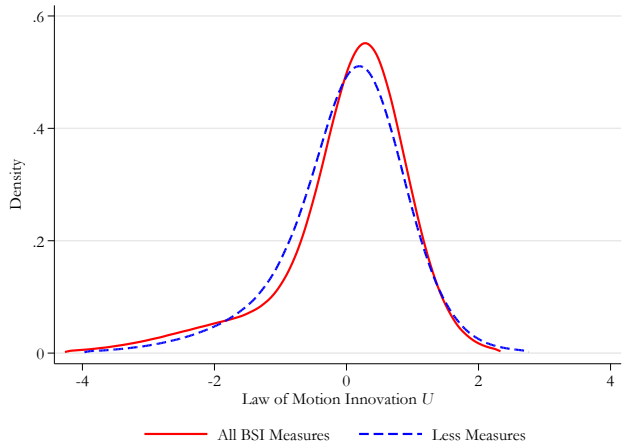
OA.1.4 Robustness to Less BSI Measures

Figure OA.1.23: Estimated Posterior Distributions of the Factors

(a) Distribution of M



(b) Distribution of U



Notes: This figure reports the estimated posterior distributions of the factors. Figure OA.1.23a displays the estimated density of mental health at baseline, while Figure OA.1.23b displays the estimated density of the law of motion innovation to mental health in year 7. Results correspond to the baseline model with all BSI measures and an alternative model using measures 2, 4, 6, and 8 of the BSI subscales (Depression, Hostility, Interpersonal Sensitivity, Paranoid Ideation), both at baseline and after seven years.