What Good are Treatment Effects without Treatment? Mental Health and the Reluctance to Use Talk Therapy*

Christopher J. Cronin†
Department of Economics, University of Notre Dame

Matthew P. Forsstrom‡
Department of Economics, Wheaton College

Nicholas W. Papageorge§
Department of Economics, Johns Hopkins University, IZA and NBER

May 24, 2024

Abstract: Evidence across disciplines suggests that talk therapy is more curative than antidepressants for mild-to-moderate depression and anxiety. Yet, few patients use it. We develop a dynamic choice model to analyze patient demand for the treatment of depression and anxiety. The model incorporates myriad potential impediments to therapy use along with links between mental health improvements and earnings. The estimated model reveals that mental health improvements are valuable, directly through utility and indirectly through earnings. However, patient reluctance to use therapy is nearly impervious to reasonable counterfactual policies (e.g., lowering prices or removing other costs). Patient behavior might reflect stigma, biases in beliefs about the effectiveness of therapy, or a distaste for discussing personal or painful issues with a stranger. More broadly, the benefits of therapy estimated in randomized trials tell only half the story. If patients do not use treatments outside of an experimental setting—and we fail to understand why or how to get them to—estimated treatment effects cannot be leveraged.

Keywords: Mental Health, Demand for Medical Care, Labor Supply, Structural Models.
JEL Classification: I10, I12, J22, J24.

*First draft: October 31, 2016. Current draft: May 24, 2024. We gratefully acknowledge helpful comments from our editor, Jérôme Adda, three anonymous referees, as well as Daniel Avdic, Victoria Baranov, Sonia Bhalotra, Pietro Biroli, David Bradford, Jeffrey Campbell, Janet Currie, Michael Dickstein, Fabrice Etile, Bill Evans, Richard Frank, George-Levi Gayle, Donna Gilleskie, Barton Hamilton, Robert Moffitt, and Michael Richards, along with seminar participants at the UPenn, Rice, Notre Dame, UNC, AHEW St. Louis, ASHEcon Philadelphia, Southeastern HESG Richmond, H2D2 Ann Arbor, SOLE Raleigh, Essen Health and Labour Conference, EWEHE Prague, Barcelona GSE Summer Forum, and DSE Bonn. A previous version of this paper was circulated as "Mental Health, Human Capital, and Labor Outcomes."

†Corresponding Author: 3053 Jenkins Nanovic Halls, Notre Dame, IN 46556. ccronin1@nd.edu.
‡matthew.forsstrom@wheaton.edu.
§papageorge@jhu.edu.

1 Introduction

Mental illness is widespread and costly. Roughly one in five adults in the US experiences mental illness each year, the most common being mild-to-moderate depression and anxiety (NSDUH). Such conditions not only impose direct costs by making daily life more of a struggle and less enjoyable, they can also have indirect costs, such as lower labor market productivity and income (Frank and Gertler, 1991). Yet, how patients choose to treat mental illness remains poorly understood. Below we describe a broad, cross-disciplinary literature which suggests that a course of talk therapy (or simply "therapy", henceforth) is more curative than antidepressants for mild-to-moderate depression and anxiety, yet the vast majority of individuals treating these conditions opt for the latter. For example, estimates from the 2011 National Survey on Drug Use and Health (NSDUH) indicate that about 11.5 percent of Americans over age 18 used an antidepressant in the past year. For comparison, an estimated 3.8 percent of Americans over the age of 18 received care in a therapist's office.[1]

This paper develops a structural model of dynamic mental health treatment choices in the context of depression and anxiety. The aim is to shed light on why patients do not use the treatment with the highest average effectiveness to improve mental health. The model captures various factors affecting patient decisions, in particular, reluctance to use therapy versus antidepressants. Taking the model to data, we find that while costs often characterized as critical barriers to use, such as the high price of therapy or time costs, help to explain patient decisions, they cannot fully explain patient reluctance to use therapy. This unwillingness is thus captured as a negative utility cost, which could reflect stigma but may also capture the fact that talking about private problems with a stranger is an arduous or odious prospect for many individuals, especially when an alternative treatment, antidepressants, is available. A consequence is that counterfactual policies that remove the costs we explicitly model do relatively little to change treatment use and mental health. This overarching finding underscores challenges to addressing what amounts to a population health crisis. It also suggests that treatment effects estimated in well-identified settings, while a useful factor to understand how to improve health, are difficult to leverage if patients are reluctant to use the treatment and we fail to understand why or how to get them to do so.

The model envisions agents making repeated dynamic choices about mental health treatment and labor supply. The labor supply decision is standard: work has a time cost but increases income and consumption. Treatment decisions are more complex. Both therapy

[1]We report utilization rates for 2011 because it is the last year of our sample period. Since 2011, antidepressant and therapy use have risen. Therapy use peaked in 2020 at about 5.5 percent but then declined in 2021 (the latest year available for the NSDUH) to about 4.7 percent. Antidepressant use in 2021 was about 13.5 percent. In other words, there is no evidence that therapy use has increased dramatically enough to approach rates of antidepressant use. The gap between the two—the focus of this paper—persists.

and antidepressants involve out-of-pocket payments: costs related to employment, such as time costs for therapy and side effects for antidepressants, and costs embodied in uncertainty about therapy treatment effects. Both treatments improve mental health, which has a direct impact on utility and can also increase earnings. Given heterogeneity in the efficacy of therapy documented elsewhere (Wampold and Owen, 2021), we assume individuals face a distribution of potential treatment effects and only learn the impact of therapy after their first session. Subsequently, they choose how many sessions to attend. This feature not only captures how, in some cases, therapy may not be very effective, it also allows the model to rationalize a consistent empirical pattern in therapy use: many patients go to one or two sessions and then stop, which the model explains as a low treatment effect draw. We model antidepressant use as a binary choice, and individuals are assumed to expect the average treatment effect, as there is little evidence of strong variation in the impact of antidepressants on future mental health. Finally, we permit two forms of unobserved heterogeneity, permanent and time-varying, the latter of which allows individuals to experience intra-period mental health shocks that can drive them to use either treatment. These shocks help explain negative selection into treatment, or why some individuals in the data appear to experience mental health declines after receiving treatment.

We use moments from several data sources in estimation. First, we use the 1996–2011 cohorts of the Medical Expenditure Panel Survey, which, apart from mental health treatments and conditions, also contain rich data on labor supply and earnings. One unique feature of these data is that they include mental health information for individuals who are unemployed, which allows us to explore links between mental health conditions, employment decisions, and related outcomes. In principle, we should be able to estimate the impact of treatment on mental health using these data; however, the selection problem mentioned previously and a lack of credible instruments make doing so difficult. Thus, our preferred specification relies on findings from a collection of randomized controlled trials summarized in the medical and psychology literatures. We use these outside data to fix the mean of the distribution of mental health treatment effects. The variance of the therapy treatment effect distribution is then estimated using variation in how many sessions individuals choose to attend.

Model estimates are generally aligned to priors and reflect descriptive patterns in the data. Individuals derive utility from mental health, which we quantify below. Mental health treatment carries a utility cost, consistently larger across demographic groups for therapy than for antidepressants. Utility costs of treatment are larger for people who are employed and smaller for individuals who have used treatment in the past. When choosing therapy, individuals face a symmetric distribution of treatment effects, implying that 35 percent receive a negative draw, while just as many receive a draw that is over twice the clinical mean. Turning to the labor market, individuals are willing to work, sacrificing leisure, because in doing so

they gain utility through consumption, governed by a coefficient of relative risk aversion parameter estimate of 0.24. Parameters that govern time-varying unobserved heterogeneity imply persistent differences in the magnitude and likelihood of negative intra-period shocks to mental health. For example, approximately 28 percent of individuals receive a negative shock each period and are subsequently 40 (30) percent more likely to use therapy (antidepressants), all else equal. Finally, parameters governing permanent unobserved heterogeneity suggests there are three types with persistent differences in mental health and utility of treatment and work. One of these types accounts for roughly 25 percent of the population that has persistently poor or fair mental health, elevated treatment levels, and low rates of employment.

We illustrate the value of mental health improvements in our first set of counterfactuals, which suppose there is a costless technology that provides a lower bound on mental health equal to the sample mean. We then compute willingness to pay (WTP) for such a technology for a single six-month period. In 2023 dollars, average WTP is \$2,536 for the full sample and \$8,536 for those with bad enough mental health to use the technology, which we deem the "sick sample." A back-of-the-envelope calculation puts total annual willingness to pay among US adults aged 26–55 at \$854B. Only 2.1 percent of this total accrues through earnings gains, driven by increases in labor supply. Average wages actually decline as very small increases in wages for existing employees are overshadowed by the movement of low-wage (formerly mentally unwell) individuals into the labor force. These indirect labor market gains are smaller compared to those found in other studies, which do not account for how people in poor mental health tend to have high disutility costs of work and low productivity even when their mental health improves. Importantly, finding large utility gains from mental health rules out one potential explanation of low therapy use: that people simply do not value mental health very much. Instead, our findings show that individuals value mental health but choose not to use the most effective treatment available. Our remaining counterfactuals are designed to shed light on why.

In particular, our second set of counterfactuals assesses the impact of assignment to therapy. The model allows us to vary this policy along several dimensions, including whether or not individuals assigned to therapy can choose how many sessions to attend and whether or not they draw from the full distribution of treatment effects, which leads to a wide range of impacts. If individuals are assigned a full 12 sessions of therapy and obtain the mean treatment effect, we predict large increases in treatment uptake in future periods and moderate gains to mental health. Mental health gains are substantial if we narrow our focus on the sickest individuals (7 percent of the population) with low mental health at the time the policy is implemented and who are prone to intra-period negative shocks to mental health.[2] In contrast, we predict far smaller impacts if

[2]Quantifying these gains is nuanced given the categorical measure of mental health we use to derive a continuous latent mental health measure. The sickest individuals experience improvements that are roughly

individuals are in relatively good mental health, which means they have little to gain from therapy; if they draw from the full distribution of treatment effects; or if they are only assigned to a single session and can choose how many sessions to attend afterward. The last point means that getting people to go to therapy is of limited value if the intensive margin is endogenous since many individuals drop out. Finally, even the largest gains to mental health are not accompanied by positive employment or wage effects. This finding runs counter to narratives suggesting that policy aimed at improving mental health could "pay for itself" through increased employment or worker productivity (Laynard et al., 2007). There are two explanations: one, among individuals with the most to gain from therapy, labor supply and wages are not very elastic to mental health. Two, therapy comes with time and other employment related costs that make working more difficult.

While random assignment to therapy, either in an experimental setting or as a counterfactual policy, can shed light on the impact of therapy, the population-level benefits of estimated treatment effects accrue only if people actually use therapy, which they rarely do. This fact leads us to explore policies that focus on individual choices versus random assignment. Specifically, in our third set of counterfactuals we assess responsiveness to several commonly-suggested policies that would presumably lower barriers to therapy and thus increase usage. Eliminating monetary costs of therapy has negligible effects on usage and mental health, while removing time and employment costs leads to a roughly 45 percent increase in usage. However, a low base rate means that this translates to about a 1.5 percentage point gain in therapy use. We also examine therapy uptake if we improve the effectiveness of the mental health treatment. Rather than simply increasing the mean treatment effect, we use the model to examine a somewhat more realistic possibility: eliminating below-average treatment effects from the distribution patients face. This would be akin to enforcing best practices and/or monitoring and retraining therapists with poor track records. This policy leads to similar increases in therapy uptake of roughly 45 percent. Importantly, none of the policy changes lead to dramatic absolute gains in therapy use (e.g., none cause the majority of individuals who could benefit from therapy to use it) that would lead to non-negligible changes to population mental health, much less earnings.

Results from these counterfactuals are somewhat bleak. The main takeaway is that factors widely viewed as critical barriers to therapy use (e.g., monetary and time/employment related costs) explain little patient reluctance to use the treatment. Even improvements to the quality of therapy by eliminating the worst treatment effects lead to very small changes to population mental health. Generating meaningful improvements in population mental health thus requires that we look elsewhere. Our estimates suggest that disutility from therapy is the primary disincentive for use. There are several reasons this might be the case. This finding may reflect that individuals do not like therapy because it is more difficult than taking a pill. Therapy

equivalent to one half of the distance between subjectively reported "fair" and "good" mental health.

requires not only time but also effort and perseverance. It may be uncomfortable and taxing to discuss personal problems with a stranger, to confront sources of mental health problems, or to avoid familiar coping strategies, thought patterns, and habits. Indeed, a patient's commitment to successful therapy is often characterized as "hard work" (Göstas et al., 2013; Werbart et al., 2019). If so, policy designed to increase therapy use would need to change the way therapy occurs, which may not be possible without sacrificing its effectiveness. Other potential explanations include stigma and biased beliefs about how effective therapy is. We return to a discussion of these possibilities in the conclusion.

The paper proceeds as follows: Section 2 discusses related research. Section 3 introduces the data used in this project and highlights several key empirical patterns that motivate our analysis. Section 4 introduces the dynamic choice model. Section 5 discusses estimation and identification. Section 6 presents parameter estimates, model fit, and counterfactual policy simulations, which use the estimated dynamic choice model. Section 7 concludes.

2 Literature

In studying mental health treatment choices, we contribute to a large literature engendered by Grossman (1972) that views medical treatment decisions as rational, dynamic choices made under uncertainty. Within this framework, medical treatment is seen as a costly investment. Rational, forward-looking patients make decisions by weighing current and future costs and benefits of different treatment options. The framework has been applied to a number of health contexts, such as chronic illness (Cronin, 2019) and infectious disease (Chan et al., 2015), and extended to incorporate additional features of healthcare decisions, such as learning and uncertainty about treatment quality (Crawford and Shum, 2005; Chan and Hamilton, 2006), drug side effects (Papageorge, 2016), risky behaviors that affect illness (Arcidiacono et al., 2007; Darden, 2017), and links between health and the labor market (Gilleskie, 1998).

A smaller, growing literature in economics also studies mental health. Generally, this literature documents that mental health is valuable and corroborates the medical literature showing that therapy improves mental health. An early contribution is Ettner et al. (1997) who provide evidence that psychiatric disorders significantly reduce employment, hours worked, and income. In a more recent and groundbreaking study, Baranov et al. (2020) find that random assignment to therapy for women with postpartum depression yields substantial mental health benefits that extend over many years. In another recent study, Jolivet and Postel-Vinay (2020) estimate a job search model where mental health and job stress play a role in selection into employment and across types of jobs. A key finding is that negative mental health shocks are very costly, equating to roughly one-third of the cost of losing a job for an average worker. Regardless, their key findings create a puzzle. If therapy is highly effective, and people value mental health, then why is

therapy rarely used? The unique contribution of this paper is to shed light on this puzzle, which we do by incorporating various costs and benefits of different mental health treatments into a unified framework to assess treatment choices, in particular, reluctance to use therapy. Doing so is an important extension of previous work on the value of mental health and the benefits of therapy, which can only be harnessed if people opt for therapy outside of experimental settings.

To our knowledge, we are the first to apply the Grossman framework in the form of a structural dynamic model of treatment and employment decisions to understand how forward-looking individuals manage their mental health.[3] This gap in the literature is itself puzzling. It likely arises from some of the econometric issues we encounter in this study, including difficulties measuring mental health and the impact of treatment due to coarse subjective mental health measures; nonrandom selection into diagnosis and into treatment; as well as limited data available to relate mental health, treatment decisions, and labor market outcomes. Another unfortunate reason for this gap in the literature is that mental health problems—perhaps due to widespread stigma or ignorance—may be seen as fundamentally different from physical health problems. The implicit suggestion is that rational choice, applied in a wide variety of medical contexts, is somehow inappropriate for an analysis of mental healthcare. This position ignores that the vast majority of mentally ill individuals manage relatively mild illnesses.[4] Moreover, it impedes progress on the fundamentally important question of why people do not use a beneficial treatment despite widespread evidence of its effectiveness, which is the focus of this paper.

Finally, we contribute to a massive and well-developed medical and public health literature on the determinants and consequences of mental health issues, the effectiveness of mental health treatment, and predictors of mental health treatment choices. Our approach is motivated by earlier work on the substitutability of mental health treatments (Elkin et al., 1989; Berndt et al., 1997) and patient price sensitivity (Frank and McGuire, 1986; Keeler et al., 1988). In addition, we contribute to research examining how mental health, treatment, education, and the labor market interact for both adolescents (Currie and Stabile, 2006) and for adults (Frank and Gertler, 1991; Ettner et al., 1997; Butikofer et al., 2020; Shapiro, 2020). Much of this literature focuses on more severe mental health problems, whereas we consider a representative sample, which includes individuals suffering from moderate, mild, or no mental illness at all. We are thus able to place focus on the relatively large set of individuals who are not severely ill

[3]Davis and Foster (2005) use the Grossman framework to study a parent's choice to seek mental health treatment for their children. Yet, as Currie and Stabile (2006) mention, the framework has generally not been applied to mental health investments.

[4]According to the National Survey on Drug Use and Health, in 2015 18 percent of US adults reported mental illness in the past year, while only 4 percent reported a *serious* mental illness, defined as those, "resulting in serious functional impairment, which substantially interferes with or limits one or more major life activities." Even among these individuals, inpatient treatment, much less institutionalization, is rare. In 2008, only 7.5 percent of individuals reporting a serious mental illness sought inpatient treatment.

but who could benefit from therapy and yet choose not to, even when the costs of doing so are drastically reduced. Finally, we relate to medical and psychological literature studying barriers to access and stigma as possible reasons why therapy uptake is low (Corrigan, 2004).

3 Data

3.1 Data Set

Our empirical analysis uses data from the Medical Expenditure Panel Survey (MEPS), which has been collected annually since 1996 by the Agency for Healthcare Research and Quality (AHRQ). Each year, a nationally representative sample of new participants (i.e., a cohort) is added to the MEPS, drawn randomly from the previous year's National Health Interview Survey (NHIS) sample. Each cohort is interviewed five times over the two years that follow January 1 of the cohort year.

Several characteristics of the MEPS make it well suited for our purposes. The MEPS contains individual-level panel data on mental health, treatment, and employment, which permits estimation of a model capturing how these choices and outcomes interact. To our knowledge, no other publicly available data set offers this unique set of features. Moreover, the MEPS offers several clear advantages over the large, administrative claims data that have become popular in the literature. For example, the MEPS consistently reports a mental health measure that does not require diagnosis, which is thus prone to selection into the sample, e.g., if one hopes to model dynamic mental health transitions as motivation for future treatment. Claims data reveal treatment decisions, but mental illness can only be measured via diagnosis, which requires a patient to endogenously choose to visit a physician. Moreover, researchers typically acquire claims data from large, self-insuring employers, meaning all observed individuals are employed and insured. Critical to our study is the relationship between mental health and employment, which may be influenced by insurance status.

Despite these advantages, the MEPS data have some drawbacks. First, the panel is short and individuals enter at various points over the life cycle; thus, we need to address endogenous initial conditions with our econometric specification. Second, while it is valuable to model state-to-state mental health transitions as a function of treatment, doing so yields unreliable estimates, often with the wrong sign due to negative selection into treatment. We thus use outside data on treatment effects and time-varying unobserved heterogeneity to account for intra-period mental health shocks, as explained in Section 5. Third, survey data are likely to contain measurement error in key variables, such as wages, medical care prices, and medical care treatment, as the data rely on accurate self-reports of events that may have occurred several months in the past. When administrative data sets report these variables, they are likely subject to less measurement error; however, many such data sets exclude variables like wages

and the price of specific medical treatments due to individual and corporate privacy concerns.

Our estimation sample is comprised of individuals 26–55 years old from the 1996–2011 MEPS cohorts.[5] The first interview period provides information for initial conditions. We exclude individuals who miss one or more interviews, as well as those with an interview period that is less than three and a half months or greater than seven months. Interview period length is determined randomly by AHRQ. The first restriction reduces the full sample by about 12 percent and the second by another 35. The resulting sample, which we refer to as Sample C, looks virtually identical to the full sample on observables and consists of 54,989 individuals and 208,113 individual-period dyads.[6] In Appendix Section A.I.1, we provide additional details regarding sample restrictions.

3.2 Mental Health and Treatment in the MEPS

The MEPS offers several ways to measure an individual's mental health and associated treatment decisions. As we are interested in mental health broadly, the primary measure that we employ is taken from the question, "In general, would you say that your mental health is excellent (5), very good (4), good (3), fair (2), or poor (1)?" In Appendix Section A.I.2, we discuss three alternative measures of mental health contained in the MEPS. Each of these alternatives correlates strongly with our preferred measure, but each has a significant downside that prevents us from using it directly in our analysis. For example, the Kessler 6 index was added to the MEPS in 2005 but is only collected in interview rounds 2 and 4.

While any particular mental health condition could produce variation in subjective mental health, the variation in our sample largely relates to a narrow set of psychological disorders that share a common set of symptoms and treatments (namely, Stress Induced Disorders (ICD-9 Codes 308 and 309), Anxiety Disorders (ICD-9 Code 300), and Depressive Disorders (ICD-9 Codes 296 and 311), which we call "SAD" disorders). Among all interview rounds in which a mental illness (i.e., ICD-9 Codes 290–319) is reported, 93 percent contain a SAD report, while only 12 percent contain a non-SAD report.[7] Furthermore, we show below that subjective mental health is highly correlated with reports of these SAD disorders.

Individuals report the date, location, and price of medical treatments, as well as the condition being treated, which is later mapped to an ICD-9 code by AHRQ. For prescription drugs, individuals report the same, as well as the name, dose, and refill information about the drug.[8]

[5]We do not use interviews after 2012 because International Classification of Disease (ICD) codes are not recorded for prescriptions and office visits, preventing us from determining which conditions are being treated.

[6]Sample statistics reported in this section are for Sample C unless noted otherwise. To decrease estimation time, we estimate the structural model using a 20 percent random sample from Sample C, which we refer to as Sample D.

[7]Examples of other mental illnesses include substance abuse disorders, dementia, schizophrenia, psychosis, ADHD, autism, and brain injuries/deformities, among others.

[8]AHRQ verifies treatment reports with providers via telephone and mail surveys. Pharmacies are contacted regarding each reported fill/refill. Physicians are contacted for reported office-based visits but are subsampled

Because self-reported mental health is not condition-specific, the medical treatments we model can be related to any mental health condition. Specifically, an office visit is coded as a therapy session if the respondent (i) visited a medical professional in person, (ii) reports receiving therapy/counseling, *and* (iii) the visit relates to a mental health condition (i.e., ICD-9 code $\in$ [290,319]). This definition includes therapy received from a psychiatrist, psychologist, or social worker but does not include, for example, a one-time visit to a psychiatrist where prescription drugs are prescribed yet no therapy is received. An individual is coded as using prescription drugs during an interview round if he or she filled a prescription for the treatment of a mental health condition (again, ICD-9 code $\in$ [290,319]).[9] Of those treating a mental health condition with prescription drugs, 94 percent report a SAD disorder (i.e., ICD-9 Codes 296, 300, 308, 309, or 311). In light of this fact, and to simplify our language, in what follows we refer to this "prescription drug use for mental health conditions" as "antidepressant use."

Columns 1–4 of Table 1 contain sample means for demographic, mental health, and labor market variables by treatment choice—antidepressants, therapy, both, or neither, respectively. Compared to those using antidepressants alone (column 1), individuals using therapy (columns 2 and 3) are younger, more likely to live in a metropolitan statistical area (MSA), less likely to be married, less likely to be white, more likely to have a college degree, and have smaller families. Those in therapy are also more likely to have public insurance and less likely to have private insurance than those taking antidepressants alone. Individuals receiving *any* type of mental health treatment have lower employment rates than those not receiving treatment. Those in treatment also have worse subjective mental health than those not receiving treatment, and those using *both* types of treatment have the worst subjective mental health, all of which suggests selection into treatment. Appendix Section A.I.3 further details the relationship between demographics, mental health, and treatment choices.

3.3 Key Empirical Patterns

3.3.1 Treatment Usage: Individuals in our sample are about three-to-five times more likely to use antidepressants than therapy in an interview period regardless of subjective mental health or reporting of a SAD disorder (see Appendix Table A.IV). To further investigate treatment use across demographics, we estimate a multinomial logit model, where the outcome categories are no treatment, antidepressants only, therapy only, and both antidepressants and

at various rates each year. Note that for prescription *refills*, the date the drug is obtained is not observable to researchers, but we know the interview round in which the prescription was refilled.

[9] Coding prescription drug use in this way has two implications. First, off-label drug use, which represents as much as 30 percent of antidepressant use (Wong et al., 2016), is intentionally ignored in our analysis, as these treatment regimens should have no impact on mental health. Second, we include some drugs outside the class of antidepressants (e.g., SSRIs, SNRIs, TCAs, etc.) and benzodiazepines (e.g., Alprazolam, Diazepam, etc.) that are commonly used to treat SAD disorders.

Table 1: Sample Means By Treatment Choice

	Antidepressants N=12,287	Therapy N=823	Both N=3,088	Neither N=191,915
Demographics				
Male	0.296	0.333	0.306	0.470
Age	43.399	41.620	42.792	40.818
Live in MSA	0.778	0.875	0.833	0.826
Married	0.558	0.450	0.381	0.660
Family Size	2.904	2.661	2.500	3.434
White (race)	0.855	0.796	0.797	0.766
Public Insurance	0.306	0.349	0.498	0.123
Private Insurance	0.626	0.575	0.463	0.661
Other HH Income	14,367	12,399	10,500	14,477
School & Work				
High School Grad.	0.575	0.482	0.542	0.529
College Grad.	0.226	0.346	0.246	0.255
Employed	0.588	0.611	0.416	0.783
Hourly Wage	23.310	27.583	24.789	23.426
Mental Health				
Subjective	3.110	2.911	2.470	4.025
SAD disorder	0.936	0.904	0.932	0.029
Any disorder	0.999	1.000	1.000	0.032

Notes: There are 208,113 observations (Sample C). The mean hourly wage excludes those who are not working. Subjective mental health is the respondent's own mental health assessment ranging from 1 (poor) to 5 (excellent). Stress, Anxiety, and Depression (SAD) indicators are based on ICD-9 codes 296, 300, 308, 309, and 311, whereas any (mental health) disorder pertains to all codes 290–319.

therapy (see Appendix Table A.VI for details). This exercise provides some insight into why individuals are unlikely to use therapy despite its significant benefits. For example, we find that insured individuals are the most likely to use treatment, which suggests that financial costs are a potential barrier to receiving care. Living outside an MSA and living in the South or Midwest are both positively associated with antidepressant use and negatively associated with therapy use, possibly suggesting cultural motivations (e.g., stigma) for treatment choice. Strikingly, based on the results from the multinomial logit, the predicted probability of using antidepressants is greater than the probability of using therapy for every observation in the sample. For 206,668 of 208,113 observations (99.3 percent), the predicted probability of using antidepressants *alone* is greater than the probability of using therapy. The consistent unpopularity of therapy across demographic groups and mental health categories suggests that a more robust choice model is needed to understand how patients choose mental health treatments. In what follows, we consider the costs and benefits of treatment that might inform decision making in such a model.

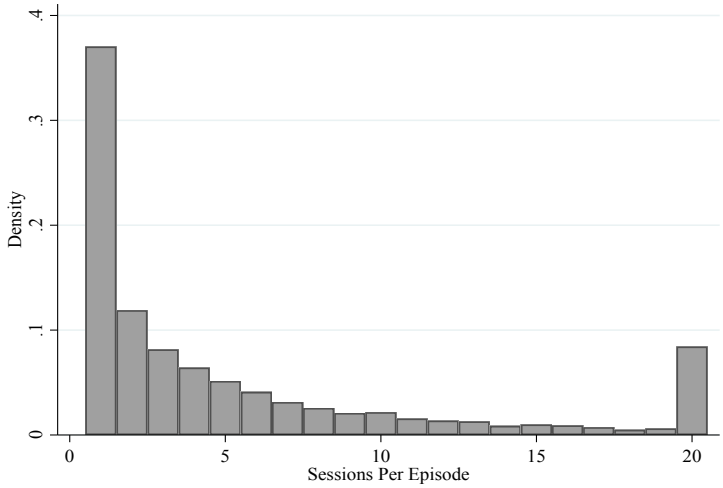
3.3.2 Costs of Treatment: High costs could explain why individuals rarely use therapy. The most obvious cost to consider is monetary. We document several price and cost-sharing patterns in our data. Like most healthcare, over time the total amount paid (from all sources) for both types of treatment has risen faster than inflation. That said, mental healthcare is somewhat unique in that the share of the price paid out-of-pocket (OOP) has fallen over time, due in part to the passage of both federal and state Mental Health Parity laws over this time period. Across all insurance types, a large fraction of therapy requires no OOP payment. For the insured, this is due to cost-sharing. For the uninsured, this is likely due to charity care and public mental health clinics. In the MEPS data, nearly 50 percent of therapy users receive treatment at no cost OOP. Conditional on paying anything, individuals pay $49 per session on average. (We report values in 2013 dollars throughout the paper unless stated otherwise.) These figures are 10 percent and $36, respectively, for a one-month supply of antidepressants. In Appendix Section A.I.4, we provide additional statistics on monetary costs across time and insurance status.[10] The dynamic model described in Section 4 features a six-month decision period, where the average number of one-month prescriptions filled by an antidepressant user is 5.7 and the average therapy patient attends 6.7 sessions; thus, the average out-of-pocket costs per period are approximately $168 for both antidepressants and therapy (s.d. of $468 and $559, respectively). While these figures do not condition on observables and ignore several selection concerns addressed in estimation, they at least suggest that the two treatments involve similar monetary costs, meaning monetary costs are unlikely to explain why patients rarely choose therapy in favor of antidepressants.

Another cost of therapy relates to uncertainty. A striking feature of the data is that a large share of the individuals using therapy attend very few sessions before stopping treatment. To show this, we define a therapy *treatment episode* as a consecutive sequence of therapy sessions occurring without a two-month gap in treatment. Figure 1 contains a histogram of the number of therapy sessions attended within each treatment episode. Notice that roughly *half* of these treatment episodes contain two or fewer sessions, meaning one or two sessions are attended without any sessions attended in the preceding or following two months. It is highly unlikely that such a course of treatment would be prescribed; rather, such behavior likely reflects what is called "discontinuation" in the psychology literature, where similarly high rates are reported (Wierzbicki and Pekarik, 1993; Swift and Greenberg, 2012). Many of those using therapy at some point during

[10]Unfortunately, the MEPS data contains no information on insurance plan characteristics or whether insurance coverage is related to employment. In reality, many insurance plans feature complicated cost-sharing arrangements that yield a non-linear schedule of OOP costs for the consumer. These costs can be a function of past spending (e.g., in the case of deductibles) or network. Research by Einav et al. (2013) and Cronin (2019) focuses on patient price sensitivity in light of this within-year cost heterogeneity. The limitations of our data requires simplifying assumptions. Namely, we assume consumers face constant OOP prices over the course of the year, which are estimated from the OOP prices observed in the data. We allow OOP prices to be influenced by public and private insurance coverage but assume that coverage is unrelated to employment.

the survey period (2,236 individuals) are observed to *only* use these very short treatment episodes (697 individuals or 31.2 percent). Relative to other therapy users, those choosing these short therapy episodes are more likely to be from the South, are less educated, are less likely to live in an MSA, and have better subjective mental health. Again, we can be sure that these are not individuals only visiting a mental health specialist to receive antidepressants; we only code individuals as receiving therapy if they explicitly state that they received therapy/counseling during their visit.

Figure 1: Therapy Sessions per Episode



Notes: This figure is produced using Sample C referenced in Table A.I. Sequences of therapy sessions are grouped into episodes according to a two-month gap rule described in the text. The figure displays the number of therapy sessions per treatment episode.

The psychology literature discusses possible causes of discontinuation. One possibility is that after the first few visits, inexperienced therapy patients learn that therapy is more costly or less beneficial than their prior belief. Another possibility is that patients visiting a new therapist experience poor "therapeutic alliance" (i.e., a bad match with their therapist), leading them to quit treatment (Ardito and Rabellino, 2011). Yet another possibility is that some patients find the first few sessions highly productive and no longer feel that they need for therapy. Likely, each of these narratives is the source of *some* discontinuation. While data constraints prevent us from precisely disentangling the relevance of each source,[11] our model permits heterogeneous treatment effects, learning about these treatment effects after the first therapy session, and nonlinear (utility) returns to additional sessions in an effort to explain discontinuation. We discuss these modeling choices further in Section 4.1. This discussion of therapy treatment episodes and discontinuation highlights a related issue about the discreteness of our data and the timing of treatment. For example, we might observe one session in interview period

[11]For example, we are not able to see the identity of the therapist, meaning we cannot distinguish new and existing matches. Moreover, we cannot observe therapy use prior to the survey period, so we cannot always distinguish new and existing patients.

two followed by nine sessions in period three, which is likely a single treatment episode that overlaps two interview periods. To appropriately count longer courses of therapy and to avoid over-counting discontinuation due to the timing of data collection, we develop an algorithm (described in Appendix Section A.III.2) that assigns therapy sessions to model periods based on surrounding sessions. We explore robustness to these decisions in Appendix Section A.IV.1.

Finally, therapy carries a significant time cost. According to the Mayo Clinic, therapy sessions are typically scheduled weekly or every other week for 45–60 minutes. Patients also must travel to and from treatment. While neither of these time costs are observable in the data, we do observe and model employment, which has a substantial effect on the available time an individual has for therapy.[12] One way to capture this cost in a structural model is to explicitly model preferences for leisure, which can be decreasing in work hours and some fixed therapy time cost (e.g., 2 hours, which would include a 50 minute session and 70 minutes of round-trip travel time). To provide some evidence that time costs are relevant in decision making, we reestimate the multinomial logit model discussed above, controlling for part- and full-time employment.[13] Consistent with these treatments having relevant time costs, we find that full-time workers are less likely than part-time workers, who are less likely than the unemployed, to use all types of treatment. Moreover, this relationship is strongest for individuals consuming both types of treatment. With that said, we do find that antidepressant use is also decreasing in employment, possibly suggesting that in addition to time costs, the well documented side-effects associated with antidepressants impair an individual's ability to work. In light of these findings, we decided not to measure leisure directly in the structural model; rather, we allow preferences for each treatment to vary with employment. This approach is more flexible in the sense that it captures time costs but also any other employment-related motivations for not using treatment.

3.3.3 Benefits of Treatment: It is possible that therapy is rarely used because it has limited benefits relative to antidepressants. To consider the benefits of treatment, we turn to the medical literature that has estimated the effects of these treatments on mental health using randomized controlled trials (RCTs). This literature reports standardized treatment effects (i.e., the mean effect is divided by the standard deviation of the outcome) making for relatively easy comparisons across research studies. In what follows, we focus on effect sizes estimated for depression and anxiety scales, such as the Hamilton Depression and Anxiety Rating Scales.

With respect to the effect of antidepressants on depression, results are remarkably consistent

[12]The 2018 National Survey on Drug Use and Health asks participants if they "needed mental health services but didn't get them," followed by "why." Among employed respondents with an unmet need, 24.4 percent report that the main reason for not receiving treatment is that they "do not have the time." Among those not employed, just 9.7 report "time" as the main constraint.

[13]Clearly, employment is endogenous in this simple model, as treatment could also impact the decision to work. As such, this exercise is only meant to be suggestive.

across the most highly cited medical research. Turner et al. (2008) performed a meta-analyses of both published and unpublished studies submitted to the Food and Drug Administration for review. Among published studies, they report a standardized effect of 0.37 with a 95 percent confidence interval from 0.33 to 0.41. Among unpublished studies, they report an effect of 0.15 with a 95 percent confidence interval from 0.08 to 0.22. A meta-analysis by Kirsch et al. (2008) finds an effect of 0.32 with a confidence interval from 0.25 to 0.40. A more recent study by Cipriani et al. (2018) found a similar effect at 0.3 with a 95 percent confidence interval from 0.26 to 0.34. Regarding the effects of anxiety medications, a meta-analysis by Mitte et al. (2005) documents effects of 0.32 and 0.30, respectively, for benzodiazepines and azapirones.

For therapy, there is a broader range of effect size estimates, but effects are consistently higher than for antidepressants.[14] Figure 2 shows effect sizes and confidence intervals for therapy from the medical literature. Gloaguen et al. (1998) directly compare the effects of therapy on depression with the effects of antidepressants. They find an effect of 0.82 for cognitive therapy relative to placebo and an effect of 0.38 for cognitive therapy relative to antidepressants. Ekers et al. (2008) report an effect size for behavioral therapy relative to placebo of 0.70 with a 95 percent confidence interval from 0.39 to 1.00.[15] Gould et al. (1997) consider studies that estimate the effects of therapy on generalized anxiety disorder and report a mean effect of 0.7 with a 95 percent confidence interval from 0.57 to 0.83. Hofmann and Smits (2008) also focus on those with generalized anxiety disorders and find an effect of 0.73 for anxiety measures and an effect of 0.45 for depression measures (among those with anxiety). Hofmann et al. (2012) review the literature on the effects of cognitive behavioral therapy (CBT) and report that papers tend to find that CBT has a "medium" (i.e., 0.5 to 0.8) effect for depression and a "medium" to "large" (i.e., 0.5 to above 0.8) effect for anxiety.

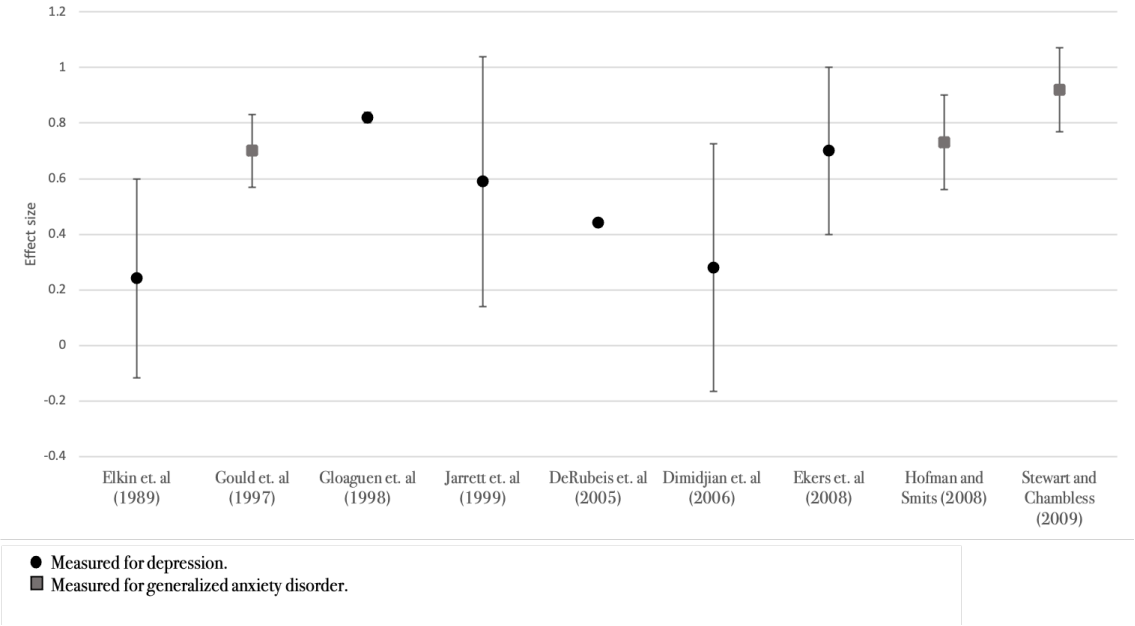
If we simply average the estimated effects across these literatures, we calculate an effect near 0.6 for therapy, which is within the 95 percent confidence interval of all studies (except one) reporting a confidence interval shown in Figure 2 and closely matches the effect estimated for depression severity at the 6-month and 12-month follow-ups in the field study by Baranov et al. (2020). For antidepressants, we obtain an effect near 0.3. As discussed below, we use these figures to calculate the baseline effect of any antidepressant use and the mean treatment effect for a therapy session.[16]

[14]The hypothesis that all forms of therapy have the same effect, sometimes called the "Dodo Bird Conjecture," is oftentimes not rejected (Wampold et al., 1997). Therefore, our literature search was focused on finding highly cited papers in the medical literature rather than focusing on specific forms of therapy.

[15]Cuijpers et al. (2010) argues that "high-quality" therapy studies typically yield smaller effect sizes. However, among the "high-quality" studies they describe that focus on a general adult population (Elkin et al., 1989; Jarrett et al., 1999; DeRubeis, Hollon, et al., 2005; Dimidjian, Hollon, et al., 2006), the average effect is roughly 0.4.

[16]A potential concern with each of the studies discussed above is that the treatment environment and patient characteristics in RCTs may not reflect the typical patient experience in the "real world." Effects estimated in controlled settings may not be externally valid if some therapists do not use empirically-validated

Figure 2: Estimated Effects of Therapy on Mental Health



Notes: This figure summarizes average standardized treatment effect sizes and 95 percent confidence intervals (when available) for a course of talk therapy for nine highly-cited individual clinical trials or meta analyses of clinical trails.

In addition to improving mental health, which individuals presumably value, it is also likely that improved mental health positively impacts labor market outcomes, meaning that treatment has important indirect benefits (see, e.g., Butikofer et al., 2020; Shapiro, 2020). Appendix Table A.V shows the results from ordinary least squares regressions of labor market outcomes on mental health, controlling for gender, age, race, marital status, whether or not one lives in an MSA, region, and education. The results indicate that better mental health is associated with significantly higher amounts of labor supply on both the extensive and intensive margins and also higher hourly wages. Mental health is clearly endogenous in these regressions, which is addressed by the structural model that follows.

4 Dynamic Model

We begin with an overview of key model features, then provide details on model specification.

4.1 Model Overview

We design the model to capture the key contemporaneous and dynamic tradeoffs associated with treatment and employment alternatives. Regarding treatment, the key benefit is improved

methods of treatment, providers that take part in efficacy trials are more skilled or better trained than the typical provider, or if patient discontinuation is common in less-controlled settings. We explore the sensitivity of our results to deviations from our baseline effects in Appendix Section A.IV.1. In Appendix Section A.I.5, we also discuss how the effects estimated in RCTs are generally consistent with those from effectiveness studies, which focus on estimating treatment effects in less-controlled settings.

future mental health, which may impact future utility through several labor and non-labor channels. Treatment is costly in that it reduces contemporaneous consumption via treatment prices, requires a time investment that reduces time that can be used for work or leisure, and may have direct negative effects on utility (e.g., physical discomfort, psychological discomfort due to stigma, etc.).[17] Regarding employment, key benefits are the receipt of wages, which allows for greater contemporaneous consumption, and the accumulation of experience, which may increase future wages. The primary cost of employment is reduced utility from lost leisure. The model also accounts for within-person changes in treatment and employment, which helps to capture barriers to transitions between treatment and employment states, such as search costs.

The therapy and antidepressant decisions differ in several important and related ways. First, we argue in Section 3.3.3 that the clinical literature finds substantially more variation in the effectiveness of therapy than antidepressants. Perhaps this is not surprising. Biologically identical antidepressant pills can be mass produced. Therapists, even those practicing the same form of therapy, will vary in personality, demographics, training, and skill. One might then expect the effectiveness of therapy to be heterogeneous. As such, the model allows the effect of therapy on mental health to vary across people and time. Second, Figure 1 suggests wide variation in the intensity of therapy use, conditional on going at all. Antidepressant use is more binary (people either use them throughout a period or not at all). As such, the model features both extensive and intensive margin therapy decisions, but only the former for antidepressants. Finally, because most individuals have never been to therapy, they are unlikely to know exactly how they will respond to it, even if they understand that therapy is highly effective on average. As such, we assume individuals enter each period knowing the distribution from which therapy treatment effects are drawn, but not their own specific draw. Upon attending a session, patients are assumed to learn their treatment effect for that period, prior to making an intensive margin decision. The combination of treatment effect heterogeneity and learning permits two plausible explanations of the empirical pattern documented in Section 3.3.2—that a large share of individuals going to therapy attend just a few sessions. Namely, (i) patients may learn that therapy is ineffective after trying it, leading to discontinuation or (ii) the first few therapy sessions may be highly effective, making additional sessions unnecessary.

Finally, we note that the model permits different forms of heterogeneity in key parameters, including permanent unobserved heterogeneity (often known as "unobserved types") in preferences, mental health, wages, and prices, along with time-varying unobserved heterogeneity

[17] We acknowledge the role that physicians play as advisors, and potential gatekeepers, in treatment choices (Arrow, 1963). Unfortunately, unlike Dickstein (2018) our data do not allow us to separately identify the incentives faced and choices made by patients and physicians. Thus, while we describe in this section an optimization problem solved by an individual, the true data generating process is likely determined by joint patient-physician optimization, and our treatment preference estimates will reflect that it is a joint decision.

in intra-period mental health shocks. These additions have theoretical implications for the variation we observe in choices and transitions, which is why we briefly mention them here—and will thus affect how we interpret estimates and results. However, the inclusion of unobserved heterogeneity is largely motivated by empirical issues specific to the data set and its relationship with our model, which affects estimation (e.g., initial conditions, measurement error, and negative selection into treatment). Thus, we relegate further discussion of heterogeneity to Section 5, which details estimation and identification, except for brief mentions related to model notation.

4.2 Model Specification

Consider an individual, $i = 1, ..., N$, who seeks to maximize expected lifetime utility in time period $t = 1, ..., T$. Each period, the individual receives utility, U_{it}, from consumption of a numeraire good, C_{it}; his or her mental health health status, M_{it}; and employment and treatment decisions, d_{it}^{rce}; where $e = 0, 1, 2$ denotes no, part-time, or full-time employment, respectively; $c = 0, ..., C$ denotes the number of therapy (i.e., "couch") sessions; and $r = 0, 1$ denotes antidepressant (i.e., "Rx") use.[18]

4.2.1 Preferences: Let vector $\boldsymbol{d_t}$ be comprised of $d_t^{rce}\ \forall\ r, c,$ and e, where $d_t^{r'c'e'} = 1$ when alternative (r', c', e') is chosen and zero otherwise. Flow utility from any decision d_t^{rce} can be expressed as

$$U_t^{rce} = \alpha_0 \frac{C_t^{1-\alpha_1} - 1}{1 - \alpha_1} + U(d_t^{rce}, \boldsymbol{d_{t-1}}, M_t, \boldsymbol{X_t}; \boldsymbol{\alpha}) + \mu_k(d_t^{rce}) + \epsilon_t^{rce} \tag{1}$$

where $\boldsymbol{X_t}$ measures a variety of exogenous, non-stochastic individual-specific observables.[19] The function $U(\cdot)$ is linear in parameters $\boldsymbol{\alpha}$ and includes interactions. The function $\mu_k(d_t^{rce})$ captures permanent, unobserved preferences for alternative (r, c, e) among type k individuals, while ϵ_t^{rce} captures any remaining unobserved, idiosyncratic preferences. We assume each individual has a permanent, unobserved type, k, which allows the unobserved determinants of choices and outcomes in the model to be correlated.

4.2.2 Budget Constraint: Gross household income in period t is calculated as $GY_t = \sum_{e=1}^{2} [d_t^{rce} * w_t^e * h_t^e] + I_t$. The bracketed term measures the individual's labor income, where w_t^e is wages from employment type e and h_t^e is the corresponding hours worked, the latter of which is held fixed across all individuals of employment type e. I_t measures all other

[18] The individual i subscript will be suppressed moving forward for notational simplicity. All variables are individual-specific unless otherwise stated.

[19] The variables in $\boldsymbol{X_t}$ are *exogenous* in the sense that other model variables are assumed to have no influence on their evolution. Moreover, we assume no structure on the random evolution of the variables. We abuse notation in using $\boldsymbol{X_t}$ as a generic vector of exogenous control variables that may include different sets of controls in different equations. In Appendix Section A.III.2, we provide a complete list of controls and indicate which controls are included in each equation.

household income sources and is assumed to evolve exogenously.

Numeraire consumption is calculated as disposable income minus treatment expenses, $C_t = D(GY_t, \boldsymbol{X_t}) - p_t^r * r_t - p_t^c * c_t$, where r_t equals one if antidepressants are used and c_t measures the number of therapy sessions attended. Because Equation 1 is non-linear in C_t, the marginal utility of treatment varies across the income distribution. As such, it is important that the model measures disposable income available for the purchase of healthcare. We approximate disposable income $D(\cdot)$ by adjusting GY_t for approximate total tax liability and housing expenses, as well as family size. $D(\cdot)$ is discussed in detail in Appendix Section A.II. We assume that the individual consumes all income in each period due to data constraints. It would be relatively straightforward to permit a savings decision with alternative data.

Wages w_t and prices p_t are stochastic and vary over time. Log wages in period t for part-time and full-time employment are expressed as $\log(w_t^e) = F(M_t, K_t, \boldsymbol{X_t}; \boldsymbol{\delta^e}) + \mu_k^{w,e} + \epsilon_t^{w,e}$ where $F(\cdot)$ is linear in parameters $\boldsymbol{\delta^e}$ and includes interactions (here and elsewhere), K_t is work experience entering period t, $\mu_k^{w,e}$ captures the permanent unobserved wage effects for individuals of type k, and $\epsilon_t^{w,e}$ is an idiosyncratic error. Out-of-pocket treatment prices for antidepressants ($x = r$) and therapy ($x = c$), somewhat complicated by the fact that insured individuals often face no out-of-pocket payments for medical care (e.g., see Appendix Section A.I.4), are written using the following latent variable structure

$$\begin{aligned} f_t^{*x} &= \boldsymbol{X_t}\boldsymbol{\eta^x} + \mu_k^{f,x} + \epsilon_t^{f,x} \\ p_t^{*x} &= \exp(\boldsymbol{X_t}\boldsymbol{\gamma^x} + \mu_k^{p,x} + \epsilon_t^{p,x}) \end{aligned} \tag{2}$$

where $p_t^x = p_t^{*x}$ if $f_t^{*x} > 0$ and zero otherwise. As before, $(\mu_k^{f,x}, \mu_k^{p,x})$ and $(\epsilon_t^{f,x}, \epsilon_t^{p,x})$ capture permanent and idiosyncratic unobserved heterogeneity, respectively.[20]

4.2.3 State Transitions: Work experience and mental health evolve over time as a function of individual employment and treatment decisions. Work experience, K_{t+1}, updates deterministically, increasing by one (one-half) each period that the individual decides to be employed full (part) time.

Self-reported mental health entering period $t+1$, M_{t+1}, measures integer values from 1

[20] As mentioned in footnote 10, we cannot observe whether privately held insurance is related to employment; thus, we allow prices to be influenced by insurance coverage (contained in $\boldsymbol{X_t}$), but assume coverage is unrelated to employment. A potential way to capture links between employment, insurance, and lower prices (e.g., changes in prices from job transitions that operate through changes in employer-provided coverage) would be to allow employment status to impact prices directly in Equation (2). The data suggest this approach would have no appreciable impact on results. For individuals continuously employed (unemployed) from period t to $t+1$, the period $t+1$ insurance rate is 81 (69) percent. The $t+1$ rate for individuals experiencing a job-separation between periods is 72 percent. And yet, the rate for individuals transitioning into employment is actually just 69 percent. Though it is true that many Americans receive insurance coverage through their employers, it does not appear to be the case that job transitions are highly predictive of overall coverage status.

(i.e., poor) to 5 (i.e., excellent). Define M^*_{t+1} as a latent, continuous measure of mental health expressed as $M^*_{t+1} = F(M_t, d^{rce}_t, \epsilon^{te}_t, \boldsymbol{X_t}; \boldsymbol{\nu}) + \mu^M_k + \psi_{jt} + \epsilon^M_{t+1}$. M_{t+1} is then assigned ordered, integer values when M^*_{t+1} falls between estimated thresholds. Therapy treatment effects, ϵ^{te}_t, vary across agents and time. The parameter ψ_{jt} captures unobserved heterogeneity in mental health transitions that varies across time and unobserved discrete types, j.

4.2.4 Dynamic Programming Problem: Given the above described choices, transitions, and payoffs, the timing of the model is as follows: an individual enters period t knowing $(M_t, K_t, \boldsymbol{X_t}, \boldsymbol{d_{t-1}})$ and their permanent unobserved type, k. Upon entry, he or she receives wage $(\epsilon^{w,e}_t)$, price $(\epsilon^{f,x}_t, \epsilon^{p,x}_t)$, and preference (ϵ^{rce}_t) draws. The individual also learns their time-varying unobserved type, j_t. At this time, the individual does *not* know how effective therapy would be for them, ϵ^{te}_t; rather, they know the distribution from which this treatment effect is drawn. With this information, which we call the state space, denoted $\boldsymbol{\Omega_t} = (M_t, K_t, \boldsymbol{X_t}, \boldsymbol{d_{t-1}}, k, w^e_t, p^x_t, j_t)$, the individual makes treatment and employment decisions, d^{rce}_t, to maximize expected lifetime utility. If the individual decides to visit a therapist, then he or she learns their true treatment effect, after which, they are able to reassess their employment and treatment decisions (though zero therapy is no longer possible), which fully determines contemporaneous utility, U^{rce}_t. Following this, experience, K_{t+1}, is updated and the individual receives a mental health shock, ϵ^M_{t+1}, before entering period $t+1$.

In each period, an individual selects alternative (r, c, e) to maximize his or her expected lifetime utility, V^{rce}, which can be written recursively as the sum of contemporaneous utility received at the time a decision is made and the expected, discounted present value of all future utility (Bellman, 1966):

$$V^{rce}(\boldsymbol{\Omega_t}, \epsilon^{rce}_t) = U^{rce}_t(\boldsymbol{\Omega_t}, \epsilon^{rce}_t) + \beta * \left[\int_{R^7_+} \sum_{m=1}^{5} P(M_{t+1} = m | \boldsymbol{\Omega_t}, d^{rce}_t, \epsilon^{te}_t) * \sum_{j=1}^{J} P(j_{t+1} = j | m) * EV(\boldsymbol{\Omega_{t+1}}) f(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon} \right]. \tag{3}$$

In this expression, β represents an exponential discount factor. The function $EV(\cdot)$ measures the expected future value of period t alternative (r, c, e) assuming optimal future behavior, sometimes called the "Emax" function. The value of $EV(\cdot)$ is influenced by several random variables that are not yet known by the agent when decisions are made in period t. First, the agent has not yet received their mental health shock entering period $t+1$; hence, the probability of all possible mental health transitions, $P(M_{t+1})$, must be accounted for, as M_{t+1} is contained in $\boldsymbol{\Omega_{t+1}}$. Second, prior to attending therapy, the agent does not know their therapy treatment effect, ϵ^{te}_t, so the agent must integrate over this distribution. Third, prior to making employment

and treatment choices in period $t+1$, the agent receives a total of six wage and price draws $(\epsilon_{t+1}^{w,1}, \epsilon_{t+1}^{w,2}, \epsilon_{t+1}^{f,r}, \epsilon_{t+1}^{f,c}, \epsilon_{t+1}^{p,r}, \epsilon_{t+1}^{p,c})$. To ease notation, these latter seven random variables are represented by $\boldsymbol{\epsilon}$ in Equation 3 and $f(\boldsymbol{\epsilon})$ represents the product of their probability density functions. Fourth, the individual learns their period $t+1$ unobserved time-varying type j_{t+1}, meaning the probability that they are any of the J types, $P(j_{t+1})$, which (we assume) is influenced by M_{t+1}, must also be accounted for. Finally, future preference shocks, ϵ_{t+1}^{rce}, are unknown in period t.

Traditionally, one would then write the Emax function as follows, where the expectation E_t is with respect to the preference shocks ϵ_{t+1}^{rce}.

$$EV(\boldsymbol{\Omega_{t+1}}) = E_t[\max_{rce} V^{rce}(\boldsymbol{\Omega_{t+1}}, \epsilon_{t+1}^{rce})]. \tag{4}$$

Our problem is complicated by the intra-period dynamics of the therapy decision; namely, if an agent selects any therapy next period, they learn ϵ_{t+1}^{te}, but then cannot "go back" to the no therapy state. As a result, the Emax function must be modified to account for this:

$$\begin{aligned} EV(\boldsymbol{\Omega_{t+1}}) = P(d_{t+1}^{r0e} = 1) * E_t \left[\max_{r0e} V^{r0e}(\boldsymbol{\Omega_{t+1}}, \epsilon_{t+1}^{r0e})\right] \\ + \left(1 - P(d_{t+1}^{r0e} = 1)\right) * E_t \left[\max_{rce} V^{rce}(\boldsymbol{\Omega_{t+1}}, \epsilon_{t+1}^{rce}) | c > 0\right]. \end{aligned} \tag{5}$$

While discussing estimation in Appendix Section A.III.1, we show that under several assumptions about the distribution of ϵ_{t+1}^{rce}, the probability of no therapy in period $t+1$, $P(d_{t+1}^{r0e} = 1)$, and both expectations above can be written as closed-form expressions of $\overline{V}^{rce}$, the deterministic part of V^{rce}.

Solving individual i's dynamic programming (DP) problem then proceeds as follows: starting in the terminal period T, we approximate the integrated Emax function using a non-stochastic, linear in parameters function $T(M_{T+1}, K_{T+1}, d_T^{rce}, age_{T+1}; \boldsymbol{\chi})$. Using Equation 3, we then calculate $\overline{V}^{rce}(\boldsymbol{\Omega_T})$ for every combination (r, c, e) in period T. We use these values in Equation 5 to determine $EV(\boldsymbol{\Omega_T})$, which is needed to calculate $V^{rce}(\boldsymbol{\Omega_{T-1}})$. We repeat this process backwards until $V^{rce}(\boldsymbol{\Omega_t})$ has been determined for every individual $i = 1, ..., N$, in every time period $t = 1, ..., T$, and for every combination (r, c, e).

5 Estimation and Identification

5.1 Estimation

The structural parameters of the dynamic model specified in Section 4 are estimated using the data described in Section 3 and a nested fixed-point algorithm (Rust, 1987). In the inner algorithm, the DP problem is solved for a given set of parameters. The outer algorithm uses the solution to calculate a likelihood function value and updates the parameter vector using

the Berndt et al. (1974) algorithm. A number of distributional assumptions are employed so that we can solve the DP problem and estimate the model using maximum likelihood. We discuss these assumptions and write the likelihood function in Appendix Section A.III.1.

We estimate the model using a 20 percent random subsample of Sample C to reduce the computation burden, which we refer to as Sample D (see Appendix Table A.I for descriptive statistics). Moreover, taking the model to data requires some decisions regarding the construction of therapy spells spanning multiple survey periods, computation of hours and the specification of the income process, use of control variables, and normalization of the numeraire consumption good. Details are found in Appendix Section A.III.2.

The number of permanent, K, and time-varying, J, unobserved types is determined in estimation. Our approach is influenced by the "upwards testing approach" recommended by Mroz (1999).[21] We alternate between adding first permanent and then time-varying types until the newest model fails a likelihood ratio test or fails to improve model fit. The latter occurs with the fourth permanent type; thus, we present results for the $(K = 3, J = 3)$ model.

5.2 Identification

There are four categories of model parameters: standard utility and transition/payoff functions (e.g., mental health, wage, price), permanent unobserved heterogeneity, mental health treatment effects, and time-varying unobserved heterogeneity. The identification of parameters in the first two categories is standard for this literature; thus, we provide a brief summary of key points here, but most commentary is relegated to appendices. Identification of these parameters relies on arguments from Magnac and Thesmar (2002), who show that, assuming rational expectations, a value for the discount factor, parametric distributions for error terms, and a utility function normalization, period-to-period transition probabilities identify parameters that govern state-to-state transitions and choice probabilities identify utility function parameters. We recount these arguments in more detail in Appendix Section A.III.3.

Our model also permits permanent unobserved heterogeneity in an effort to solve several identification and measurement error challenges in estimation.[22] Magnac and Thesmar (2002) estab-

[21] Working with just one layer of unobserved heterogeneity, Mroz (1999) recommends determining the number of types by first estimating all model parameters assuming one unobserved type, which produces a likelihood function value, LF_1 and a set of maximizing parameters, $\widehat{\mathbf{\Theta}}_\mathbf{1}$. The model is then re-estimated with two unobserved types using the previously estimated parameters, $\widehat{\mathbf{\Theta}}_\mathbf{1}$, as starting values, which produces a new likelihood function value, LF_2, and a new set of maximizing parameters, $\widehat{\mathbf{\Theta}}_\mathbf{2}$. A likelihood ratio (LR) test is used to determine whether the additional unobserved type led to a significant improvement in the likelihood function. Mroz suggests continuing to add types until the likelihood function does not improve.

[22] For example, permanent unobserved heterogeneity allows us to extract permanent mental health "types" that remain similar even though individuals may exhibit fluctuations in their subjective well-being due to a bad day. Poor initial mental health is then allowed to increase the probability that an individual is of the "low health type" in future periods, which addresses the endogeneity of the initial health state.

lish non-parametric identification in such models with additional restrictions; for example, they impose that the permanent component is additive separability and does not affect terminal value functions. Though we impose these restrictions, identification in our parametric setting is less demanding because, for example, each choice-state pair does not have a different utility parameter. We assume a finite number of types, which requires a normalization and data on repeated choices and outcomes over time (i.e., panel data on treatment and employment choices, mental health transitions, and price and wage outcomes). We discuss this further in Appendix Section A.III.4.

A key challenge to identification of our model is negative selection into treatment. Our approach for addressing this challenge includes the use of outside data from clinical trials, permitting time-varying unobserved heterogeneity in mental health, and allowing therapy treatment effects to vary across individuals. In what follows, we provide evidence of selection in the data. We then discuss how our approach aids in the credible identification of utility parameters, as well as how each feature of our approach itself is identified from the data.

Negative selection into treatment is best illustrated in Table 2. Holding M_t constant, those using either treatment in period t appear *worse off* entering period $t+1$ than those not going to treatment. The explanation likely relates to the timing of data collection. On average, M_t and M_{t+1} are reported six months apart. Treatment decisions, d_t^{rce}, are made between these reports. If, for example, mental health shocks occur over the course of the period and those selecting into treatment are those experiencing large negative shocks, then this exact data pattern would emerge. Put differently, mental health outcomes for those not being treated in Table 2 (columns 1 and 3) provide an inadequate control group to compare those being treated (columns 2 and 4), should they have opted out of treatment.

Table 2: Mental Health Transitions by Treatment Choice

	Mental Health at $t+1$			
	Antidep. Use		Therapy Use	
	No	Yes	No	Yes
$M_t = 5$	4.547	4.089	4.539	3.608
$M_t = 4$	4.008	3.607	3.992	3.360
$M_t = 3$	3.469	3.018	3.438	2.840
$M_t = 2$	2.782	2.333	2.695	2.226
$M_t = 1$	2.095	1.691	1.993	1.558

Notes: The figures presented are calculated from Sample C. Subjective Mental Health is the respondent's subjective assessment of own mental health and ranges from 1 (poor) to 5 (excellent). The first column contains mean subjective mental health entering period $t+1$ among those *not* using antidepressants in period t. Others columns are similarly defined.

We address negative selection into treatment first by leveraging data on treatment effects

established in well-identified settings; namely, the clinical trials literature.[23] Consider Section 3.3.3, where we argue that the average standardized impact of antidepressants and therapy on mental health are approximately 0.3 and 0.6, respectively. Standardized effects for an outcome, Y, are measured as $(Y_1 - Y_0)/SD(Y)$, where Y_1 measures the outcome for the treated group and Y_0 for the untreated group. We utilize these estimates within our model, despite having a different measure of mental health, by scaling the clinical estimates by the standard deviation of the latent mental health variable in Section 4.2.3.[24] Specifically, we calculate the antidepressant treatment effect as $\nu_{0,1} = 0.3 * SD(M^*(\boldsymbol{\Theta}_s))$. In Section 4, we discuss our motives for allowing heterogeneity in therapy treatment effects. We thus assume that the average standardized effect size determines the *mean* effect of a full course of therapy (12 sessions) or that $\nu_{0,2} = 0.6 * SD(M^*(\boldsymbol{\Theta}_s))/12$ is the effectiveness of one session. We then estimate the variance of the therapy treatment effect, $\nu_{0,3}$. We explore the robustness of these assumptions in Appendix Section A.IV.1.

Using treatment effects from the clinical literature ensures well-identified mean treatment effects in our model. That said, because the underlying data suffer from negative selection into treatment, imposing these treatment effects could induce bias into estimates of utility parameters. For example, use of therapy by a healthy individual who suffered an unobserved negative shock could lead to an upwardly biased estimate of the value of excellent mental health. Moreover, some counterfactual policy analysis considers dynamic impacts, which is more credible if our estimated model is able to reproduce dynamic mental health patterns.[25] To address these concerns, we first allow time-varying unobserved heterogeneity (TVUH) that approximates the intra-period health shocks discussed above. That is, we assume at the start of each period, t, each individual, i, learns they are of type $j = 1, ..., J$, where each type receives a different mental health shock ψ_j. The estimation procedure seeks to determine (i) the number of unobserved types in the population, J; (ii) the share of the population that is described by each type, ι_j; and (iii) the impact that each unobserved type j has on the mental health transition, ψ_j,

[23] Selection problems are not entirely unique in comparable structural work and can at times be addressed with robust error correlation and exclusion restrictions, strategies that we explored in depth. Briefly, we used county-level variation in psychiatrists per capita, mental health clinics per capita, and low-price Walmart pharmacy locations, as well as state-level variation in mental health parity laws, as plausibly exogenous shifters of treatment choices. Within the structural model, we repeatedly found negative treatment effects regardless of error structure. Using more traditional reduced-form approaches to estimate treatment effects with these variables as instruments, we recovered positive treatment effects for some sub-samples, but the instrument set was consistently too weak for credible estimation. Additional findings are available upon request from the corresponding author.

[24] As M_t is modeled as an ordered logit, the variance of the error term ϵ_t^M is fixed; however, the explained variance of M_t^* is influenced by the estimated parameters $\boldsymbol{\nu_0}$ and cut points (ν_1, ν_2, ν_3) in particular, which are a function of the data. Because the standard deviation of M_t^* is a function of model parameters, it must be calculated at every iteration of the parameter set $\boldsymbol{\Theta}_s$.

[25] Though we estimate the impact of M_t on M_{t+1}, our likelihood function is not designed to capture *all* mental health dynamics, for instance, mean mental health three periods post-treatment. The unobserved heterogeneity we describe can aid in reproducing these dynamics, but this is not critical for identification.

for $j = 1, ..., J$. This structure allows that there are some types j that receive poor mental health shocks prior to the treatment decision and, therefore, select into treatment. Furthermore, as bad shocks are likely more common among those already in poor health, we allow the probability an individual i is of type j in a period t to be influenced by their mental health entering the period M_t.

Allowing a distribution of treatment effects also gives the model some capacity to explain the patterns in Table 2, e.g., by assigning some patients negative treatment effects. More importantly, by allowing heterogeneous treatment effects, we avoid erroneously attributing variation in treatment effectiveness to utility parameters, e.g., estimating an upwardly biased distaste for therapy that arises from a small or negative treatment effect. Absent information on providers, we are unable to use variation in observables, such as treatment modalities or therapists' demographic information, training, or experience treating patients. To capture variation, we allow for a distribution of treatment effects, the mean of which is taken from clinical literature as described above, and we restrict the distribution to be symmetric around the mean. We then estimate the variance.

Identification arguments for both treatment effect heterogeneity and TVUH are straightforward in a setting where mean treatment effects are estimated. Treatment effect heterogeneity is identified by differences in the variation of M_{t+1} across those using therapy and not, all else equal. TVUH can be viewed as a simple random effect, in the traditional panel data sense, that is specific to the time dimension of the panel.[26] Similarly, allowing M_t to influence ι_j is akin to estimating the impact of M_t on M_{t+1} as a random effect. Identification requires that the mental health transition be observed in multiple periods, that the distribution of TVUH be the same in each period, that type probabilities sum to one (i.e., $\sum_{j=1}^{J} \iota_j = 1$), and the normalization that $\psi_1 = 0$.

Imposing mean treatment effects does not change these formal identification arguments, but choice patterns will influence the magnitude of associated treatment effect and TVUH parameters differently than if mean treatment effects were estimated. For example, that observed mental health does not improve for many people using either type of treatment, despite positive mean treatment effects, could be explained by a subset of the population (i.e., a particular TVUH type) that receives a bad pre-treatment mental health shock selecting into treatment. If such a type best fits the pattern of treatment choices and health transitions, the estimation routine will generate it. For therapy, attending few sessions with no (or negative) mental health improvement, for example, would be well explained by a negative therapy draw in our model; thus, discontinuation influences the treatment effect variance parameter in a way that makes such draws more likely.

[26]Readers may be more accustomed to thinking of random effects specific to the individual (or panel) dimension of the data, which we model as well, using permanent unobserved heterogeneity as is described in Appendix Section A.III.4. Yang et al. (2009) and Darden et al. (2018) also use both permanent and time-varying discrete random effects to allow error correlation in systems of dynamic equations, as does Adda et al. (2022) in a dynamic structural model.

6 Results

6.1 Parameter Estimates

Parameter estimates can be found in Appendix Section A.IV. Tables contain estimates for models without ($K = 1, J = 1$) and with ($K = 3, J = 3$) unobserved heterogeneity. Permanent unobserved heterogeneity parameters ($\boldsymbol{\mu}, \boldsymbol{\theta}$) are reported together in Appendix Table A.IX. Parameter signs meet *a priori* expectations with few exceptions. Below, we briefly describe some parameter estimates for the model *with* unobserved heterogeneity, our preferred specification.

Regarding medical treatments: (i) Holding future mental health constant, individuals derive disutility from both antidepressants and therapy. (ii) For the baseline individual,[27] a single therapy session yields more disutility ($\alpha_{1,0}+\alpha_{1,8}+\alpha_{1,9}$) than six months worth of antidepressant use ($\alpha_{2,0}$); additional therapy sessions increase disutility at an increasing rate. (iii) Any *past* therapy use lowers the disutility of any *current* therapy use, consistent with search costs. (iv) *Past* antidepressant use lowers the disutility of *current* therapy use, indicating dynamic complementarities. The same effects exist for antidepressant use,[28] though past antidepressant use nearly removes the contemporaneous disutility of current antidepressant use, reflecting the ease with which prescriptions can be refilled within the American healthcare system. (v) The contemporaneous disutility of any treatment is largest for full-time employees ($\alpha_{1,4}, \alpha_{2,4}$), followed by part-time employees ($\alpha_{1,3}, \alpha_{2,3}$). For antidepressants, this likely reflects difficulty working while experiencing well documented side-effects (e.g., nausea, tiredness, headaches, etc.). For therapy, this likely reflects time costs, but this interpretation requires some nuance, as the disutility of therapy while working increases more slowly in the number of sessions attended for the employed ($\alpha_{1,8}, \alpha_{1,9}, \alpha_{1,10}, \alpha_{1,11}$) than the unemployed ($\alpha_{1,8}, \alpha_{1,9}$).[29]

[27] Baseline here refers to a male, living outside an MSA, who is 26 years of age, unemployed, and did not consume either type of treatment in the prior period.

[28] That the treatments act as dynamic complements is sensible in this setting. In many cases, patients first reveal depressive symptoms to their general practitioner who may prescribe antidepressants or refer the patient to a specialist who can administer/prescribe either type of treatment. Using one type of treatment reveals a willingness to treat symptoms and opens a dialogue with medical professionals, each of which facilitates greater use of alternative treatments in the future.

[29] There are several possible explanations. First, as is also true for the unemployed, therapy starting costs ($\alpha_{1,0}$) are more substantial than continuing costs ($\alpha_{1,0}+\alpha_{1,1}$). The employment disutility we uncover may reflect a time or effort cost to find a therapist or to shuffle one's work schedule around therapy appointments. Another possibility is that this employment disutility captures workplace stigma related to therapy. If an agent's employer is self-insuring (something we cannot observe), then the employer will observe the agent's therapy attendance, which may impact the possibility of promotion. Each of these interpretations can be thought of us startup costs particular to employed versus unemployed people. After incurring startup costs, employed individuals face a slightly lower per-session cost compared to unemployed individuals. A possible explanation is that an employed individual may appreciate (or find relatively less costly) the break from work to spend time in therapy to focus on themselves and their private issues. To be clear, this does not change the fact that additional sessions imply a cost. Rather, the per-session cost is slightly lower for the employed versus the unemployed. The counterfactual we consider in Section 6.4.3 sets ($\alpha_{1,3}, \alpha_{1,4}, \alpha_{1,10}, \alpha_{1,11}$) to zero and could be interpreted as the impact of bringing a therapist

Regarding employment: (i) Holding earnings constant, individuals derive disutility from part-time employment, but greater disutility from full-time employment, indicating the value of leisure. (ii) Past employment lessens employment disutility, indicating switching costs. (iii) The disutility from employment increases as mental health worsens.[30] (iv) We uncover several general employment patterns that have been documented elsewhere: women derive greater disutility from full-time work than men; employment is increasing, and then decreasing in age; full-time employment is increasing (decreasing) in family size for men (women).

Regarding mental health: (i) Utility is increasing in mental health. (ii) Poor mental health in the past lowers current mental health; M_t is a stock. (iii) Wages decrease as mental health worsens, though these effects are very small and diminishing. (iv) Therapy has heterogeneous treatment effects such that 35 percent of individual-periods have treatment effects that are twice the mean clinical level and another 35 percent are negative.

Regarding numeraire consumption: Utility is increasing in consumption at a decreasing rate; the coefficient of relative risk aversion (CRRA) is 0.24. This estimate is quantitatively similar to the estimates of Cronin (2019) and Blau and Gilleskie (2008), both near 0.1, but smaller than estimates from French (2005) and French and Jones (2011), both above 2.[31]

6.2 Model Fit

To assess the model's ability to explain unique features of the data, we use the model and estimated parameters to simulate new data sets and compare key moments of the observed and simulated data. Simulated data are constructed by sampling from the joint error distribution,

into the workplace. Setting all parameters to zero may underestimate the benefit of such a policy, since doing so only lowers the startup costs but may not affect the relatively low per-session cost, in which case the last two parameters should remain unchanged. However, results remain about the same if we only set $(\alpha_{1,3}, \alpha_{1,4})$ to zero.

[30] De Quidt and Haushofer (2016) posit that depression acts as an exogenous shock to an agent's beliefs about the returns to effort, whereby pessimistic beliefs may reduce expected productivity for a fixed amount of effort or the expected cost of effort, ultimately reducing labor supply. Note that we find virtually no impact of mental health on wages; thus, disutility from employment increasing as mental health worsens could be interpreted either as a bias in beliefs about own productivity or rational expected costs of effort.

[31] The latter two papers referenced above include a savings decision, while the former two do not. Recall that our model omits savings due to data limitations, which has several implications for our results, including the estimation and interpretation of the CRRA parameter. For example, in a model with savings, the marginal utility of consumption is estimated at lower levels of consumption. Thus, the "flatness" in the utility-consumption profile that we estimate at high levels of consumption would be estimated at lower levels in the alternative model, meaning a larger CRRA parameter. Moreover, the CRRA parameter does not measure the elasticity of intertemporal substitution in a model without savings. Other parameters could be similarly affected by omitting savings. For example, the employment decision involves weighing the disutility of work against consumption gains. Without savings, the relative consumption gains associated with working are too large, meaning the counterbalancing disutility of work that we estimate may also be too large. A smaller disutility of work might mean that agents are more responsive to improvements in mental health in terms of their willingness to supply labor. That said, under the variety of parameter combinations and modeling assumptions we have explored, our main results remain intact: labor supply remains similarly inelastic to changes in mental health, and use of mental health treatment remains similarly inelastic to price differences.

permanent and time-varying unobserved heterogeneity distributions, therapy treatment effect distribution, and the parameter covariance matrix 50 times for each individual, then forward simulating from the observed initial conditions of our estimation sample. Table 3 shows that the model matches the data on most key moments both with and without unobserved heterogeneity, including mean treatment and employment choices, as well as mental health, wage, and price distributions. Figure 3 shows that we also match these moments for men and women separately across the age distribution. For most of these comparisons, the model without unobserved heterogeneity does nearly as well as the model with it.

Without unobserved heterogeneity, the model struggles to match mean treatment and employment levels across the mental health distribution. The second column of Table 3 shows that while simulated treatment is decreasing and employment increasing in mental health, as is true in the data, we under-predict both types of treatment and over-predict full-time employment when individuals are in the lowest two mental health states; the opposite is true in the best mental health states. The disparity in treatment patterns is the product of the negative selection problem we describe in Section 5: those in the MEPS who receive treatment in period t have worse average mental health in period $t+1$ than those not receiving treatment, even when conditioning on mental health in period t (see Table 2). And yet, we impose (positive) mean treatment effects from the clinical literature. With these treatment effects, it is difficult for our model to produce the empirical fact that most treatment is consumed by very sick individuals, because the model says these individuals should be better in the following period, meaning they no longer need treatment, but the data suggest otherwise.[32]

Section 5 describes how both permanent and time-varying unobserved heterogeneity aid in modeling selection. The last two columns of Table 3, show how integrating unobserved heterogeneity improves the model's ability to fit both treatment choices and full-time employment across the mental health distribution. For example, therapy, antidepressant, and full-time employment rates in the lowest mental health state in the data are 22, 51, and 12 percent, respectively. The simulated rates go from 13, 25, and 28 without unobserved heterogeneity to 16, 31, and 17 with it; not a perfect fit, but a significant improvement.[33]

[32]Without unobserved heterogeneity, treatment and employment moments across the mental health distribution match the data nearly perfectly in the first simulation period. The same is true in future periods if dynamic state space variables ($M_t, K_t, \boldsymbol{d_{t-1}}$) are not update in simulation; see Table A.XIV. These findings indicate that without unobserved heterogeneity, there is dynamic correlation between mental health, treatment, and employment that the model cannot reproduce—sign of unmodeled correlation in the unobserved determinants of these variables. The role of unobserved heterogeneity is to allow such correlation.

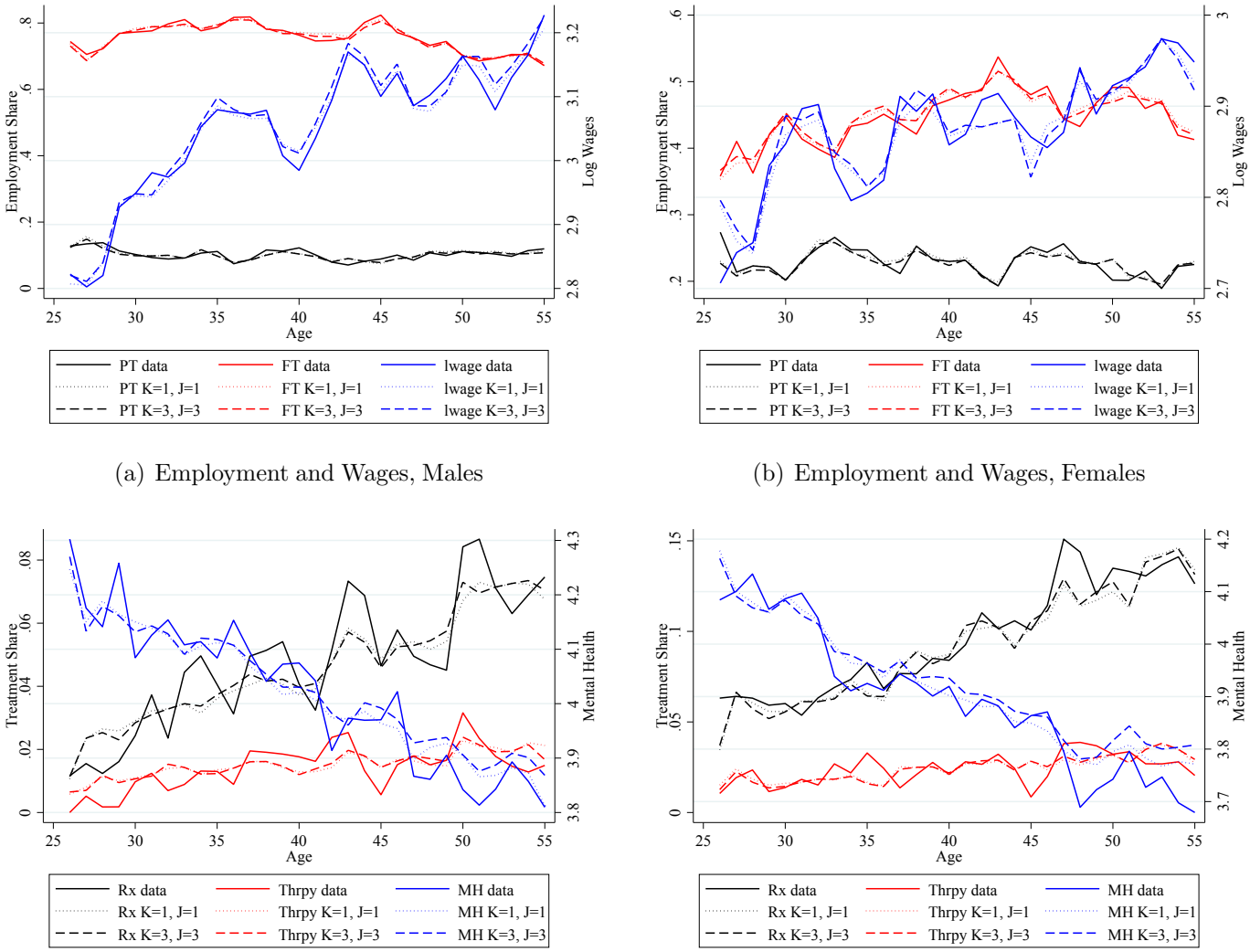
[33]This is also the main dimension upon which time-varying unobserved heterogeneity improves model fit. Were one to only include permanent unobserved heterogeneity, the figures are 12, 27, and 25. In other words, time-varying unobserved heterogeneity makes it easier for the model to explain how individuals with poor health in period t can be the most likely to go to treatment, yet remain in poor health in period $t+1$.

Table 3: Model Fit

Variable	Est. Sample Mean	Sim, K=1, J=1 Mean	Sim, K=1, J=1 S.E.	Sim, K=3, J=3 Mean	Sim, K=3, J=3 S.E.
Treatment					
Any therapy	0.0202	0.0204	0.0003	0.0205	0.0003
if $c_{t-1} = 1$	0.4651	0.4401	0.0043	0.4399	0.0035
if $r_{t-1} = 1$	0.1946	0.1914	0.0027	0.1947	0.0026
if $M_{t-1} = 5$	0.0029	0.0150	0.0003	0.0129	0.0002
if $M_{t-1} = 4$	0.0085	0.0143	0.0002	0.0155	0.0002
if $M_{t-1} = 3$	0.0256	0.0219	0.0003	0.0230	0.0003
if $M_{t-1} = 2$	0.1164	0.0541	0.0008	0.0613	0.0010
if $M_{t-1} = 1$	0.2215	0.1258	0.0017	0.1574	0.0030
Share with $p_t^c = 0$	0.5106	0.4552	0.0032	0.4784	0.0029
$p_t^c \| p_t^c > 0$	41.2218	47.4623	0.7589	44.9732	0.8495
Sessions $\| c_t \neq 0$	6.0731	5.3868	0.0322	5.5081	0.0318
Any Rx	0.0752	0.0723	0.0003	0.0727	0.0004
if $c_{t-1} = 1$	0.7136	0.6737	0.0039	0.6780	0.0037
if $r_{t-1} = 1$	0.7483	0.7205	0.0018	0.7249	0.0017
if $M_{t-1} = 5$	0.0224	0.0545	0.0004	0.0493	0.0004
if $M_{t-1} = 4$	0.0500	0.0588	0.0004	0.0610	0.0005
if $M_{t-1} = 3$	0.1073	0.0845	0.0004	0.0869	0.0005
if $M_{t-1} = 2$	0.2956	0.1620	0.0011	0.1879	0.0013
if $M_{t-1} = 1$	0.5099	0.2527	0.0025	0.3109	0.0035
Share with $p_t^r = 0$	0.1142	0.1044	0.0009	0.1103	0.0010
$p_t^c = 0 \| p_t^r > 0$	163.7217	179.7108	1.1977	181.0007	1.4402
Employment					
PT	0.1696	0.1702	0.0004	0.1684	0.0004
if $PT_{t-1} = 1$	0.8974	0.8966	0.0007	0.8936	0.0008
if $FT_{t-1} = 1$	0.0085	0.0041	0.0001	0.0087	0.0001
if $M_{t-1} = 5$	0.1696	0.1712	0.0006	0.1722	0.0006
if $M_{t-1} = 4$	0.1740	0.1744	0.0005	0.1721	0.0005
if $M_{t-1} = 3$	0.1785	0.1729	0.0007	0.1689	0.0006
if $M_{t-1} = 2$	0.1327	0.1473	0.0012	0.1344	0.0013
if $M_{t-1} = 1$	0.0819	0.1093	0.0022	0.0874	0.0025
Mean: W_t^1	21.1150	20.3415	0.0277	21.2345	0.0473
SD: W_t^1	17.7599	16.5463	0.0533	19.3601	0.1103
FT	0.5918	0.5942	0.0004	0.5933	0.0005
if $PT_{t-1} = 1$	0.0377	0.0137	0.0002	0.0276	0.0004
if $FT_{t-1} = 1$	0.9572	0.9556	0.0002	0.9555	0.0003
if $M_{t-1} = 5$	0.6591	0.6437	0.0006	0.6507	0.0006
if $M_{t-1} = 4$	0.6349	0.6184	0.0007	0.6220	0.0007
if $M_{t-1} = 3$	0.5281	0.5451	0.0009	0.5408	0.0009
if $M_{t-1} = 2$	0.3282	0.4288	0.0018	0.3610	0.0014
if $M_{t-1} = 1$	0.1168	0.2847	0.0027	0.1671	0.0023
Mean: W_t^0	24.3149	24.4731	0.0121	25.2139	0.0209
SD: W_t^0	15.6820	16.2158	0.0162	18.1166	0.0397
Mental Health					
$MH_t = 5$	0.3781	0.3851	0.0005	0.3775	0.0004
$MH_t = 4$	0.3029	0.2980	0.0004	0.3056	0.0004
$MH_t = 3$	0.2447	0.2429	0.0004	0.2534	0.0004
$MH_t = 2$	0.0586	0.0592	0.0002	0.0513	0.0002
$MH_t = 1$	0.0157	0.0148	0.0001	0.0123	0.0001

Notes: The simulated data are constructed using the process described in Section 6.2. All moments are calculated over all four simulation periods. We present model fit for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Figure 3: Life Cycle Model Fit



(a) Employment and Wages, Males

(b) Employment and Wages, Females

(c) Treatment and Mental Health, Males

(d) Treatment and Mental Health, Females

Notes: The simulated data are constructed using the process described in Section 6.2. All moments are calculated over all four simulation periods. We present model fit for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Appendix Tables A.XV and A.XVII characterize both the permanent and time-varying types via simulation, clarifying how model fit improves. Among the permanent types described in Table A.XV, $k = 2$, which represents 24 percent of the population, has the worst mental health, highest treatment levels, and lowest employment rates. Recovering this pattern offers one explanation for the contradiction described above—those who consistently select into treatment (and out of employment) are also consistently more likely to experience negative mental health shocks (e.g., due to chronic illness) and so, we the econometricians, are less likely to observe mental health improvements for them, despite their use of treatment. Table A.XVII

further illustrates the role that time-varying unobserved heterogeneity plays. Individuals with poor mental health states entering period t, M_t, are most likely to be unobserved type $j_t = 2$, which receives the largest negative mental health shock in period t prior to treatment (see ψ_2 in Table A.XI). Table A.XVII shows that conditional on, M_t, treatment rates are highest among $j_t = 2$, and yet, M_{t+1} is lowest, regardless of treatment. In other words, individuals entering a period with poor mental health are most likely to receive large negative mental health shocks, leading them to select into treatment, but this yields little *observable* mental health improvements, precisely because of the large negative shocks that elicited treatment.

6.3 Robustness

Modeling treatment choices requires many decisions. We have experimented, estimating models with a number of alternative assumptions: One, we drop individuals that report "excellent" (i.e., $M_t = 5$) mental health in all five sample periods, as these individuals may never consider treatment. Two, we drop individuals reporting non-SAD mental health conditions, to test whether those with more severe/rare mental health conditions carry disproportionate weight in our findings. Three, we ignore the algorithm used in Appendix Section A.III.2 to assign therapy sessions to interview periods, instead leaving all sessions in the period they are reported in. Four, we test alternative assumptions about treatment effects, including: (i) allowing no heterogeneity in the therapy treatment effect, (ii) halving and doubling mean clinical treatment effects, and (iii) allowing antidepressant treatment effects to be a function of lagged mental health, which is supported by some of the clinical literature.

In general, our results remain robust to these assumptions; all models yield similar parameter estimates, unobserved types, and model fit comparisons. As such, we leave a more robust description of each alternative specification and findings to Appendix Section A.IV.1. This section also contains a discussion of the limitations of our work.[34]

6.4 Counterfactuals

We use the estimated model to perform a series of counterfactuals. Our first counterfactual assesses the personal and economic value of improvements to mental health. We establish that individuals indeed value mental health improvements. Our second and third counterfactuals focus on the extent to which mental health can be improved via policies that encourage therapy use. Just as in our analysis of model fit, all counterfactuals take the initial conditions of our esti-

[34]For example, we assume that individuals have rational expectations and full information, which implies that when making treatment decisions, they understand average treatment effects as reported in the medical literature and act accordingly. It is possible that individuals make treatment decisions with incorrect expectations. Additionally, our model assumes patients have full agency in their medical decision-making. However, it is likely that doctors advise patients on these decisions, meaning estimated preferences likely reflect those of patients and their physicians. Finally, we do not simulate long-run effects because our model, which includes permanent unobserved heterogeneity, is estimated using just two years of data.

mation sample, which is nationally representative, as the starting point for forward simulation. Unless stated otherwise, each simulation is conducted using the same process described in Section 6.2. Throughout our analysis, we consider separately the full sample used in estimation, as well as a sicker sample, which is comprised of individuals with initial observed mental health less than 4 (or worse than "good"). The latter sample contains about 30 percent of the population. We consider the sick sample separately because these individuals are most in need of mental health treatment and, therefore, their response to policy is most relevant for policy-makers.

6.4.1 The Value of Mental Health: To measure the personal and economic value of improvements to mental health, we begin by simulating behaviors and outcomes over a two-year period while assuming that an individual's mental health cannot fall below the baseline sample mean, which is $M_t = 4$. To fix ideas, this would be like inventing a new, costless treatment (i.e., no financial cost, no time cost, no utility cost, etc.) guaranteed to return an individual's mental health to the population average; as such, those below the average would always take the treatment and those above never would. We then compare this counterfactual data to the baseline simulated data, which assumes no new technology.

For the full sample, we find that average mental health per-period improves by 11 percent, employment improves by two percent (part-time and full-time), while wages fall by one percent. As mental health has almost no direct effect on wages (see Appendix Table A.XII), the small, negative simulated impact of improved mental health on wages is driven by negative selection into employment—those newly employed as a result of improved mental health are low-wage workers. For the sicker sample defined above, average mental health over the next two years improves by 22 percent, employment increases by 5.5 percent, and wages fall by two percent. To understand these findings, consider an individual with initial mental health equal to 3 (or "fair") in period t—the healthiest of individuals in the sick sample. This new technology would raise their mental health to 4 (or "good") in period t; a (4-3)/3 = 33 percent one-period improvement. That the average improvement in mental health for the sick sample over the two-year period is just 22 percent illustrates that many individuals with low initial mental health see improvements over time, even without the new technology.

We next consider what individuals would be willing to pay for access to such a technology. In addition to providing a single summary measure of the total benefit to society of the technology, because mental health is *the* mechanism through which treatment can yield welfare improvements, this counterfactual establishes a sort of upper bound on the possible gains from treatment inducing policies, the subject of the next two subsections. Note that the new mental health improving technology has the potential to be very valuable for some. Furthermore, large WTP calculations involve numeraire consumption values that are potentially outside the range

of values considered in estimation; recall, we measure consumption at six month intervals and there are no savings. As such, we measure WTP for the new technology in just one (the first) simulation period, requiring a new simulation that assumes the technology no longer exists in the second period. This new exercise requires that we take a stand on the nature of the technology; specifically, whether the technology affects the flow of mental health, or flow and stock. If the latter, then the technology alters mental health entering period $t + 1$, increasing the likelihood of better mental health further into the future, which the agent should be willing to pay even more for. We ultimately provide both calculations.

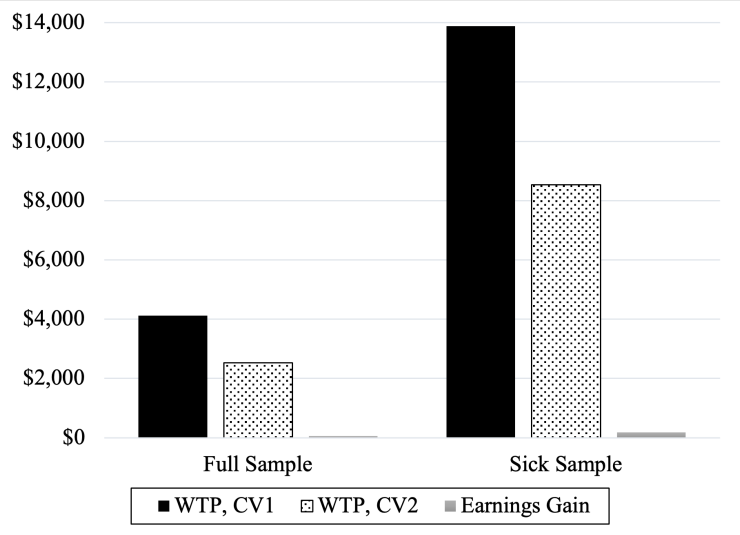
Our willingness to pay calculations are summarized in Figure 4. All calculations are made for the full and sick samples. Our first calculation, "WTP, CV1", measures with a black bar the amount of money that would need to be taken from the average individual after the introduction of the new technology, to return them to the level of utility they experienced before the new technology was introduced (i.e., compensating variation), assuming that the new technology alters mental health flow as well as stock entering the following period. This figure is \$4,126 in the full sample and \$13,886 in the sick sample. Our second simulation, "WPT, CV2", assumes that the new technology only alters mental health flows in period 1. Measured with speckled bars in the figure, WTP under this assumption falls to \$2,536 and \$8,536 for the full and sick samples, respectively. With the final grey bar, we measure the period t earnings gain produced by the mental health improvements as \$53 and \$177 for the full sample and sick sample, respectively. This exercise highlights that individuals, particularly low mental health individuals, would be willing to sacrifice a lot to improve their mental health. That said, the value of improved mental health is derived overwhelmingly from direct utility gains and not indirect increases in earnings.

Finally, we can use these model predictions and supplemental data to calculate the total economic and welfare cost of below average mental health in the US.[35] For example, the US had approximately 129.6 million residents aged 26–55 in 2023 (Census). Our findings suggest that in this year, US residents were willing to pay about \$854B (2023 USD) to avoid below average mental health and that just 2.1 percent (\$17.9B) of this value is attributable to labor market gains.[36] To put these figures in perspective, Kessler, Heeringa, et al. (2008) argue that serious mental illness cost the US \$193B in lost earnings alone in 2002; a figure commonly cited by the American Psychological Association. If we use our estimates to make a similar inflation and populated adjusted estimate for 2002, we get a much smaller earnings loss of just \$10.1B. The difference in our findings is not surprising, as our study accounts for the endogeneity of mental illness. Their implicit

[35]This exercise comes with the caveats that (i) labor market changes reflect a partial equilibrium response and (ii) we are assuming that individuals are only willing to pay for their own mental health improvements.

[36]To generate these aggregate figures, the average WTP and earnings gain is calculated for each individual from the CV2 simulation. We then multiply each figure by (i) 2 to annualize gains, (ii) 1.3 to account for inflation between 2013 and 2023 (BLS), and (iii) 129,587,395, the 26–55 year old population estimate.

Figure 4: The Value of Mental Health Improvements



Notes: The black and speckled bars measure the money that would need to be taken from an individual after the new technology is introduced to return them to their pre-policy utility level (i.e., compensating variation). Both calculations are made for one period. The black bar, CV1, assumes that the policy-induced mental health improvement in period t influences period $t+1$ mental health. The speckled bar, CV2, assumes that mental health entering $t+1$ returns to baseline. Earnings gains result from increased employment and wages in period t resulting from improved mental health. The Sick Sample is limited to individuals entering period t with mental health lower than 4 (i.e., worse than "good").

assumption is that people with serious mental health conditions would exhibit similar employment decisions and wage outcomes as people without such conditions if they were in better mental health, which ignores how mental health may correlate with omitted factors affecting labor market outcomes, which our model accommodates.[37] In particular, we find (i) little if any relationship between mental health and wages and (ii) the permanent-unobserved type with the worst mental health (i.e., $k = 2$) also has very strong preferences against working; thus, mental health improvements among the sickest in the population yield relatively small employment gains.

6.4.2 Assignment to therapy: We have established that mental health has direct utility and (small) indirect earnings benefits and relied on myriad findings showing that therapy is the most effective mental health treatment. Our second counterfactual assigns individuals to therapy in the first period of the model. We assume individuals attend the initial therapy session, but allow individuals to discontinue treatment or continue with additional sessions after learning their treatment effect. This design mimics an RCT that assigns a treatment group to therapy, but cannot guarantee follow-up visits or a uniform treatment effect.

We present baseline and counterfactual simulation results in Table 4. Statistics are presented

[37] Baseline estimates in Kessler, Heeringa, et al. (2008) are calculated as the difference in earnings and wages for those with serious mental illnesses vs. well individuals, while controlling for age, gender, race, census region, and urbanicity. When additional controls for just education, marital status, and household size are added, the figure is reduced by more than a quarter to $144B.

separately for the full and sick sample for each simulated time period. For the full sample, period one therapy participation increases from 2 to 100 percent (columns 1 and 2, row 1). One thing assigned therapy accomplishes is that it reduces search costs in the following period (note $\alpha_{1,1}$ in Appendix Table A.X). As a result, a full two years after assigned therapy, individuals in both samples are using therapy nearly 100 percent more than in the baseline simulation (column 12, rows 1 and 8). Nearly as large is the increase in antidepressant use two years later—an 86 (74) percent increase in the full (sick) sample (column 12, rows 3 and 10). Again, this is due to the dynamic complementary of the treatments. Mental health in both samples increases immediately following assigned therapy, but the gains diminish with time so that two years later, even the sick sample experiences just a four and a half percent improvement (column 12, row 11). Short-run effects on employment are small and negative (column 6, row 5 and 6), as the time cost of therapy and side-effects associated with increased antidepressant use outweigh the positive impact that improved mental health has on employment. Wages and long-run employment are mostly unaffected (column 12, rows 6 and 7).[38]

It is surprising that the improvement in mental health immediately following assigned therapy, which is assumed to be effective (on average) in the model, is quite moderate—just four percent above baseline for the full sample. This finding merits exploration. First, note that our discrete measure of mental health has a maximum value of five and that a large share of the population is in very good mental health. For example, in the baseline simulation 38 percent of the full sample enters the second period with $M_t = 5$ and, therefore, we measure no treatment induced improvement in their predicted mental health, despite these individuals realizing improvements in latent mental health. This fact helps to explain why the one-period, treatment-induced mental health improvement in the sick sample is a larger 9.6 percent. Yet even for this sample that enters period one with $M_t < 4$, 11.5 percent have $M_t = 5$ entering the following period in the baseline simulation.[39]

Second, the moderate increase in mental health that results from assigned treatment relates closely to the heterogeneous nature of treatment effects and an individual's ability to make intensive margin treatment decisions. To illustrate this, we conduct an additional assigned therapy counterfactual where we (i) force all individuals into a full course of therapy (i.e., 12 sessions) in the first period and (ii) assume no heterogeneity in treatment effects so that all

[38]These findings are qualitatively similar to Baranov et al. (2020), who randomly assign therapy to Pakistani women suffering postpartum depression. Their study also documents mental health improvements that diminish with time. Our finding of positive spill-overs of therapy to other choices (e.g., antidepressant use) but not secondary outcomes (e.g., wages) is also consistent with Baranov et al. (2020), who find that therapy increased mothers' financial empowerment and parental investments, but had no impact on children.

[39]Related to this point regarding the discrete nature of our mental health metric, there are individuals who experience increases in latent mental health due to therapy, but the increase simply isn't enough to push them past a threshold so that their discrete mental health measure improves.

individuals receive the sample mean. We plot the average increase in mean mental health entering period two on the left hand side of Figure 5, labeled "12, no het". Mean mental health increases by 8 (13) percent above baseline for the full (sick) sample. For this analysis, we add a third "very sick" sample, defined as those with initial mental health less than 4, of permanent unobserved heterogeneity type $k = 2$, and of time-varying unobserved heterogeneity type $j_1 = 2$; just seven percent of the period one sample is thus categorized as very sick. Forcing these individuals into 12 sessions with positive treatment effects raises average mental health by over 20 percent. This large response is partly explained by the fact that average baseline mental health entering period two in the very sick sample is low, so large percentage improvements are easier to achieve. The large gains are further explained by the fact that fewer individuals in this group achieve $M_t = 5$ entering period two at baseline (less than one percent), which is in large part due to these individuals being of time-varying unobserved heterogeneity type $j_1 = 2$ (see Table A.XVII). All in all, mental health improvements are much more pronounced when treatment is effective for all and individuals are assigned 12 sessions, with stronger impacts for the very sick.

Moving to the center bars in Figure 5, we amend the previous counterfactual by allowing individuals to discontinue treatment or continue with additional sessions after being initially forced to go; labeled "any, no het". Compared to the previous counterfactual, average mental health gains decline for all three groups. This decline is explained by the utility and financial cost of therapy, both of which are increasing in the number of sessions attended. Once individuals are exposed to these costs, some will elect to discontinue therapy, despite the consequences for their mental health.

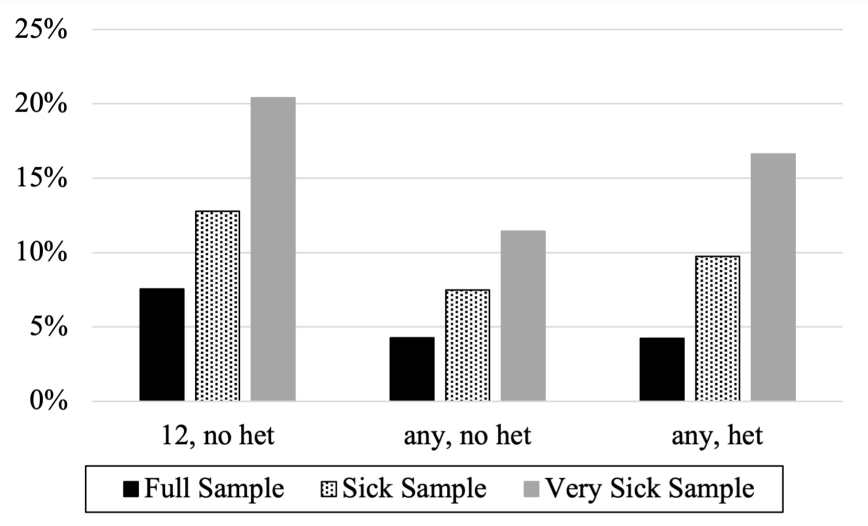
Finally, the far right of Figure 5, labeled "any, het", amends the prior counterfactual by allowing heterogeneous treatment effects (i.e., the Table 4 counterfactual). Compared to "any, no het", both the sick and very sick samples realize larger mental health gains. The primary mechanism by which this occurs is positive selection into additional therapy sessions—those learning that therapy is very effective for them upon being assigned treatment decide to attend additional sessions. Further evidence of this selection can be observed in Figure 6, which plots the distribution of treatment effects for individuals attending two sessions (i.e., discontinuing treatment as soon as it is possible) and greater than two sessions. The mean treatment effect is 0.233 larger (per session) for those attending more than two sessions.

In summary, Table 4 illustrates that a policy that assigns individuals to therapy, but that allows discontinuation of therapy and heterogeneous treatment effects, has a fairly large and positive impact on future treatment decisions, a moderate positive impact on mental health that diminishes with time, a small negative effect on employment that also diminishes with time, and virtually no impact on wages. The analysis summarized in Figure 5 shows that while forcing 12 sessions and positive treatment effects can produce large mental health improvements,

the moderate mental health effects that resulted from the initial policy are explained by a combination of factors: (i) because many people start in very good mental health, improving it is not really possible; (ii) even if individuals are assigned to therapy, making them stay in therapy is another more challenging issue; and (iii) not everyone benefits from therapy and forcing therapy on already healthy people can be detrimental.

Assignment to therapy can be viewed as a very strong one-time public policy. A clear implication of these findings is that more sustained interventions are needed to reach individuals not treating mental health conditions. Moreover, more realistic policy interventions must be considered. We explore such interventions next.

Figure 5: Mental Health Improvement from Assignment to Therapy in the First Period



Notes: The simulated data are constructed using the process described in Section 6.2. This figure reports the percentage increase in average mental health at the end of the first period due to therapy assignment across three simulations and three samples. The first simulation (12, no het) assigns individuals 12 therapy sessions and holds the treatment effect of each therapy session at the sample mean. The second simulation (any, no het) assigns two therapy sessions, but then allows the individual to decide if they want to attend additional sessions; treatment effects are again fixed at the sample mean. The third simulation (any, het) again assigns just two sessions and allows heterogeneous treatment effects. For completeness, we also conducted a fourth simulation, (12, het), that assigns individuals 12 therapy sessions and allows heterogeneity in the treatment effect; the corresponding percent changes in mental health levels are 3, 9, and 19 percent for the full, sick, and very sick samples, respectively. The Full Sample contains all individuals and average baseline mental health at the end of the first period is 3.98. The Sick Sample contains individuals with initial mental health under 4 (i.e., worse than "good"); baseline average mental health is 3.21. The Very Sick sample contains individuals with (i) initial mental health is less than 4, (ii) $k = 2$ (i.e., the sickest permanent type), and (iii) $j_1 = 2$ (i.e., the sickest time-varying type); baseline average mental health is 2.39.

6.4.3 Lower Costs of therapy: Our third set of counterfactuals explores several policies designed to reduce the costs of therapy. We again focus on mean outcomes for the full and sick samples, which are presented in Table 5. The table's first column contains baseline sample means for the simulated data, averaged over the four interview periods. For each policy, we present the corresponding sample mean and percent change from baseline.

The first policy that we consider (i.e., Cost Reduction Policy 1) eliminates the financial cost associated with therapy. Note that over the past three decades, several major US policies have

Table 4: Assigned therapy in First Period

	t=1			t=2			t=3			t=4		
	Base	Sim	% Δ	Base	Sim	% Δ	Base	Sim	% Δ	Base	Sim	% Δ
Full Sample												
Any therapy	0.0218	1.0000	44.9225	0.0223	0.1321	4.9140	0.0194	0.0596	2.0720	0.0185	0.0367	0.9783
Sessions | therapy> 0	5.8762	5.9722	0.0163	5.6141	5.5178	-0.0172	5.3746	5.3129	-0.0115	5.0884	5.0371	-0.0101
Medication	0.0712	0.1285	0.8041	0.0748	0.1860	1.4864	0.0751	0.1622	1.1590	0.0697	0.1295	0.8591
Avg MH	4.0321	4.0321	0.0000	3.9882	4.1567	0.0423	3.9647	4.0870	0.0309	3.9539	4.0416	0.0222
Working PT	0.1660	0.1633	-0.0162	0.1679	0.1667	-0.0074	0.1697	0.1696	-0.0007	0.1699	0.1706	0.0046
Working FT	0.5963	0.5857	-0.0177	0.5956	0.5859	-0.0164	0.5949	0.5867	-0.0138	0.5864	0.5795	-0.0118
Avg Wage	18.8731	18.6684	-0.0108	18.8715	18.7248	-0.0078	18.8360	18.7266	-0.0058	18.5108	18.4252	-0.0046
Sick Sample												
Any therapy	0.0525	1.0000	18.0416	0.0501	0.2033	3.0604	0.0396	0.1054	1.6595	0.0310	0.0608	0.9598
Sessions | therapy> 0	6.0188	6.3959	0.0626	5.6917	5.5686	-0.0216	5.3935	5.3358	-0.0107	5.1124	5.0297	-0.0162
Medication	0.1526	0.2368	0.5519	0.1476	0.2940	0.9914	0.1371	0.2582	0.8831	0.1168	0.2027	0.7361
Avg MH	2.7033	2.7033	0.0000	3.2172	3.5302	0.0973	3.4867	3.7189	0.0666	3.6394	3.8079	0.0463
Working PT	0.1607	0.1575	-0.0200	0.1647	0.1626	-0.0123	0.1686	0.1677	-0.0053	0.1717	0.1719	0.0015
Working FT	0.4627	0.4527	-0.0216	0.4667	0.4578	-0.0191	0.4737	0.4662	-0.0159	0.4762	0.4705	-0.0120
Avg Wage	13.3968	13.2010	-0.0146	13.5323	13.4061	-0.0093	13.6174	13.5111	-0.0078	13.5708	13.5011	-0.0051

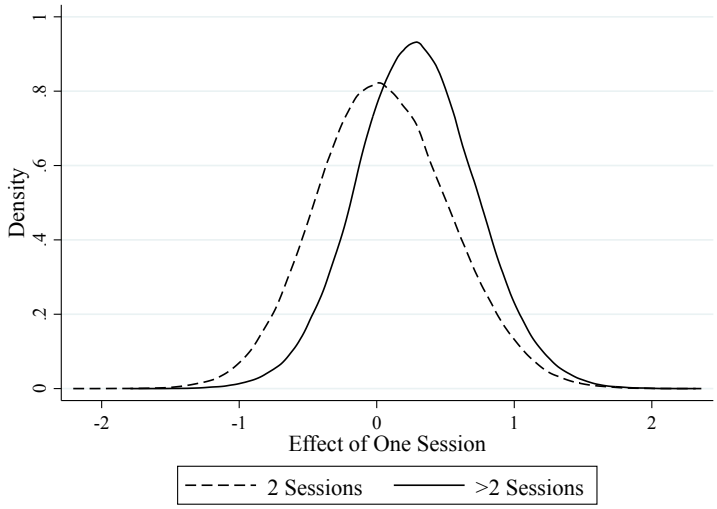
Notes: The simulated data are constructed using the process described in Section 6.2. In this table, we compare choices and outcomes in each of the four simulation periods, using the baseline model and an alternative model where individuals are assigned to therapy in period $t = 1$. The Sick Sample is limited to individuals with initial mental health that is lower than 4 (i.e., worse than "good").

Table 5: Counterfactual Policy Simulations

	Base	Policy 1		Policy 2		Policy 3		Policy 4	
	Level	Level	% Δ	Level	% Δ	Level	% Δ	Level	%Δ
Full Sample									
Any therapy	0.0205	0.0207	0.0110	0.0299	0.4581	0.0300	0.4640	0.0443	1.1602
Sessions | therapy> 0	5.5081	5.5327	0.0045	5.5716	0.0115	6.4695	0.1745	6.5602	0.1910
Medication	0.0727	0.0727	0.0003	0.0749	0.0301	0.0751	0.0330	0.0804	0.1058
Avg MH	3.9847	3.9848	0.0000	3.9864	0.0004	3.9950	0.0026	4.0006	0.0040
Working PT	0.1684	0.1684	-0.0001	0.1688	0.0028	0.1683	-0.0004	0.1692	0.0050
Working FT	0.5933	0.5933	0.0000	0.5934	0.0001	0.5932	-0.0001	0.5936	0.0005
Avg Wage	18.7729	18.7728	0.0000	18.7796	0.0004	18.7715	-0.0001	18.7880	0.0008
Sick Sample									
Any therapy	0.0433	0.0438	0.0103	0.0581	0.3427	0.0668	0.5414	0.0907	1.0932
Sessions | therapy> 0	5.6165	5.6404	0.0042	5.6759	0.0106	6.5316	0.1629	6.6316	0.1807
Medication	0.1385	0.1386	0.0009	0.1434	0.0351	0.1463	0.0560	0.1572	0.1347
Avg MH	3.2617	3.2619	0.0001	3.2657	0.0012	3.2889	0.0084	3.3014	0.0122
Working PT	0.1664	0.1664	-0.0001	0.1675	0.0063	0.1662	-0.0011	0.1684	0.0122
Working FT	0.4698	0.4698	-0.0001	0.4709	0.0023	0.4697	-0.0003	0.4718	0.0043
Avg Wage	13.5293	13.5284	-0.0001	13.5552	0.0019	13.5261	-0.0002	13.5753	0.0034

Notes: The simulated data are constructed using the process described in Section 6.2. In this table, we compare choices and outcomes for individuals using the baseline model and four alternative models representing the following policy interventions. Policy 1: Remove the financial cost of therapy. Policy 2: Remove the time/employment cost of therapy (i.e., $\alpha_{1,3} = \alpha_{1,4} = 0$). Policy 3: Remove bad therapists (i.e., the minimum therapy treatment effect is the sample mean, taken from the clinical literature). Policy 4: Combine Policies 1, 2, and 3. Sample moments are aggregated across all four simulated periods. The Sick Sample is limited to individuals entering the first period with mental health less than 4 (i.e., worse than "good").

Figure 6: Treatment Effects of therapy Users



Notes: The simulated data are constructed using the process described in Section 6.2. All individuals are assigned therapy in the first simulation period. The figure reports the distribution of simulated therapy treatment effects for those discontinuing after two therapy sessions and those going to more than two sessions. The means of the distributions are 0.053 and 0.286, respectively.

attempted to encourage mental health treatment by forcing insurers to share in the monetary cost of treatment, effectively lowering the out-of-pocket price for individuals.[40] Appendix Table A.VII provides some evidence that these policies have been effective at reducing costs as, in our data, the share of individuals paying no out-of-pocket cost has risen and average out-of-pocket costs, conditional on paying anything, have fallen over time. Table 5 suggests that both the full and sick samples are mostly unaffected by the policy change, as therapy increases by roughly one percent (column 3, rows 1 and 8), meaning there is virtually no movement in mental health or employment outcomes either. Though not shown, we also explore how the "very sick" sample from the previous section, a group policy makers might be most interested in helping, respond to this policy.[41] Despite being a relatively poor and sickly population, the price reduction has a similar effect on therapy use, roughly a one percent increase, which is due in large part to the fact that the very sick are of permanent unobserved heterogeneity type $k = 2$ and this group already pays very little for therapy (see Table A.XV; 67.5 percent pay nothing out of pocket and those who pay anything average just $32 per session). This is an important point to make. Those most in need of therapy cannot be induced into treatment via price reductions because they can already

[40] Examples include state-level mental health parity laws passed throughout the 1990s and early 2000s; the (federal) 1996 Mental Health Parity Act and 2008 Mental Health Parity and Addiction Equity Act; and the 2010 Patient Protection and Affordable Care Act, which made mental health an essential health benefit.

[41] Recall that the very sick sample is defined as having initial mental health less than 4, permanent unobserved heterogeneity type $k = 2$, and time-varying unobserved heterogeneity type $j_t = 2$. The last inclusion criteria necessarily means that an individual is defined to be very sick period-by-period. As such, in this section, we measure treatment and mental health effects for this sub-sample in the very sick period only.

access inexpensive care. Given the emphasis on mental health coverage and the expansion of public health insurance over the past several decades, this finding is maybe not all that surprising.

The second policy we consider (i.e., Cost Reduction Policy 2) eliminates the time/employment cost of therapy. In practice, we simulate the model with $(\alpha_{1,3}; \alpha_{1,4}; \alpha_{1,10}; \alpha_{1,11}) = 0$. This counterfactual is meant to explore the value of bringing therapists into the workplace, allowing employees to attend therapy during work hours. This policy, which is already provided by some large employers (McLeod, 2001), could not only reduce time and effort costs associated with beginning therapy, but could also help destigmatize therapy in the workplace. Table 5 shows that this policy increases therapy use by 46 percent in the full sample and 34 percent in the sick sample (column 5, rows 1 and 8). The policy is less effective in the sick sample because many are of unobserved heterogeneity type $k = 2$, which has a strong preference for unemployment (see Table A.XV; just 20 percent are employed). For the same reason, the response among the very sick is even weaker, just a six percent increase. While the increases in therapy use for the full and sick samples are sizable in percentage terms, the low rate of therapy use at baseline means the absolute increase is quite small. As a result, there is essentially no improvement in population level mental health as a result of the policy (column 5, rows 4 and 11).

The third policy that we consider (i.e., Cost Reduction Policy 3) eliminates the possibility that the therapy treatment effect could fall below the sample mean, which again, is determined by estimates from the clinical literature. While this policy is likely the least plausible of the three, it is meant to capture several ideas. First, it is likely that some therapists simply aren't very good at their jobs and with tighter regulation on licensing and continuing education, some bad therapy experiences could be avoided. Second, some low-productivity therapy sessions could result from poor therapeutic alliance (i.e., bad therapist-patient match quality), a problem that could potentially be improved in the era of big data and machine learning. This counterfactual can be thought to provide an upper bound on the potential gains of such efforts. In response to the policy, we find that (any) therapy use increases by 46 and 54 percent, respectively, for the full and sick samples (column 7, rows 1 and 8), while sessions conditional on any use increase by 17 and 16 percent (column 7, rows 2 and 9). Furthermore, the dynamic complementary of the treatments means that antidepressant use also increases by 3.3 and 5.6 percent (column 7, rows 3 and 10). Similar to the previous policy, despite these being relatively large percentage changes, because the absolute increases in treatment are small, improvements in aggregate mental health are also small; just under one percent in the sick sample (column 7, row 11).

Finally, our fourth experiment (i.e., Cost Reduction Policy 4) implements all three policies—eliminating the financial costs of therapy, the time/employment cost of therapy, and the possibility that therapy treatment effects could deliver less-than-average returns. The result of this experiment is unsurprising given the findings above, but illustrates an important point: for

the population at large, a wildly ambitious, expensive, and likely unrealistic effort to encourage use of therapy would have a relatively small impact on aggregate mental health and employment. For the full sample, the combination of policies increases the likelihood of any therapy use by 116 percent, sessions conditional on any use by 19 percent, and antidepressant use by just over 10 percent (column 9, rows 1, 2, and 3); however, average mental health over the two-year period improves by just a half of a percent in the full sample and 1.2 percent in the sick sample (column 9, rows 4 and 11). The impact on part-time employment is similar, with almost no impact on full-time employment or wages (column 9, rows 5, 6, and 7). As was previously shown, virtually all of this benefit is produced by the elimination of time/employment costs and unproductive treatment effects, but not financial costs.

These results show that while factors widely viewed as critical barriers to therapy use, such as monetary and time costs, explain some reluctance to use the treatment, entirely eliminating these costs has a marginal effect on population-level mental health and does not raise the level of therapy use to that of antidepressant use. Moreover, even when therapy use is increased, via incentives or counterfactual assignment to treatment, the resulting improvement in mental health does not yield increases in employment or wages. As our work shows traditional policy tools that address these costs to be mostly ineffective, a natural follow-up question is: Could new approaches to policy do more to increase therapy take-up? A potentially fruitful approach that our work points towards are policies designed to reduce the stigma associated with mental health treatment, such as those discussed by Corrigan (2004) and Lannin et al. (2013).

As stigma cannot be measured directly in our data, the impact of stigma on treatment decisions is unobserved and therefore one of a number of factors contributing to the disutility of treatment. By lowering utility costs in simulation, we obtain an upper-bound on the potential effect of destigmatization policy, since other costs (e.g., effort or fatigue from therapy) would remain even absent stigma.[42] Our final set of counterfactuals reduce the disutility of therapy by between 2.5 and 30 percent.[43] Results are in Appendix Table A.XVIII. What is made clear in these results is that seemingly modest reductions in treatment disutility produce large increases in treatment use. For example, a 15 percent reduction in the disutility of therapy more than doubles baseline therapy use to 4.3 percent (column 2, row 7), which is nearly the same

[42]We acknowledge that mental illness may be stigmatized in a way that the same social concerns that prevent some from attending therapy may also lead them to under-report their mental distress on a survey like the one we use in estimation. Unfortunately, both doctors and researchers of mental illness are constrained by the fact that, to date, virtually all mental illness is diagnosed/measured by asking patients questions in the hopes that they will respond honestly. An interesting implication is that real life attempts to address therapy under-utilization via destigmatization programs could be difficult to evaluate if the programs change reporting practices, i.e., if individuals become more truthful about their mental health after it is destigmatized.

[43]To account for heterogeneity in therapy preferences, we multiply all therapy utility parameters (i.e., main and interaction effects) by 1-x, where x ranges from 0.025 to 0.3.

aggregate take-up rate induced by the wildly impractical "Cost Reduction Policy 4" that makes therapy free, removes the time/employment cost, and removes the possibility of treatment effects below the sample mean (see Table 5, column 8, row 1).

7 Conclusion

Studies from several disciplines, including economics (Baranov et al., 2020), show that random assignment to a course of psychotherapy improves mental health. Yet, individuals rarely choose to go to therapy. To understand why, we develop and estimate a structural model of mental health treatment usage and labor supply decisions. The model is designed to capture key trade-offs associated with therapy and antidepressants so that it can be used to evaluate how counterfactual policies affect treatment decisions and population mental health. Some model features designed to address problems specific to our context could be used in other settings where similar issues are present. In particular, we address negative selection into treatment, which complicates estimation in many contexts, using a multi-pronged approach. Our approach includes using outside data on treatment effects estimated in well-identified settings, and also modeling intra-period shocks to health. Doing so means the model is able to capture a negative correlation between therapy and future health arising from negative shocks to mental health that induce people to seek treatment. The estimated model is able to reproduce empirical patterns well, including moments we explicitly match in estimation, along with those we do not. Counterfactual policy evaluation performed using the estimated model shows that individuals value mental health. However, the benefits of therapy, which are increasingly indisputable, are difficult to leverage. People forgo therapy even when costs commonly thought to be key barriers to use, such as monetary and time costs, are removed.

Improving population mental health thus requires that we look beyond commonly suggested policies thought to increase therapy use. It is possible that individuals simply dislike therapy and that the utility costs we estimate should be taken at face value. Therapy entails talking with a stranger about personal and troubling issues. It requires commitment and may cause discomfort. Indeed, it is often referred to as "doing the work." Another potential source of disutility is stigma. Individuals may feel ashamed that they need professional help to process their emotions. Indeed, an extensive literature has discussed stigma in the context of mental health treatment (e.g., Corrigan, 2004; Lannin et al., 2013), and there is evidence that stigma plays an important role in other contexts like social welfare programs (Moffitt, 1983) and HIV testing (Yu, 2023). Consistent with stigma, our estimates show that the disutility of therapy is larger for people living in small towns, where receiving treatment anonymously is more challenging and stigma may thus play a larger. However, this interpretation requires caution since it is not possible to separate this possibility from differences in the supply of therapists. Still, if stigma plays a role, newer forms of

therapy delivery, in particular, telehealth, may reduce its impact. While telehealth is certainly on the rise, it is too early to understand its impact on therapy uptake and population mental health.

More generally, efforts to improve mental health require continued exploration of the reluctance to use talk therapy. To that end, it would be useful to collect data on why people avoid treatment for mental health and prefer antidepressants to therapy. Ideally, such information would be collected as a module in an existing data set (such as the MEPS) so that it could be analyzed alongside treatment choices, mental health, employment, and other sources of heterogeneity across individuals. Initial data collection efforts could include open-ended questions to help identify potentially unknown factors that contribute to reluctance. Answers could inform future models designed to assess policies aiming to improve population mental health.

The data and code underlying this research is available on Zenodo at https://doi.org/10.5281/zenodo.10607646.

References

Adda, J., C. Dustmann, and J.S. Görlach. 2022. "The dynamics of return migration, human capital accumulation, and wage assimilation." *The Rev. of Economic Studies* 89 (6): 2841–2871.

Arcidiacono, Peter, Holger Sieg, and Frank Sloan. 2007. "Living rationally under the volcano? An empirical analysis of heavy drinking and smoking." *International Econ. Rev.* 48 (1): 37–65.

Ardito, R.B., and D. Rabellino. 2011. "Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research." *Frontiers in psychology* 2.

Arrow, K.J. 1963. "Uncertainty and the welfare economics of medical care." *American Economic Review* 53 (5): 941–973.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko. 2020. "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial." *American Economic Review* 110 (3): 824–59.

Bellman, Richard. 1966. "Dynamic programming." *Science* 153 (3731): 34–37.

Berndt, E.R., R.G. Frank, and T.G. McGuire. 1997. "Alternative Insurance Arrangements and the Treatment of Depression: What are the Facts?" *American J. of Managed Care* 3:243–250.

Berndt, E.R., B.H. Hall, R.E. Hall, and J.A. Hausman. 1974. "Estimation and inference in nonlinear structural models." In *Annals of Econ. and Social Measurement, v.3, n.4,* 653–665.

Blau, David M, and Donna B Gilleskie. 2008. "The role of retiree health insurance in the employment behavior of older men." *International Economic Review* 49 (2): 475–514.

Butikofer, Aline, Christopher Cronin, and Meghan Skira. 2020. "Employment Effects of Healthcare Policy: Evidence from the 2007 FDA Black Box Warning on Antidepressants." *Journal of Health Economics,* no. 73.

Chan, Tat Y, and Barton H Hamilton. 2006. "Learning, private information, and the economic evaluation of randomized experiments." *Journal of Political Economy* 114 (6): 997–1040.

Chan, T.Y., B.H. Hamilton, and N.W. Papageorge. 2015. "Health, risky behaviour and the value of medical innovation for infectious disease." *The Review of Economic Studies* 83 (4): 1465–1510.

Cipriani, A., Toshi A Furukawa, Georgia Salanti, et al. 2018. "Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder." *Lancet* 391 (10128): 1357–1366.

Corrigan, Patrick. 2004. "How stigma interferes with mental health care." *American psychologist* 59 (7): 614–625.

Crawford, B Y Gregory S, and Matthew Shum. 2005. "Uncertainty and Learning in Pharmaceutical Demand." *Econometrica* 73 (4): 1137–1173.

Cronin, C.J. 2019. "Insurance-Induced Moral Hazard: A Dynamic Model of Within-Year Medical Care Decision Making Under Uncertainty." *International Econ. Rev.* 60 (1): 187–218.

Cuijpers, P., A. van Straten, E. Bohlmeijer, S.D. Hollon, and G. Andersson. 2010. "The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size." *Psychological Medicine* 40 (2): 211–223.

Currie, Janet, and Mark Stabile. 2006. "Child mental health and human capital accumulation: the case of ADHD." *Journal of Health Economics* 25 (6): 1094–1118.

Darden, M. 2017. "Smoking, expectations, and health: a dynamic stochastic model of lifetime smoking behavior." *Journal of Political Economy* 125 (5): 1465–1522.

Darden, Michael, Donna B Gilleskie, and Koleman Strumpf. 2018. "Smoking and mortality: New evidence from a long panel." *International economic review* 59 (3): 1571–1619.

Davis, Morris A, and E Michael Foster. 2005. "A stochastic dynamic model of the mental health of children." *International Economic Review* 46 (3): 837–866.

De Quidt, J., and J. Haushofer. 2016. "Depression for Economists." *NBER,* no. 22973.

DeRubeis, Robert J, Steven D Hollon, et al. 2005. "Cognitive therapy vs medications in the treatment of moderate to severe depression." *Archives of General Psychiatry* 62 (4): 409–416.

Dickstein, Michael. 2018. "Efficient provision of experience goods: Evidence from antidepressant choice." Working Paper.

Dimidjian, S., S. Hollon, et al. 2006. "Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression." *Journal of consulting and clinical psychology* 74 (4): 658.

Einav, L., A. Finkelstein, S.P. Ryan, P. Schrimpf, and M.R. Cullen. 2013. "Selection on moral hazard in health insurance." *American Economic Review* 103 (1): 178–219.

Ekers, D., D. Richards, and S. Gilbody. 2008. "A meta-analysis of randomized trials of behavioural treatment of depression." *Psychological Medicine* 38 (5).

Elkin, Irene, Tracie Shea, John Watkins, et al. 1989. "NIMH Treatment of Depression Collaborative Research Program." *Archives of General Psychiatry* 46:971–982.

Ettner, S.L, R.G. Frank, and R.C. Kessler. 1997. "The Impact of Psychiatric Disorders on Labor Market Outcomes." *Industrial and Labor Relations Rev.* 51 (1).

Frank, R., and T. McGuire. 1986. "A review of studies of the impact of insurance on the demand

and utilization of specialty mental health services." *Health Services Research* 21:241–265.

Frank, R.G., and P. Gertler. 1991. "An Assessment of Measurement Error Bias for Estimating the Effect of Mental Distress on Income." *Journal of Human Resources* 26 (1): 154–164.

French, Eric. 2005. "The effects of health, wealth, and wages on labour supply and retirement behaviour." *The Review of Economic Studies* 72 (2): 395–427.

French, Eric, and John Bailey Jones. 2011. "The effects of health insurance and self-insurance on retirement behavior." *Econometrica* 79 (3): 693–732.

Gilleskie, Donna B. 1998. "A dynamic stochastic model of medical care use and work absence." *Econometrica,* 1–45.

Gloaguen, V., J. Cottraux, M. Cucherat, and I.M. Blackburn. 1998. "A meta-analysis of the effects of cognitive therapy in depressed patients." *Journal of affective disorders* 49 (1): 59–72.

Göstas, M.W., B. Wiberg, K. Neander, and L. Kjellin. 2013. "'Hard work' in a new context: Clients' experiences of psychotherapy." *Qualitative Social Work* 12 (3): 340–357.

Gould, Robert A, Michael W Otto, Mark H Pollack, and Liang Yap. 1997. "Cognitive behavioral and pharmacological treatment of generalized anxiety disorder: A preliminary meta-analysis." *Behavior therapy* 28 (2): 285–305.

Grossman, M. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy* 80 (2): 223–255. ISSN: 0022-3808.

Hofmann, S.G., A. Asnaani, I.J. Vonk, A.T. Sawyer, and A. Fang. 2012. "The efficacy of cognitive behavioral therapy: A review of meta-analyses." *Cognative Therapy and Research* 36 (5): 427–440.

Hofmann, S.G., and J.A. Smits. 2008. "Cognitive-Behavioral Therapy for Adult Anxiety Disorders: A Meta-Analysis of Randomized Placebo-Controlled Trials." *Journal of Clinical Psychiatry* 69 (4): 621–632.

Jarrett, R.B., M. Schaffer, D. McIntire, A. Witt-Browder, D. Kraft, and R.C. Risser. 1999. "Treatment of atypical depression with cognitive therapy or phenelzine: a double-blind, placebo-controlled trial." *Archives of General Psychiatry* 56 (5).

Jolivet, Gregory, and Fabien Postel-Vinay. 2020. *A Structural Analysis of Mental Health and Labor Market Trajectories.* IZA WP No. 13518.

Keeler, E.B., W.G. Manning, and K.B. Wells. 1988. "The demand for episodes of mental health services." *Journal of Health Economics* 7 (4): 369–392.

Kessler, Ronald, Steven Heeringa, et al. 2008. "Individual and societal effects of mental disorders on earnings in the United States: results from the national comorbidity survey replication." *American Journal of Psychiatry* 165 (6): 703–711.

Kirsch, I., B.J. Deacon, T.B. Huedo-Medina, A. Scoboria, T.J. Moore, and B.T. Johnson. 2008. "Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration." *PLoS Med* 5 (2): e45.

Lannin, D.G., M. Guyll, D.L. Vogel, and S. Madon. 2013. "Reducing the stigma associated with seeking psychotherapy through self-affirmation." *Journal of Counseling Psychology* 60 (4): 508.

Laynard, R., D. Clark, Mm Knapp, and G. Mayraz. 2007. "Cost-benefit analysis of psychological

therapy." *National Institute Economic Review* 202 (1): 90–98.

Magnac, Thierry, and David Thesmar. 2002. "Identifying dynamic discrete decision processes." *Econometrica* 70 (2): 801–816.

McLeod, John. 2001. *Counselling in the workplace: The facts: A systematic study of the research evidence.* British Association for Counselling / Psychotherapy.

Mitte, Kristin, Peter Noack, Regina Steil, and Martin Hautzinger. 2005. "A meta-analytic review of the efficacy of drug treatment in generalized anxiety disorder." *Journal of Clinical Psychopharmacology* 25 (2): 141–150.

Moffitt, Robert. 1983. "An economic model of welfare stigma." *American Economic Review* 73 (5): 1023–1035.

Mroz, Thomas A. 1999. "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome." *Journal of Econometrics* 92 (2): 233 –274.

Papageorge, Nicholas W. 2016. "Why Medical Innovation Is Valuable: Health, Human Capital, and the Labor Market." *Quantitative Economics* 7 (3): 671–725.

Rust, John. 1987. "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher." *Econometrica: Journal of the Econometric Society,* 999–1033.

Shapiro, Bradley. 2020. "Promoting wellness or waste? evidence from antidepressant advertising." *American Economic Journal: Microeconomics,* no. Forthcomming.

Swift, J., and R. Greenberg. 2012. "Premature discontinuation in adult psychotherapy: A meta-analysis." *Journal of consulting and clinical psychology* 80 (4): 547.

Turner, E.H., A.M. Matthews, E. Linardatos, R.A. Tell, and R. Rosenthal. 2008. "Selective publication of antidepressant trials and its influence on apparent efficacy." *New England Journal of Medicine* 358 (3): 252–260.

Wampold, B., G. Mondin, M. Moody, F. Stich, K. Benson, and H. Ahn. 1997. "A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, all must have prizes." *Psychological bulletin* 122 (3): 203.

Wampold, Bruce E, and Jesse Owen. 2021. "Therapist effects: History, methods, magnitude, and characteristics of effective therapists." In *Bergin and Garfield's handbook of psychotherapy and behavior change: 50th anniversary edition,* edited by M. Barkham, W. Lutz, and L. G. Castonguay, 297–326. John Wiley & Sons, Inc.

Werbart, Andrzej, Peter Missios, Fredrik Waldenström, and Peter Lilliengren. 2019. ""It was hard work every session": Therapists' view of successful psychoanalytic treatments." *Psychotherapy Research* 29 (3): 354–371.

Wierzbicki, Michael, and Gene Pekarik. 1993. "A meta-analysis of psychotherapy dropout." *Professional psychology: research and practice* 24 (2): 190.

Wong, J., A. Motulsky, T.S. Eguale, D.L. Buckeridge, M. Abrahamowicz, and R. Tamblyn. 2016. "Treatment indications for antidepressants prescribed in primary care in Quebec, Canada, 2006-2015." *Jama* 315 (20): 2230–2232.

Yang, Z., D.B. Gilleskie, and E.C. Norton. 2009. "Health insurance, medical care, and health

outcomes: A model of elderly health dynamics." *Journal of Human Resources* 44 (1): 47–114.

Yu, Hang. 2023. "Social stigma as a barrier to HIV testing: evidence from a randomized experiment in Mozambique." *Journal of Development Economics* 161:103035.

Appendix

A.I Data

A.I.1 Estimation Sample Construction: We begin with individuals from the 1996–2011 MEPS cohorts. We then restrict the sample to those between the ages of 26 and 55 to focus on those for whom education is unlikely to change and retirement is unlikely to be a viable employment alternative. We also remove round one observations, as lags of several variables are used as controls in our econometric specification. Demographic information for this subsample (Sample A) is provided below in Table A.I. We then limit this sample to those who complete each of the five possible interview rounds (Sample B). This restriction allows us to avoid integrating over the likelihood function when sample periods are missing, which reduces the (already substantial) computational burden of estimation and simulation.

MEPS interview periods vary in length. They are 5.2 months long on average and approximately 75 percent are between 3.5 and 7 months long. Figure A.I shows the distribution of period lengths, rounded to the nearest half-month. Period length was randomly allocated as a part of the survey design. The estimation of our structural model requires that each interview period covers an approximately equal amount of time; thus, we eliminate observations where the length of time between interviews is less than 3.5 months or greater than 7 months. To avoid needing to integrate over missing time periods in the estimation of the structural model, we use the following process to eliminate individuals and observations from the data: (i) drop any observation where length is less than 3.5 months; (ii) drop any observation where length is greater than 7 months; and (iii) drop any individual whose 2nd, 3rd, or 4th interview is dropped in (i) or (ii). These restrictions produce Sample C. As Table A.I indicates, Sample C looks virtually identical to Sample A, which is nationally representative. In light of these similarities, we believe the sample restrictions described above are justified.

Finally, to decrease estimation time, we estimate the structural model using a 20 percent random sample of Sample C, which we refer to as Sample D.

A.I.2 Alternative Measures of Mental Health: We use a subjective mental health report throughout our analysis. There are three other potential measures of mental health in the MEPS data, but each has a significant downside that prevents us from using it as our primary measure. First, in every round, individuals are able to report mental health conditions, which are then given ICD-9 codes by professional coders. These reports almost certainly suffer from non-classical measurement error, as they are likely to be influenced by past, unobserved interactions with medical professionals. We also have information on two indices used to measure mental health via survey questions: the Kessler 6 index (K-6) and the Mental Component Summary (MCS). The K-6 is a commonly used mental health scale that is calculated from responses to six

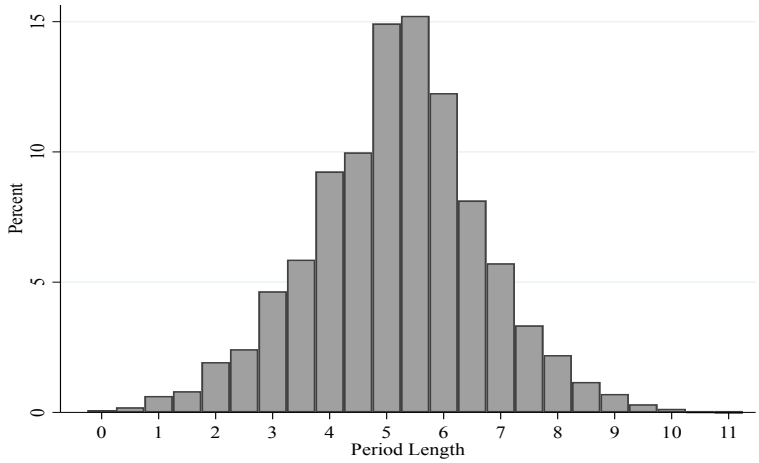
Table A.I: Sample Statistics Across Limiting Samples

	Sample A	Sample B	Sample C	Sample D
Demographics				
Male	0.465	0.456	0.457	0.458
Age	40.586	40.942	41.003	40.995
Live in MSA	0.830	0.828	0.824	0.818
Married	0.637	0.657	0.649	0.648
Family Size	3.416	3.432	3.386	3.381
White (race)	0.766	0.771	0.772	0.772
Problem Child	0.296	0.296	0.296	0.294
Health Insurance				
Public Insurance	0.133	0.132	0.140	0.147
Private Insurance	0.649	0.665	0.656	0.654
Schooling & Employment				
High School Grad.	0.542	0.538	0.532	0.526
College Grad.	0.250	0.257	0.253	0.256
Employed	0.770	0.776	0.765	0.761
Hourly Wage	23.344	23.686	23.445	23.602
Other H.H. Income	14,583	14,713	14,403	14,387
Treatment Decisions				
Therapy (round)	0.016	0.018	0.019	0.020
Therapy (ever)	0.040	0.046	0.047	0.047
Antidepressants (round)	0.062	0.070	0.074	0.075
Antidepressants (ever)	0.107	0.121	0.125	0.125
Mental Health				
Subjective	3.947	3.953	3.944	3.943
SAD (round)	0.095	0.096	0.100	0.101
SAD (ever)	0.174	0.178	0.182	0.180
Individuals	103,893	85,829	54,989	11,071
Observations	389,722	343,316	208,113	41,895

Notes: Problem child is measured in rounds two and four as the average response to 13 questions regarding problems with a child in the house. Examples are "(child has) problem getting along with Mom" and "(child has) problem behavior in school." Larger values indicate more problems. We measure the most problematic child in the household. Other Household Income is the weighted sum of an individual's own non-labor income and total income (labor and non-labor) of household members. Spousal income is given full weight, while non-spousal household income is weighted at a third of its full value. The mean hourly wage excludes the unemployed. Subjective MH is the respondent's subjective assessment of own mental health and ranges from 1 (poor) to 5 (excellent). The stress, anxiety, depression (SAD) indicator is based on the ICD-9 codes associated with reported disorders, including 296, 300, 308, 309, and 311.

questions of the form: "During the past 30 days, how often did you feel ... [nervous, hopeless, restless or fidgety, so depressed that nothing could cheer you up, that everything was an effort, worthless]?" For each question, an integer [0,4] is assigned to answers ranging from "none of the time" to "all of the time." The K-6 is calculated by summing the integers, generating a 0–24 scale, with lower scores indicating better mental health. The MCS is calculated from the standardized SF-12 health screening questions, where mental health questions (6–9) are weighted more heavily. The index range is 1–78, where higher scores indicated better mental health. The K-6 has only

Figure A.I: The Distribution of Period Lengths in MEPS



Notes: Data from Sample B (see Appendix Section A.I.1); 343,316 observations.

been collected since 2005 and the MCS since 2001. Most importantly, these two measures are only collected in survey rounds two and four. Estimation of the structural model we present in Section 4, which includes two types of unobserved heterogeneity that is identified partly by observed mental health transitions, requires more than one such transition per individual.

In Table A.II, we show that the subjective mental health score (MH) captures a significant amount of the variation in the two mental health indices and is highly correlated with the diagnosis of depression, anxiety, and stress disorders.[44]

Table A.II: Association between Subjective Mental Health and Other Measures

	MH=5	MH=4	MH=3	MH=2	MH=1
Mental Health					
SAD diagnosis	0.022	0.054	0.120	0.393	0.613
Kessler-6	1.975	2.955	4.613	10.057	15.012
Mental Component Summary	54.084	51.361	47.574	37.574	29.897

Notes: Subjective Mental Health (MH) is the respondent's subjective assessment of own mental health and ranges from 1 (poor) to 5 (excellent). SAD disoders are based on ICD-9 codes 296, 300, 308, 309, and 311. K-6 ranges from 0–24, while MCS ranges from 1–78. A higher (lower) score indicates greater mental distress for the K-6 (MCS) measure. Sample means reflect K-6 and MCS scores in rounds two and four only, measure from 2005 and 2001, respectively, onward.

A.I.3 Additional Descriptive Statistics: Table A.III shows how subjective mental health and treatment decisions differ by age. The first two columns highlight that as people age, subjective mental health worsens and the likelihood of reporting a SAD condition increases. For example, 6.0 percent of 26–30 year-olds report a SAD disorder, while the same is true of 13.2

[44]We define SAD disorders in the Section 3.2 as Stress Induced Disorders (ICD-9 Codes 308 and 309), Anxiety Disorders (ICD-9 Code 300), and Depressive Disorders (ICD-9 Codes 296 and 311).

percent of 51–55 year-olds. Similarity in these age patterns is one piece of evidence that both subjective mental health and the diagnosis of a condition capture variation in latent mental health. The correlation between these and two other measures of mental health is further discussed in Appendix Section A.I.2. Columns 3 and 4 of Table A.III show (any) antidepressant and therapy usage, respectively, by age group. Three patterns are evident from the table. One, use of mental health treatment rises with age. Two, an individual is about three-to-five times more likely to use antidepressants than therapy in an interview period. Three, the relative popularity of antidepressants holds across age groups, and therapy becomes even less popular (relative to antidepressants) as individuals age.

Table A.III: Mental Health and Treatment Decisions By Age

	Subjective MH	SAD Disorder	Antidepressants	Therapy
Ages 26–30	4.114	0.060	0.040	0.012
Ages 31–35	4.052	0.077	0.054	0.015
Ages 36–40	3.982	0.091	0.066	0.018
Ages 41–45	3.911	0.107	0.078	0.021
Ages 46–50	3.844	0.121	0.093	0.022
Ages 51–55	3.800	0.135	0.105	0.023

Notes: Data from Sample C (see Appendix Sect. A.I.1); 208,113 obs. Subjective mental health ranges from 1 (poor) to 5 (excellent). SAD disoders are based on ICD-9 codes 296, 300, 308, 309, and 311.

Table A.IV presents sample means by level of subjective mental health. The following are associated with worse subjective mental health: being female, older ages, living outside an MSA, being unmarried, having a smaller family, having a problematic child, not being white, having public health insurance, and having lower household income. Interestingly, high school graduation rates are constant across subjective mental health states, but those in the worst mental health states are the least likely to have a college degree. Employment falls as mental health declines. Across levels of subjective mental health, individuals are more likely to choose antidepressants than therapy. For example, individuals in the lowest subjective mental health category remain more than twice as likely to use antidepressants as they are to use therapy and those in the second-to-lowest subjective mental health category are almost three times as likely to use antidepressants. There is a close relationship between subjective mental health and reporting a mental health condition; however, even at low levels of subjective mental health, a large share of individuals do not report a condition. This could result from undiagnosed conditions or from subjective mental health being an imperfect measure of latent mental health.

A.I.4 Treatment Prices across Time and Insurance Status: Table A.VII shows how inflation-adjusted prices for individual therapy sessions and a one-month supply of antidepressants have changed over the sample period. The growth in total expenditures from all sources is shown in columns 3 and 6. This growth is consistent with medical prices in general out-pacing

Table A.IV: Sample Means By Subjective Mental Health

	MH=5 N=78,513	MH=4 N=63,261	MH=3 N=51,537	MH=2 N=11,877	MH=1 N=2,925
Demographics					
Male	0.484	0.458	0.432	0.393	0.394
Age	40.187	40.957	41.718	42.852	43.751
Live in M.S.A.	0.844	0.828	0.801	0.787	0.746
Married	0.699	0.673	0.611	0.445	0.328
Family Size	3.404	3.423	3.434	3.026	2.726
Problem Child	0.443	0.569	0.666	0.965	1.149
White (race)	0.766	0.793	0.769	0.722	0.709
Public Insurance	0.087	0.102	0.178	0.415	0.615
Private Insurance	0.734	0.699	0.568	0.386	0.245
Other HH Income	16373	15321	12037	8733	6393
School & Labor					
High School Grad.	0.521	0.542	0.539	0.525	0.523
College Grad.	0.327	0.271	0.162	0.111	0.073
Employed	0.832	0.809	0.712	0.460	0.220
Hourly Wage	25.627	23.525	20.244	18.603	19.388
Treatment Decisions					
Therapy	0.003	0.009	0.024	0.103	0.210
Antidepressants	0.023	0.053	0.103	0.299	0.476
Mental Health					
SAD disorder	0.034	0.071	0.140	0.392	0.600
Any disorder	0.038	0.077	0.148	0.418	0.645

Notes: Data from Sample C (see Appendix Sect. A.I.1); 208,113 obs. The mean hourly wage excludes the unemployed. Subjective mental health (MH) ranges from 1 (poor) to 5 (excellent). SAD disoders are based on ICD-9 codes 296, 300, 308, 309, and 311. Any (mental health) disorder refers to ICD-9 codes 290–319.

Table A.V: Mental Health and Labor Market Outcomes

	Employment		ln(Wage)		Hours	
	Coef.	SE	Coef.	SE	Coef.	SE
Mental Health						
Excellent	0.536	0.007	0.124	0.020	3.832	0.432
Very Good	0.525	0.007	0.082	0.020	3.430	0.433
Good	0.459	0.007	0.023	0.020	3.045	0.434
Fair	0.230	0.008	-0.025	0.021	1.789	0.454
Observations	N=208,113		N=159,284		N=159,284	

Notes: The excluded mental health group is "poor". All models control for sex, age, race, marital status, MSA, region, year, and education. Hours and hourly wage models are estimated on those who are working.

inflation (Peter G. Peterson Foundation, 2020). Columns 1 and 4 show how the proportion of individuals paying nothing out of pocket has grown over time, while columns 2 and 5 show that average out-of-pocket expenditures, conditional on spending anything, have fallen. Both patterns are consistent with third-party payers (i.e., government and insurers) paying a larger share of the ever growing price of treatment over time.

Table A.VI: A Multinomial Logit for Treatment Choices

	Antidepressants		Therapy		Both	
	Coef.	SE	Coef.	SE	Coef.	SE
Constant	-2.486	0.084	-3.772	0.282	-2.481	0.146
Male	-0.716	0.021	-0.482	0.075	-0.602	0.041
Age	0.032	0.001	0.003	0.004	0.016	0.002
MSA	-0.086	0.025	0.374	0.108	0.205	0.053
Married	-0.265	0.021	-0.582	0.076	-0.628	0.042
Mental Health						
Excellent	-3.133	0.054	-3.570	0.171	-4.967	0.100
Very Good	-2.336	0.051	-2.624	0.154	-3.870	0.079
Good	-1.667	0.050	-1.901	0.145	-2.614	0.065
Fair	-0.554	0.052	-0.643	0.147	-0.971	0.063
Region						
Midwest	0.299	0.032	-0.222	0.103	0.007	0.058
South	0.218	0.030	-0.581	0.099	-0.338	0.055
West	-0.136	0.033	-0.434	0.101	-0.371	0.058
Race						
Black	-1.060	0.033	-0.618	0.106	-0.843	0.056
Other (non-white)	-0.678	0.048	-0.508	0.152	-0.638	0.087
Education						
High School	0.456	0.027	0.536	0.102	0.634	0.051
College	0.602	0.033	1.400	0.117	1.338	0.064
Insurance						
Public	1.118	0.028	1.139	0.100	1.492	0.053
Private	0.603	0.027	0.367	0.100	0.483	0.054

Notes: Data from Sample C (see Appendix Sect. A.I.1); 208,113 obs. The base outcome is no treatment. The excluded mental health category is poor, the excluded region is the north, the excluded race is white, the excluded education level is less than high school, and the excluded insurance status is uninsured.

The second half of the table displays treatment prices by insurance status. Publicly insured individuals are found to be the most likely to pay nothing out-of-pocket for antidepressants, while facing the highest total price. These findings are consistent with both the generosity of Medicaid, as well as the federal government's inability to negotiate for drug prices, which influences Medicaid drug prices. That the uninsured face the lowest total antidepressant prices likely reflects selection into generic medication, while the relatively high out-of-pocket prices reflects the fact that there are few opportunities for reduced-price drugs. For therapy, the most generous type of coverage is public insurance, which is a somewhat misleading indicator that therapy is affordable and attainable for all publicly insured individuals. In reality many therapist simply do not accept Medicaid patients, which can make it difficult for patients to receive therapy. Finally, a somewhat surprising finding is how little the uninsured pay for therapy. Several factors contribute to this finding. First, these figures suggest selection into treatment (i.e., those facing

the lowest prices for care are the most likely to select it); thus, the low prices observed among the uninsured population partly reflects the fact that only those who can find lower prices choose treatment. Second, it is widely known that many psychologists use "sliding scale" pricing, meaning low-income and/or uninsured patients are charged less, or nothing at all, for treatment.

Finally, while single session and refill prices are presented in the table, prices enter the model as per-round expenditure levels (see Section A.III.2). The average number of therapy sessions attended per round (for someone attending at all) is 6.7. The average antidepressant user has 5.7 refills per-period, which means average out-of-pocket costs in a period are similar for the two treatments, about $168. That said, the combination of high discontinuation rates and a relatively large proportion of individuals receiving free therapy masks how much more therapy is for some. For example, an individual that does not receive any free care and pays the average (non-zero) per-session, out-of-pocket price for 12 sessions pays almost $600 for therapy.

Table A.VII: Treatment Prices

	Antidepressants			Therapy		
	% zero paid OOP	Ave. paid OOP (if non-zero)	Ave. paid (all sources)	% zero paid OOP	Ave. paid OOP (if non-zero)	Ave. paid (all sources)
Full Sample						
Mean	0.102	35.757	109.511	0.487	48.613	130.725
S.D.	0.302	61.125	137.317	0.500	103.069	187.810
Mean by Year						
1997	0.091	37.298	94.221	0.450	49.931	102.640
1998	0.075	38.404	106.699	0.461	39.189	104.100
1999	0.073	38.327	97.347	0.363	87.435	140.311
2000	0.047	43.418	109.678	0.490	40.712	110.447
2001	0.089	42.303	114.423	0.470	44.961	154.468
2002	0.079	36.040	105.614	0.445	48.945	122.817
2003	0.064	48.704	109.927	0.515	46.984	123.135
2004	0.117	39.513	116.046	0.477	55.017	120.922
2005	0.111	36.591	113.272	0.500	47.764	141.323
2006	0.076	39.296	106.697	0.486	48.021	123.438
2007	0.091	31.758	112.517	0.450	45.844	125.256
2008	0.091	30.295	116.540	0.524	44.487	124.461
2009	0.133	30.331	110.231	0.504	49.852	175.319
2010	0.168	32.817	115.735	0.552	35.871	133.393
2011	0.146	30.063	115.014	0.516	40.782	131.308
Mean by Insurance						
Private	0.025	32.457	105.99	0.199	50.542	134.135
Public	0.246	28.275	121.747	0.793	31.660	132.282
None	0.050	75.441	91.267	0.482	59.413	114.302

Notes: Table reports average inflation-adjusted price per therapy visit and average inflation-adjusted price per (one month) prescription. All prices are in 2013 dollars. Calculations use Sample C referenced in Table A.I

A.I.5 Effect sizes of mental health treatment: In Section 3.3.3, we summarize the medical literature that estimates the effect of therapy and antidepressant use on mental health

via randomized controlled trial (RCT) and mention that mean effects from this literature are used in our structural model. A potential concern is that these clinical effects sizes do not represent the effectiveness of these treatments in the "real world."

Many studies attempt to estimate the impact of treatment in less-controlled settings (often referred to as effectiveness studies). For example, Stewart and Chambless (2009) conduct a meta-analysis of studies that estimate the effect of cognitive behavioral therapy on anxiety in "less-controlled, real-world circumstances" and find effects above 0.8 for almost all forms of anxiety considered. Similarly, Hans and Hiller (2013) conduct a meta-analysis of non-randomized studies estimating the effect of cognitive behavioral therapy on unipolar depression in "routine clinical practice." They find effects above 1.0 for depression severity and effects ranging from 0.67 to 0.88 for secondary outcomes, such as dysfunctional cognition, general anxiety, psychological distress, and functional impairment. Teachman et al. (2012) note that for psychological treatments, "the evidence for successful translation from efficacy to effectiveness has increased dramatically in recent years, with numerous studies showing comparable effect sizes across settings." Both Hans and Hiller (2013) and Swift and Greenberg (2012) highlight that patient discontinuation is one of the most significant reasons that therapy can be less effective in practice, which is something that we explicitly account for in our analysis. With respect to antidepressants, some reviews of the literature have suggested that the real-world effectiveness of antidepressants is lower than the small effects shown in efficacy trials, which has led some to suggest that antidepressants should not be used to treat depression (Pigott et al., 2010). If the efficacy-effectiveness gap is greater for antidepressants than for therapy, this makes it even more puzzling that patients are far more likely to use antidepressants.

A.II Model—Disposable Income

The disposable income function $D(\cdot)$ in Section 4.2.2 adjusts gross household income, GY_t, for approximate total tax liability, housing expenses, and family size. To calculate $D(\cdot)$, we first separate households into income quintiles. We then calculate disposable income as $D(GY_t, \boldsymbol{X_t}) = GY_t * (1 - Tr_q) * (1 - Hr_q) * (1 - (1 - \sqrt{\frac{2}{(1+FS)}}))$, where Tr_q and Hr_q approximate the average total (federal, state, and local) tax rate and housing cost rate by income quintile. Wamhoff and Gardner (2019) estimate the following tax rates for the lowest to highest income quintiles: (20.7, 23.2, 26.5, 28.9, 32.0). Calculations are made prior to the 2017 Tax Cuts and Jobs Act (see Table 2, p. 5 in the referenced paper). We calculate the following after-tax housing cost rates, again for the lowest to highest income quintiles, using the American Community Survey micro-data: (50.14, 33.24. 22.98, 17.48, 13.04).[45] We then adjust for family size. Similar

[45]To calculate housing costs as a percent of household income, we use the 2011 1-year PUMS from the American Community Survey. Housing costs are based on the reported first mortgage payment for those

to Eckstein et al. (2019), the fraction of income that is spent on other family members is calculated as $1 - \sqrt{2/(1+FS)}$, where FS is family size; thus, a single person has $FS = 1$ and consumes 100 percent of their disposable income, a married individual with one child has $FS = 3$ and consumes 69.7 percent of his or her disposable income.

A.III Estimation and Identification

A.III.1 Likelihood Function: Observed decisions (d_t^{rce}), stochastic state transitions (M_t), and stochastic payoffs (p_t^e, w_t^x) are partly determined by a set of random variables, $\boldsymbol{\epsilon_t}$, that agents observe, but that we, the econometricians, do not. Constructing the likelihood function requires that we assume to know the distribution from which these unobservables are drawn. We begin by assuming that the unobservables affecting mental health, ϵ_t^M, are drawn from a logistic distribution, making $P(M_t)$ an ordered logit probability. We further assume that log-wage errors are normally distributed, $\epsilon_t^{w,e} \sim N(0, \sigma_{w,e}^2)$, non-zero price errors, $\epsilon_t^{f,x}$, are drawn from a logistic distribution, and log-price errors (conditional on prices being non-zero) are drawn from a normal distribution, $\epsilon_t^{p,x} \sim N(0, \sigma_{p,x}^2)$.[46]

We assume that the unobservables impacting treatment and employment decisions, ϵ_t^{rce}, are drawn from a Type 1 Extreme Value (T1EV) distribution. This assumption is popular in the DP literature because it yields closed form expressions for both the choice probabilities and Emax function. Starting with the choice probabilities, recall that (i) when agents are deciding whether or not to go to therapy, they do not know their therapy treatment effect, yet (ii) upon going to therapy once, they know their treatment effect, but cannot decide to "go back" and consume no therapy. Because extensive and intensive margin therapy decisions are made with different information, the resulting choice probabilities will be slightly different. Namely, the probability of choosing any set of alternatives containing no therapy is

$$P(d_t^{r0e} = 1|\boldsymbol{\Omega_t}) = \frac{\exp\left[\overline{V}^{r0e}(\boldsymbol{\Omega_t})\right]}{\sum_{r=0}^{1}\sum_{c=0}^{C}\sum_{e=0}^{2}\exp\left[\overline{V}^{rce}(\boldsymbol{\Omega_t})\right]} \tag{A.1}$$

where $\overline{V}$ is the deterministic part of the value function from Equation 3 in Section 4.2.4. Note, this value function integrates over the distribution of ϵ_t^{te}, consistent with the notion of the

households that own a home and on the gross rent payment for those who rent a home. We exclude from our calculation the 3.8 percent of households in the survey for whom total housing costs (based on 12 months of the mortgage or rent payment) are greater than household income.

[46] Under these assumptions, the price probability density function is as follows, where $\Lambda(\cdot)$ is the standard logistic CDF and $\phi(\cdot)$ is the standard normal pdf:

$$h_x(p_t^x|\boldsymbol{\Omega_t}) = \left(1 - \Lambda(\boldsymbol{X_t\eta^x} + \mu_k^{f,x})\right)^{\mathbb{1}_{[p_t^x=0]}} \left(\left(\Lambda(\boldsymbol{X_t\eta^x} + \mu_k^{f,x}\right)\frac{1}{\sigma_{p,x}}\phi\left(\frac{log(p_t^x) - \boldsymbol{X_t\gamma^x} - \mu_k^{p,x}}{\sigma_{p,x}}\right)\right)^{\mathbb{1}_{[p_t^x\neq 0]}}.$$

non-therapy user not knowing their treatment effect.

The probability of choosing a set of alternatives that involves therapy is somewhat more complicated. First, note that for therapy patients,

$$P(d_t^{rce}=1)=\left(1-P(d_t^{r0e}=1)\right)*P(d_t^{rce}=1|c>0,\epsilon_t^{te}). \tag{A.2}$$

In words, the probability that an individual selects an alternative with positive therapy use is the product of (i) the probability that they do not select zero therapy when treatment effects are unknown (i.e., this is one minus the probability in Equation A.1) and (ii) the probability of selecting c sessions when c must exceed zero and treatment effects are known to the agent. The latter probability is written

$$P(d_t^{rce}=1|c>0,\epsilon_t^{te})=\int \frac{\exp\left[\overline{V}^{rce}(\boldsymbol{\Omega_t},\epsilon_t^{te})\right]}{\sum_{r=0}^{1}\sum_{c=1}^{C}\sum_{e=0}^{2}\exp\left[\overline{V}^{rce}(\boldsymbol{\Omega_t},\epsilon_t^{te})\right]}f(\epsilon_t^{te})d\epsilon_t^{te}. \tag{A.3}$$

Here, the denominator excludes the possibility that $c=0$. Moreover, the deterministic part of the value function $\overline{V}$ is written as a function of ϵ_t^{te}, because the agent knows this value when making an intensive margin choice (i.e., this is the value function in Equation 3 in Section 4.2.4 *without* the integral over the therapy treatment effect). That said, we the econometricians do not know the true treatment effect; thus, we integrate over the treatment effect distribution when calculating the choice probability.

Next, we turn to the Emax function (see Equation 5 in Section 4.2.4). Assuming T1EV preference shocks means each of this function's three components has a closed form. The first component, $P(d_{t+1}^{r0e}=1)$, is given in Equation A.1. The other two components are

$$\begin{aligned} E_t[\max_{r0e} V^{r0e}(\boldsymbol{\Omega_{t+1}})] &= \gamma+\log\left(\sum_{r=0}^{1}\sum_{e=0}^{2}\exp\left[\overline{V}^{r0e}(\boldsymbol{\Omega_{t+1}})\right]\right) \\ E_t[\max_{rce} V^{rce}(\boldsymbol{\Omega_{t+1}})|c>0] &= \gamma+\log\left(\sum_{r=0}^{1}\sum_{c=1}^{C}\sum_{e=0}^{2}\exp\left[\overline{V}^{rce}(\boldsymbol{\Omega_{t+1}})\right]\right). \end{aligned} \tag{A.4}$$

where γ is Euler's constant.[47]

Let the variables $e_t=\{0,1,2\}$, $r_t=\{0,1\}$, and $c_t=\{0,...,C\}$ represent observed, period t employment, antidepressant use, and therapy sessions, respectively. Let the vector $\widetilde{\boldsymbol{\Omega_t}}=(M_t,K_t,\boldsymbol{X_t},\boldsymbol{d_{t-1}})$, which is the subset of the state-space known at the beginning of period

[47]That additive T1EV preference shocks produce an Emax function having this form is well known in the DP literature. The earliest reference to this result that we can find is in Ben-Akiva and Lerman (1985).

t, but before period t prices, wages, preferences, and time-varying unobserved heterogeneity are learned. Under the above assumptions, an individual's contribution to the likelihood function for a given realization of the parameter set $\mathbf{\Theta}$ can be expressed as

$$L_{i,t}(\mathbf{\Theta}|\mathbf{\Omega_t}) = g_1(w_t^1|\widetilde{\mathbf{\Omega_t}})^{\mathbb{1}[e_t=1]} g_2(w_t^2|\widetilde{\mathbf{\Omega_t}})^{\mathbb{1}[e_t=2]} h_r(p_t^r|\widetilde{\mathbf{\Omega_t}})^{\mathbb{1}[r_t=1]} h_c(p_t^c|\widetilde{\mathbf{\Omega_t}})^{\mathbb{1}[c_t=1]} \prod_{r=0}^{1}\prod_{c=0}^{C}\prod_{e=0}^{2}\left[P(d_t^{rce}=1|\mathbf{\Omega_t})\prod_{m=1}^{5}P(M_t=m|\mathbf{\Omega_t},d_t^{rce})^{\mathbb{1}[M_t=m]}\right]^{\mathbb{1}[r_t=r,c_t=c,e_t=e]}. \quad \text{(A.5)}$$

The first line measures wage and price contributions, which exist only if the individual was employed and/or sought treatment; hence, the indicator functions, $\mathbb{1}$. The first probability in the second row measures the choice contribution, which comes from Equation A.1 for those not using therapy and Equation A.2 for those using therapy.[48] The second probability in the second line measures the mental health contribution, where the probability of observing health state m is allowed to vary by the observed choice vector, (r_t, c_t, e_t).

The individual likelihood contribution is conditional on $\mathbf{\Omega_t}$, which contains both the permanent unobserved type k and the time-varying unobserved types j_t. Thus, constructing the log-likelihood function below requires calculating $L_{i,t}(\mathbf{\Theta}|\mathbf{\Omega_t})$ for each k and j_t, then weighting appropriately.

$$\mathscr{L} = \sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\theta_k(\widetilde{\mathbf{\Omega_0}})\prod_{t=1}^{T}\left[\sum_{j=1}^{J}P(j_{i,t}=j)L_{i,t}(\mathbf{\Theta}|\mathbf{\Omega_t})\right]\right) \quad \text{(A.6)}$$

A.III.2 Taking the Model to the Data: We describe Sample D in Section A.I.3. Within this sample, the average interview period length is 5.4 months; thus, for simplicity, we assume that all decision periods in the model are six months in length. The six month period length has several implications for estimation. First, employment-specific hours are set to 1,100 for full-time workers and 650 for part-time workers, which reflects 25 weeks of 44 and 26 hours worked, respectively.[49]

The number of therapy sessions attended in a period ranges from 0 to 48 in the data. Allowing for this many alternatives would be computationally burdensome, so we restrict the

[48] Note that choice probabilities are written as a function of $\mathbf{\Omega_t}$, which contains wages and prices information the agent is assumed to know when making decisions. In practice, we the econometricians only observe wages and/or prices when individuals are employed and/or consume treatment; thus, in the absence of employment/treatment, choice probabilities are calculated by integrating over wage and price distributions, $g(\cdot)$ and $h(\cdot)$. By the same logic, we must integrate over the therapy treatment effect distribution when calculating $P(M_t)$ any time therapy is used.

[49] Anyone in the data working over 37.5 hours per week is categorized as full-time. Among these individuals, the average number of hours is 44, while the average for those working under 37.5 hours per week is 26.

number of therapy sessions to $(0, 2, 6, 12, 20)$. In the likelihood function, individuals attending 1 to 3 sessions are coded as attending 2; 4 to 8 sessions are coded as attending 6; 9 to 16 sessions are coded as attending 12; and above 16 are coded as attending 20. We assume therapy is priced per session and that prices do not change within a period. We extract therapy prices from the data by calculating the average out-of-pocket price per (observed) session. Prescription drug choices reflect the decision to consume any antidepressants during a period, meaning prices reflect total expenditure levels. We, therefore, extract antidepressant prices by summing over all observed out-of-pocket payments within the period.

As noted in Section 3.3.2, we define a therapy treatment episode as a consecutive sequence of therapy sessions occurring without a two-month gap in visits. Many episodes contain few sessions of therapy, a phenomenon often referred to as discontinuation. In the data, some therapy episodes take place across multiple interview periods, which means that some interview periods have a small number of sessions that are either the beginning or a continuation of a longer therapy episode. To separate these occurrences in the data from discontinuation, we identify those episodes that span multiple interview periods, identify any interview periods that contain three or fewer sessions from an episode that spans multiple interview periods, and then roll these sessions into the adjacent interview period with the most therapy sessions. Of the 54,989 people in Sample C, 2,559 ever report going to therapy. There are 4,511 total therapy episodes across these 2,559 people. Without any adjustment, 942 of these episodes (or 20.9 percent) would span multiple interview periods. So, for roughly 80 percent of episodes all of the therapy sessions in the episode are assigned to the interview period during which the session took place. For the 942 episodes that span multiple periods, roughly 76 percent of sessions remain in the interview period during which the session took place, while 24 percent are moved to an adjacent interview period (these are roughly 6 percent of all sessions). As discussed in Appendix Section A.IV.1, we have also estimated the model without these adjustments and the results are nearly identical.

The fact that individuals enter the data at various ages also has implications for the model. Most notably, we do not observe experience over the entire career, which is a key determinant of wages. As such, we condition the wage distributions on the observed wage in the first period of the data and measure experience, K_t, earned since the first period.

Mental health, M_t, is observed in the data as a categorical variable: five levels ranging from "excellent" to "poor." As such, in Section 4.2.3, we model mental health as an ordered, discrete outcome. Elsewhere in the model, mental health operates as an independent variable in 12 separate locations. In all such instances, we treat our mental health variable as having a cardinal interpretation, allowing both linear and quadratic effects. There are two reasons: One, estimating the impact of five mental health categories in each location requires 48 parameters; our approach requires just 24 parameters. Given the very large size of the existing parameter

space, this reduction translates to significant computational savings. Two, very few people are in the two worst mental health states: 5.8 percent for $M_t = 2$, or "fair", and 1.4 percent for $M_t = 1$, or "poor". In several places, we interact mental health with other variables, meaning, were mental health measured using indicators, there would be very few observations identifying the low mental health effects. In other words, assuming cardinally aids in identification.

The full set of exogenous, non-stochastic control variables that comprise $\boldsymbol{X_t}$ are as follows: initial wage; gender; age; calendar year; lives in an MSA; lives in the midwest, the south, or the west (northeast omitted); has public insurance or private insurance (uninsured omitted); has high school or college education (less than high school omitted); nonwhite (race); married; family size; has a problem child; and household income, which excludes own-labor income. All variables are allowed to evolve over time, except race. In Table A.VIII below, we indicate for each of these variables the outcome variables it is allowed to influence.

Table A.VIII: Exogenous Controls Allowed to Influence Each Outcome

	Utility, Treat.	Utility, Emp.	Mental Health	Wages	Prices$\vert > 0$	Prices$\neq 0$
US region			X	X	X	X
Education			X	X	X	X
Race			X	X	X	X
Gender	X	X	X	X	X	X
Age	X	X	X	X	X	X
Lives in an MSA		X	X	X	X	X
Marriage Status		X	X	X		
Family Size		X	X	X		
Insurance Coverage					X	X
Calendar Year					X	X
Has Problem Child			X			
Initial Wage				X		
Household Income	X	X				

Notes: Columns 1 and 2 represent the marginal utility of treatment and employment. Columns 3-6 represent endogenous, stochastic variables in the structural model. An "X" indicates that the control variable is allowed to influence the corresponding outcome variable directly. Note that household income influences the marginal utility of treatment and employment through the budget constraint.

Average gross household income, GY_t, from all sources is about \$33,000 per period in the data; however, the data contain unemployed individuals that report zero household income. We set a gross income floor of \$2,500. Because estimation requires that value functions are calculated for every treatment and employment alternative, numeraire consumption can become negative when income is low and a lot of treatment is consumed and/or prices are high. In Equation 1 of Section 4.2.1, we assume CRRA preferences, allowing diminishing marginal returns to consumption (for values of α_1 below 1). This function is not defined for negative consumption values and the slope of the function can become very steep when consumption is

less than 1. To allow for negative consumption, while avoiding dramatic shifts in the marginal utility of consumption at any particular level, we replace the first term in Equation 1 with $\alpha_2 C_t$ when C_t drops below 1. Moreover, to avoid computation errors than can arise with large consumption values, we divide numeraire consumption by 1,000, meaning the shift from CRRA to linear preferences in consumption occurs at $C_t = 1,000$.

A.III.3 Standard Identification Arguments: Our focus here is on the identification of utility parameters since our discussion in the main text only briefly summarizes this issue. Identification relies on standard arguments discussed in Magnac and Thesmar (2002). Given data on state-specific choice probabilities and choice-and-state specific transition probabilities along with normalizations (i.e., that utility from one alternative is fixed or known), a fixed discount factor, an assumed belief structure, and distributional assumptions on error terms, it is possible to estimate utility parameters for each (not-normalized) choice-state pair.

Note that identification requires that the modeled beliefs about the impacts of treatment be a correct representation of agents' beliefs. Consider an agents' reluctance to use therapy; it could be because they believe it is effective, but have a distaste for it. Alternatively, agents may have biased beliefs and expect therapy to be less productive than it actually is. Absent belief data, either narrative could explain the same data pattern. In much of the literature that includes the estimation of a dynamic structural model, when beliefs data are available, rational expectations are a widely-used identifying assumption (Magnac and Thesmar, 2002). We follow this approach and impose rational expectations as an identification restriction, which in our case includes specific assumptions about timing: agents know the distribution of treatment effects, learn their specific treatment effect draw once they go to treatment, and are aware of intra-period shocks to mental health, including their likelihood of experiencing future shocks, which can affect how long they expect to remain in good mental health if they choose treatment. We consider robustness to relaxing some of these assumptions in Section 6.3.[50]

We give some specific examples of how utility parameters are identified. Agent treatment choices, the values of which are functions of both current period costs and expected future mental health benefits, identify preferences for mental health treatments. Preferences for work are similarly identified by joint work decisions and income. For example, not all agents work. If they forgo little income by not working, the disutility of work need not be very large to rationalize

[50] A final point is that our discussion of rational expectations and identification is instructive in clarifying what is meant by *identification*. Different assumptions about beliefs would lead to different utility parameter estimates that fit the data, which means we must make an assumption about beliefs to secure identification in the "rank-order" sense. The wrong assumption about beliefs, while still allowing us to estimate a unique set of parameters, might however lead to biases in estimates. Thus, unbiased parameter estimates, an additional notion of identification, requires that our model of beliefs is a good approximation of what individuals in our data assume when making choices.

observed behavior. If they forgo a lot of income by not working, the disutility of work parameter must be larger. In this sense, both income and choices identify work parameters. Consumption utility parameters are separately identified by differences in treatment decisions across the price distributions, as well as differences in work decisions across the household income and wage distributions.[51] Preferences over mental health itself are identified by observed shifts in treatment choices across mental health states along with the restriction that treatment preferences do not vary across these states. The identifying variation is displayed in Table A.IV, which shows that treatment increases as mental health worsens. Conditional on productive treatment, higher treatment uptake in worse health states reflects that individuals dislike being in poor health and are, thus, willing to incur the various costs of treatment to improve their mental health as it worsens.

A.III.4 Permanent Unobserved Heterogeneity: We assume in estimation that each individual has a permanent, unobserved type, which allows correlation between the unobserved determinants of choices, outcomes, and transitions in the model. The strategy decomposes all model unobservables into two additively separable components: an i.i.d. serially-uncorrelated random component, ϵ_t, and a persistent component, $\boldsymbol{\mu_k}$, that varies across individuals of K different types. We assume that the distribution of persistent unobserved heterogeneity can be approximated by a discrete function, which is sometimes referred to as a discrete factor model (DFM) (Heckman and Singer, 1984; Mroz, 1999). Thus, the estimation procedure seeks to determine (i) the number of unobserved types in the population, K; (ii) the share of the population that is described by each type, θ_k, where $\sum_{k=1}^{K} \theta_k = 1$; and (iii) the impact that each unobserved type k has on all model choices, outcomes, and transitions, $\boldsymbol{\mu_k}$ for $k = 1, \ldots, K$. We describe a similar process by which time-varying unobserved heterogeneity in mental health is approximated using a discrete function in Section 5.2.

The DFM offers two advantages over a popular alternative, which is to assume a joint parametric distribution (e.g., multivariate normal) over the model's errors. First, the DFM is more flexible. Mroz (1999) uses Monte Carlo simulation in a two-equation, joint MLE setting to show that when the true error distribution is joint normal, DFM estimates are comparable to

[51] Specifically, note that α_1 captures how the marginal utility of consumption changes across the consumption distribution, while α_0 captures the importance of consumption gains relative to other things affecting utility. Household income not related to one's own labor places individuals at different locations in the consumption distribution. From this location, sensitivity in treatment decisions to variation in prices and sensitivity in employment decisions to variation in wages informs the consumption utility parameters. For example, if high income agents are just as sensitive to high prices as low-income agents, α_1 takes a value near zero, implying linear utility with respect to consumption. Curvature in the utility-consumption profile is then reflective of high income individuals being less sensitive to price and wage variation. If, across the income distribution, individuals simply are not sensitive to prices or wages, α_0 takes a value closer to zero, reflecting that agents simply do not care much about consumption. As prices are only observed for those engaged in treatment and wages for those working, variables that only affect consumption through their impact on prices (e.g., insurance status) or wages (e.g., initial wages and experience), or "exclusion restrictions", aid in the identification of these consumption utility parameters.

those derived using the correct distribution. However, when the true error distribution does not match the joint MLE distribution, the DFM outperforms all (tested) alternatives. Second, the DFM is almost certainly faster than assuming a joint parametric distribution, which typically requires that the parametric distribution be simulated in estimation.

In Section 5, we mention that allowing for permanent unobserved heterogeneity resolves several identification and measurement error challenges in estimation. Consider two individuals in the data. One is observed to be in a poor health state in virtually every period. The second is observed to be in a poor health state in just one period. With permanent unobserved heterogeneity in mental health, the first individual is likely to have a high probability of being an unobserved type that is persistently unwell, while the latter will likely have a low probability of being this type. Such assignment has two advantages. First, because the latter type is not permanently unhealthy, it relieves pressure on the model to explain why such an individual wouldn't go to treatment; namely, the one poor health report represents a sort of measurement error in the observable. The individual isn't chronically unwell, they likely just had a bad day when reporting. Second, that the former individual is categorized as persistently unwell allows the model an additional mechanism for explaining the negative selection pattern that we describe in Section 5.2. In particular, those selecting into treatment are the same permanent type as those persistently receiving poor mental health shocks.

Another identification challenge is the endogeneity of initial conditions. Recall, as individuals enter the data at various points in their life, it is unlikely that all initial state variables are exogenous. For example, consider someone who is observed to have poor mental health entering our data. This individual's personal history and particular life circumstances, some of which are outside the scope of our model, likely contributed to that poor health state, and also makes this individual more susceptible to bad mental health shocks moving forward. (Several other initial conditions pose similar endogeneity concerns, including employment, treatment, education, and household income.) As such, we condition type probabilities, θ_k, on the initial state vector, $\boldsymbol{\Omega_0}$, which includes all endogenous initial conditions, as well as means (across the four model periods) of the exogenous variables in $\boldsymbol{X}$. Thus, in the example above, the fact that this individual entered the data with poor mental health is allowed to influence the probability that they are of an unobserved type, k, that experiences worse mental health shocks. In using this strategy, we assume that all initial conditions are exogenous, conditional on unobserved type k. To our knowledge, this strategy was first used to address endogenous initial conditions in Keane and Wolpin (1997) and was later formalized by Wooldridge (2005).

Both Magnac and Thesmar (2002) and Kasahara and Shimotsu (2009) establish the conditions for identification of dynamic structural models with permanent unobserved heterogeneity. However, these models are "non-parametric" in the sense that (i) the models are fully saturated

(i.e., utilities and transitions can be estimated for every choice-state pair and type) and (ii) both unobserved heterogeneity and component distributions are assumed to have no particular functional form. In our case, where we estimate many fewer parameters and assume functional forms for composite distributions, the identification criteria are much weaker.[52] Permanent unobserved heterogeneity in such settings is best viewed as a simple random effect, in the traditional panel data sense, that is specific to individuals and common across equations (see, e.g., Yang et al., 2009). As such, identification simply requires repeated choice and outcome data at the individual-level, and two restrictions: that type probabilities sum to one and that type effects, μ, are restricted to zero for one type.

We describe below how patterns in the data allow the model to distinguish between the role of permanent unobserved types and other structural relationships. The first such attribute is repeated individual-level observations. Assume that a subset of the population is persistently healthy and that this persistence cannot be explained by observable variables. The estimation procedure, then, identifies a type, k', that corresponds to that subset. The larger the subset, the larger the share, $\theta_{k'}$, assigned to that type. The better their mental health, the larger the factor loading, $\mu^M_{k'}$, on that type. Second, as discussed above, unobserved type shares are estimated conditional on endogenous initial conditions. To understand how this affects, say $\partial M_t/\partial M_{t-1}$, again assume that unobserved type k' is persistently healthy. The model allows initial mental health to influence the probability that an individual is of type k'. An individual who is observed to have perfect mental health in each period, including $t = 0$, then contributes little to the estimation of $\partial M_t/\partial M_{t-1}$, as their data is best explained by them being type k' with high probability. Third, non-linearities and exclusion restrictions aid in determining whether the relationship between two endogenous variables is causal or due to common unobservables. For example, note that wages and mental health are positively correlated (see Table A.IV). This relationship could be due to a causal effect of mental health on wages (i.e., $\partial w_t/\partial M_t > 0$) or permanent unobserved heterogeneity (i.e., for unobserved reasons, the people who are most likely to fall into poor mental health may also receive the lowest wage offers). Exclusion restrictions, such as having a problematic child and observed treatment choices, help to distinguish these competing explanations by generating unique variation in mental health that cannot be entirely explained by permanent unobserved heterogeneity. If variation in these exclusion restrictions is also associated with changes in wages, then it suggests a direct relationship between mental health and wages, and the unobserved heterogeneity parameters must adjust accordingly.

[52]For example, in our setting, unobserved types shift mean mental health (as well as mean wages and prices), but marginal effects (e.g., the impact of M_t on M_{t+1}) are unaffected by type. Such restrictions are not imposed in Kasahara and Shimotsu (2009), meaning factors like (i) the number of values covariates take and (ii) the extent to which the impact that covariates have on choices varies across types is much more important for identification in their setting than in ours.

A.IV Parameter Estimates and Model Fit

Table A.IX: Permanent Unobserved Heterogeneity Parameter Estimates

		k=1	k=2		k=3	
Equation	Param.	Est.	Est.	S.E	Est.	S.E
Utility						
any therapy	$\mu_k^{U,0}$	0.0000	0.4873	0.2683	0.5454	0.2646
any rx	$\mu_k^{U,1}$	0.0000	-0.0479	0.1251	0.0457	0.1122
pt emp.	$\mu_k^{U,2}$	0.0000	-0.4380	0.0879	0.5774	0.0857
ft emp.	$\mu_k^{U,3}$	0.0000	-0.4008	0.0803	0.6054	0.0783
Mental health	μ_k^{M}	0.0000	-0.8687	0.1338	0.0356	0.1283
PT wage	$\mu_k^{w,1}$	0.0000	-1.3806	0.0123	-0.6680	0.0087
FT wage	$\mu_k^{w,2}$	0.0000	-1.2166	0.0045	-0.6072	0.0033
Therapy cost (any)	$\mu_k^{f,c}$	0.0000	-8.5868	30.9695	-7.8121	30.9623
Therapy cost	$\mu_k^{p,c}$	0.0000	-1.3770	0.3752	-1.3192	0.3864
Rx cost (any)	$\mu_k^{f,r}$	0.0000	-0.6961	0.4894	-0.8109	0.4709
Rx cost	$\mu_k^{p,r}$	0.0000	0.5821	0.1558	0.3773	0.1499
Type prob. param.						
Constant	θ_k^0	0.0000	3.4035	0.5098	1.1786	0.4423
Initial pt emp.	θ_k^2	0.0000	-2.4804	0.2166	2.3363	0.2071
Initial ft emp.	θ_k^3	0.0000	-2.0861	0.1810	3.0762	0.1827
Initial mental health	θ_k^4	0.0000	-0.0684	0.0792	-0.0509	0.0678
Initial therapy	θ_k^5	0.0000	-0.4085	0.6013	-0.6765	0.5515
Initial rx	θ_k^6	0.0000	0.2880	0.3259	-0.1061	0.3149
female	θ_k^7	0.0000	0.7767	0.1466	0.5827	0.1270
Initial age	θ_k^8	0.0000	0.0795	0.0924	-0.0140	0.0776
Initial year	θ_k^9	0.0000	0.0242	0.0147	0.0165	0.0122
Mean msa status	θ_k^{10}	0.0000	-0.0382	0.1934	0.1013	0.1561
Mean pub. ins. status	θ_k^{11}	0.0000	1.1450	0.2785	0.4252	0.2572
Mean priv. ins. status	θ_k^{12}	0.0000	-0.2032	0.1837	0.6956	0.1521
Mean edu	θ_k^{13}	0.0000	-0.7031	0.1125	-0.6953	0.0930
Nonwhite	θ_k^{14}	0.0000	0.0327	0.1721	0.2052	0.1520
Mean marriage status	θ_k^{15}	0.0000	0.2371	0.1786	-0.0850	0.1571
Mean log(hh inc.)/10	θ_k^{16}	0.0000	-0.0103	0.2080	-0.1115	0.1800
Mean problem child	θ_k^{17}	0.0000	0.0309	0.1465	0.0088	0.1260
Share of time in Midwest	θ_k^{18}	0.0000	-0.0037	0.2300	0.1237	0.1915
Share of time in South	θ_k^{19}	0.0000	0.1628	0.2095	0.0380	0.1762
Share of time in West	θ_k^{20}	0.0000	-0.2601	0.2126	-0.0789	0.1768
logit probabilities		0.0399	0.2400		0.7202	

Notes: Permanent unobserved heterogeneity parameters are discussed in Section A.III.4. We use Sample D from Table A.I to estimate the structural model. All k=1 parameters are normalized to zero. Logit probabilities are calculated as $\frac{\exp(\theta_k \Omega_0)}{\sum_{k'=1}^{3} \exp(\theta_{k'} \Omega_0)}$.

Table A.X: Utility Function Parameter Estimates

		K=1, J=1		K=3, J=3	
Variable	Param.	Est.	S.E	Est.	S.E
CRRA	α_0	0.0385	0.0521	0.2403	0.0549
Scale	α_1	0.0466	0.0084	0.0829	0.0140
Negative Consumption	α_2	0.2247	0.0679	0.2965	0.0774
Any Therapy	$\alpha_{1,0}$	-5.5523	0.1629	-6.0694	0.3099
$\times\ c_{t-1}$	$\alpha_{1,1}$	2.3689	0.0871	2.3602	0.0884
$\times\ r_{t-1}$	$\alpha_{1,2}$	1.8995	0.0985	1.9295	0.1004
$\times$ PT$_t$	$\alpha_{1,3}$	-0.3133	0.1552	-0.3303	0.2128
$\times$ FT$_t$	$\alpha_{1,4}$	-0.5350	0.1058	-0.5448	0.1865
$\times$ female	$\alpha_{1,5}$	-0.0872	0.0596	-0.0960	0.0607
$\times$ age$_t$	$\alpha_{1,6}$	-0.0552	0.0427	-0.0491	0.0433
$\times$ msa$_t$	$\alpha_{1,7}$	0.2713	0.0816	0.2724	0.0834
Therapy Sessions	$\alpha_{1,8}$	-0.0732	0.0268	-0.0934	0.0253
$\times$ Sessions (squared)	$\alpha_{1,9}$	-0.0031	0.0012	-0.0031	0.0012
$\times$ PT$_t$	$\alpha_{1,10}$	0.0108	0.0215	0.0202	0.0222
$\times$ FT$_t$	$\alpha_{1,11}$	0.0365	0.0143	0.0418	0.0143
Any Rx	$\alpha_{2,0}$	-4.9440	0.1019	-4.9614	0.1474
$\times\ c_{t-1}$	$\alpha_{2,1}$	1.2894	0.1012	1.2789	0.1030
$\times\ r_{t-1}$	$\alpha_{2,2}$	4.3336	0.0469	4.3365	0.0472
$\times$ PT$_t$	$\alpha_{2,3}$	-0.2604	0.0455	-0.2270	0.0655
$\times$ FT$_t$	$\alpha_{2,4}$	-0.3185	0.0372	-0.2943	0.0623
$\times$ female	$\alpha_{2,5}$	0.2157	0.0397	0.2294	0.0402
$\times$ age$_t$	$\alpha_{2,6}$	0.1280	0.0377	0.1562	0.0335
$\times$ msa$_t$	$\alpha_{2,7}$	-0.0445	0.0447	-0.0457	0.0454
PT$_t$	$\alpha_{3,0}$	-6.4256	0.8114	-4.9225	0.8469
$\times\ PT_{t-1}$	$\alpha_{3,1}$	4.6622	0.0504	4.0276	0.0563
$\times\ (5-M_t)$	$\alpha_{3,2}$	0.0100	0.0575	-0.0232	0.0599
$\times\ (5-M_t)^2$	$\alpha_{3,3}$	-0.0528	0.0193	-0.0366	0.0198
$\times$ female	$\alpha_{3,4}$	0.0049	0.1143	-0.0402	0.1259
$\times$ age$_t$	$\alpha_{3,5}$	0.3396	0.2336	0.4603	0.2611
$\times$ age$_t\times$ female$_t$	$\alpha_{3,6}$	0.1912	0.1547	0.1856	0.1713
$\times$ age$_t^2$	$\alpha_{3,7}$	-0.0392	0.0427	-0.0715	0.0477
$\times$ age$_t^2\times$ female$_t$	$\alpha_{3,8}$	-0.0477	0.0517	-0.0425	0.0580
$\times$ family size	$\alpha_{3,9}$	0.0125	0.0148	0.0156	0.0164
$\times$ female $\times$ married	$\alpha_{3,10}$	-0.0386	0.0379	0.0117	0.0429
$\times$ female $\times$ family size	$\alpha_{3,11}$	-0.0486	0.0186	-0.0438	0.0205
FT$_t$	$\alpha_{4,0}$	-8.6984	1.6107	-5.9455	1.6764
$\times\ FT_{t-1}$	$\alpha_{4,1}$	5.0041	0.0418	4.1484	0.0495
$\times\ (5-M_t)$	$\alpha_{4,2}$	0.0925	0.0522	0.0828	0.0552
$\times\ (5-M_t)^2$	$\alpha_{4,3}$	-0.0859	0.0179	-0.0823	0.0190
$\times$ female	$\alpha_{4,4}$	-0.1287	0.0946	-0.1899	0.1071
$\times$ age$_t$	$\alpha_{4,5}$	0.6819	0.4084	0.7629	0.4599
$\times$ age$_t\times$ female$_t$	$\alpha_{4,6}$	0.1004	0.1269	0.1413	0.1448
$\times$ age$_t^2$	$\alpha_{4,7}$	-0.0860	0.0328	-0.1028	0.0378
$\times$ age$_t^2\times$ female$_t$	$\alpha_{4,8}$	0.0004	0.0419	-0.0134	0.0482
$\times$ family size	$\alpha_{4,9}$	0.0516	0.0105	0.0552	0.0121
$\times$ female $\times$ married	$\alpha_{4,10}$	-0.0963	0.0359	-0.0390	0.0416
$\times$ female $\times$ family size	$\alpha_{4,11}$	-0.0684	0.0140	-0.0691	0.0161
$(5-M_t)$	α_5	-0.0109	0.2334	-0.5306	0.3384
$(5-M_t)^2$	α_6	-0.1398	0.0628	0.0247	0.0720

Notes: Utility parameters are discussed in Section 4. We use Sample D from Table A.I to estimate the structural model. We present estimates for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Table A.XI: Mental Health Parameter Estimates

		K=1, J=1		K=4, J=3	
Variable	Param.	Est.	S.E	Est.	S.E
Constant	$\nu_{0,0}$	6.8226	0.0842	14.4301	0.3738
Any Rx	$\nu_{0,1}$	0.7233	***	1.0772	***
Therapy Sessions (mean)	$\nu_{0,2}$	0.1205	***	0.1795	***
Therapy Sessions (s.d.)	$\nu_{0,3}$	0.3749	0.0232	0.4749	0.0318
$(5 - M_{t-1})$	$\nu_{0,4}$	-1.1071	0.0216	0.6969	0.0936
$(5 - M_{t-1})^2$	$\nu_{0,5}$	-0.1271	0.0075	-0.4631	0.0211
problem child$_t$	$\nu_{0,6}$	-0.2724	0.0182	-0.4416	0.0348
female	$\nu_{0,7}$	-0.1450	0.0766	-0.2692	0.1691
age$_t$	$\nu_{0,8}$	-0.1845	0.0744	-0.4650	0.1657
age$_t^2$	$\nu_{0,9}$	0.0064	0.0243	0.0417	0.0526
female $\times$ age$_t$	$\nu_{0,10}$	-0.0701	0.0991	-0.3275	0.2121
female $\times$ age$_t^2$	$\nu_{0,11}$	0.0236	0.0323	0.1194	0.0673
nonwhite	$\nu_{0,12}$	0.0178	0.0226	0.0456	0.0461
married$_t$	$\nu_{0,13}$	0.2146	0.0345	0.4145	0.0743
msa$_t$	$\nu_{0,14}$	0.1123	0.0265	0.1849	0.0550
high degree, high school$_t$	$\nu_{0,15}$	0.3060	0.0245	0.3475	0.0469
high degree, college$_t$	$\nu_{0,16}$	0.5416	0.0289	0.8701	0.0624
midwest$_t$	$\nu_{0,17}$	-0.0588	0.0310	-0.1032	0.0646
south$_t$	$\nu_{0,18}$	0.0243	0.0280	0.0724	0.0584
west$_t$	$\nu_{0,19}$	-0.0004	0.0295	-0.0003	0.0602
family size$_t$	$\nu_{0,20}$	0.0386	0.0101	0.0954	0.0224
family size$_t$ $\times$ female	$\nu_{0,21}$	0.0257	0.0131	0.0848	0.0273
married$_t$ $\times$ female	$\nu_{0,22}$	-0.0385	0.0439	0.0207	0.0900
cut$_1$	ν_1	2.2515	0.0442	2.8506	0.0579
cut$_2$	ν_2	4.8272	0.0499	7.8363	0.2275
cut$_3$	ν_3	6.6297	0.0513	12.4461	0.3016
Time-Varying Unobserved Heterogeneity					
Type 2—MH effect	ψ_2	0.0000	***	-9.6546	0.3163
Type 2—$P(constant)$	ι_2^1	0.0000	***	-2.2886	0.0959
Type 2—$P(5 - M_{t-1})$	ι_2^2	0.0000	***	2.3417	0.0864
Type 2—$P((5 - M_{t-1})^2)$	ι_2^3	0.0000	***	-0.3018	0.0236
Type 3—MH effect	ψ_3	0.0000	***	-5.3323	0.1622
Type 3—$P(constant)$	ι_3^1	0.0000	***	-1.1576	0.0457
Type 3—$P(5 - M_{t-1})$	ι_3^2	0.0000	***	3.0530	0.0848
Type 3—$P((5 - M_{t-1})^2)$	ι_3^3	0.0000	***	-1.0243	0.0395

Notes: Mental health transition parameters are discussed in Section 4. We use Sample D from Table A.I to estimate the structural model. We present estimates for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity. The impact of antidepressants ($\nu_{0,1}$) and the mean impact of therapy ($\nu_{0,2}$) on mental health are taken from the clinical literature. In Section 5.2, we explain how effect sizes from the clinical literature are scaled for our measure of mental health. The mean effect of therapy reported above is for one session. The standard deviation of this single-session effect ($\nu_{0,3}$) is estimated. All unobserved heterogeneity parameters for $j_t = 1$ agent-periods are normalized to zero. Logit probabilities are calculated as $\frac{\exp(\iota_j(M_t))}{\sum_{j'=1}^{3}\exp(\iota_{j'}M_t)}$. Type probabilities are approximately 38 percent, 28 percent, and 34 percent on average. For individuals with $M_t = 1$, type probabilities are 10 percent, 89 percent, and 1 percent; with $M_t = 5$, type probabilities are 70 percent, 8 percent, and 22 percent.

Table A.XII: Log Wage Parameter Estimates

		K=1, J=1				K=3, J=3			
		PT		FT		PT		FT	
Variable	Param.	Est.	S.E	Est.	S.E	Est.	S.E	Est.	S.E
constant	δ^e_1	0.4738	0.0353	0.6426	0.0206	1.1721	0.0389	0.9312	0.0148
$(5 - M_t)$	δ^e_2	-0.0264	0.0099	-0.0131	0.0040	-0.0110	0.0091	-0.0112	0.0020
$(5 - M_t)^2$	δ^e_3	0.0054	0.0036	0.0027	0.0016	0.0033	0.0032	0.0039	0.0010
$log(W_0)$	δ^e_4	0.7956	0.0175	0.6853	0.0122	0.7278	0.0221	0.8242	0.0085
$log(W_0)^2$	δ^e_5	0.0027	0.0029	0.0257	0.0019	0.0340	0.0034	0.0194	0.0014
$\mathbb{1}_{W_0=0}$	δ^e_6	2.0350	0.0296	2.0169	0.0192	2.4993	0.0373	2.6181	0.0136
married_t	δ^e_7	0.0325	0.0122	0.0227	0.0036	0.0152	0.0130	0.0136	0.0039
$\text{high degree, high school}_t$	δ^e_8	0.0520	0.0089	0.0604	0.0035	0.0029	0.0096	0.0190	0.0034
$\text{high degree, college}_t$	δ^e_9	0.1590	0.0099	0.1457	0.0038	0.0208	0.0112	0.0490	0.0038
nonwhite	δ^e_{10}	0.0011	0.0083	-0.0042	0.0027	-0.0105	0.0091	0.0028	0.0028
female	δ^e_{11}	-0.0441	0.0134	-0.0266	0.0049	-0.0046	0.0130	-0.0090	0.0060
age_t	δ^e_{12}	0.0175	0.0153	0.0052	0.0064	-0.0374	0.0050	-0.0078	0.0050
age_t^2	δ^e_{13}	0.0006	0.0045	-0.0039	0.0020	0.0094	0.0030	0.0009	0.0017
msa_t	δ^e_{14}	0.0290	0.0082	0.0262	0.0032	0.0191	0.0093	0.0116	0.0032
K_t	δ^e_{15}	0.0096	0.0048	-0.0015	0.0014	0.0013	0.0035	-0.0011	0.0010
$K_t \times \text{age}_t$	δ^e_{16}	-0.0045	0.0028	0.0016	0.0007	-0.0002	0.0020	0.0006	0.0005
midwest_t	δ^e_{17}	-0.0220	0.0103	-0.0040	0.0036	-0.0248	0.0108	0.0018	0.0041
south_t	δ^e_{18}	-0.0623	0.0091	-0.0148	0.0031	-0.0443	0.0088	-0.0001	0.0034
west_t	δ^e_{19}	-0.0003	0.0089	-0.0040	0.0033	-0.0239	0.0096	-0.0037	0.0036
family size_t	δ^e_{20}	-0.0027	0.0035	-0.0070	0.0010	-0.0008	0.0034	-0.0023	0.0010
$\text{family size}_t \times \text{female}$	δ^e_{21}	0.0059	0.0044	-0.0039	0.0016	0.0046	0.0041	-0.0005	0.0017
$\text{married}_t \times \text{female}$	δ^e_{22}	-0.0096	0.0147	0.0047	0.0053	-0.0188	0.0153	0.0018	0.0056
variance	$\sigma^{w,e}$	0.3618	0.0014	0.2543	0.0004	0.2662	0.0009	0.1849	0.0003

Notes: Wage parameters are discussed in Section 4. We use Sample D from Table A.I to estimate the structural model. We present estimates for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Table A.XIII: Price Parameter Estimates

		K=1, J=1				K=3, J=3			
		Rx		Therapy		Rx		Therapy	
Variable	Param.	Est.	S.E	Est.	S.E	Est.	S.E	Est.	S.E
Non-zero Price									
constant	η_1^x	2.0675	0.3051	-0.5952	0.5725	2.7172	0.4579	7.6698	30.8579
female	η_2^x	-0.1415	0.1096	0.0421	0.1808	-0.1321	0.1099	0.0400	0.1913
age$_t$	η_3^x	0.0053	0.0711	-0.0593	0.1122	0.0110	0.0732	-0.0743	0.1176
nonwhite	η_4^x	-0.1120	0.1083	-0.2751	0.2129	-0.1209	0.1107	-0.1807	0.2249
high degree, high school$_t$	η_5^x	0.2396	0.1127	0.1870	0.2275	0.2512	0.1137	0.1012	0.2465
high degree, college$_t$	η_6^x	0.1769	0.1834	0.5861	0.2842	0.1750	0.1849	0.4599	0.3060
msa$_t$	η_7^x	-0.2220	0.1332	-0.3205	0.2173	-0.2039	0.1361	-0.4729	0.2162
year$_t$	η_8^x	0.0826	0.0302	0.0728	0.0403	0.0850	0.0376	0.0832	0.0433
pub. ins.$_t$	η_9^x	-0.9662	0.3520	-0.7356	0.4734	-0.9591	0.4535	-0.2849	0.4977
priv. ins.$_t$	η_{10}^x	1.9886	0.4260	2.3479	0.4831	2.0594	0.5112	2.2877	0.5145
pub. ins.$_t$ $\times$ year$_t$	η_{11}^x	-0.1005	0.0307	-0.0725	0.0450	-0.1019	0.0377	-0.0821	0.0482
priv. ins.$_t$ $\times$ year$_t$	η_{12}^x	-0.1494	0.0343	-0.0774	0.0449	-0.1518	0.0396	-0.0747	0.0483
midwest$_t$	η_{13}^x	0.5188	0.1564	-0.6618	0.2827	0.5289	0.1585	-0.8184	0.2996
south$_t$	η_{14}^x	0.4826	0.1306	0.1855	0.2658	0.4687	0.1363	0.0067	0.2788
west$_t$	η_{15}^x	0.0669	0.1476	0.2947	0.2649	0.0458	0.1521	0.1721	0.2778
Log(Price)									
constant	γ_1^x	4.7333	0.1664	1.2293	0.4496	4.3206	0.2239	2.4123	0.6068
female	γ_2^x	0.0060	0.0441	-0.1336	0.1077	-0.0131	0.0447	-0.0853	0.1116
age$_t$	γ_3^x	0.0241	0.0287	0.0775	0.0772	0.0197	0.0291	0.0590	0.0807
nonwhite	γ_4^x	-0.0233	0.0587	-0.1341	0.1284	-0.0336	0.0599	-0.0848	0.1289
high degree, high school$_t$	γ_5^x	-0.0895	0.0572	0.3311	0.1609	-0.0647	0.0585	0.2875	0.1661
high degree, college$_t$	γ_6^x	-0.0164	0.0715	0.7217	0.1946	0.0247	0.0736	0.6708	0.2009
msa$_t$	γ_7^x	0.1321	0.0505	0.3250	0.1490	0.1319	0.0509	0.2225	0.1486
year$_t$	γ_8^x	-0.0394	0.0129	0.1079	0.0327	-0.0413	0.0132	0.1130	0.0346
pub. ins.$_t$	γ_9^x	0.1766	0.1435	-0.0719	0.3736	0.0581	0.1537	0.0317	0.4076
priv. ins.$_t$	γ_{10}^x	-0.6130	0.1470	1.2071	0.3677	-0.6001	0.1504	1.3960	0.3766
pub. ins.$_t$ $\times$ year$_t$	γ_{11}^x	-0.0937	0.0135	-0.1061	0.0360	-0.0907	0.0139	-0.1042	0.0370
priv. ins.$_t$ $\times$ year$_t$	γ_{12}^x	0.0239	0.0137	-0.1099	0.0350	0.0261	0.0140	-0.1100	0.0365
midwest$_t$	γ_{13}^x	0.0215	0.0653	0.2479	0.1945	0.0175	0.0669	0.2328	0.2026
south$_t$	γ_{14}^x	0.4150	0.0574	0.3128	0.1680	0.3885	0.0587	0.3061	0.1761
west$_t$	γ_{15}^x	-0.0417	0.0664	0.3184	0.1814	-0.0491	0.0680	0.3484	0.1925
variance	$\sigma^{p,x}$	1.4000	0.0166	1.0684	0.0381	1.3986	0.0170	1.0384	0.0388

Notes: Price parameters are discussed in Section 4.2.1. We use Sample D from Appendix Table A.I to estimate the structural model. We present estimates for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Table A.XIV: Static Model Fit

	Est. Sample	Sim, K=1, J=1		Sim, K=3, J=3	
Variable	Mean	Mean	S.E.	Mean	S.E.
Treatment					
Any Therapy	0.0202	0.0218	0.0002	0.0216	0.0002
if $c_{t-1} = 1$	0.4651	0.4943	0.0043	0.4821	0.0042
if $r_{t-1} = 1$	0.1946	0.2098	0.0019	0.2086	0.0020
if $M_{t-1} = 5$	0.0029	0.0077	0.0001	0.0077	0.0001
if $M_{t-1} = 4$	0.0085	0.0118	0.0002	0.0120	0.0002
if $M_{t-1} = 3$	0.0256	0.0251	0.0003	0.0242	0.0003
if $M_{t-1} = 2$	0.1164	0.0971	0.0009	0.0578	0.0000
if $M_{t-1} = 1$	0.2215	0.2305	0.0022	0.2312	0.0027
Share with $p_t^c = 0$	0.5106	0.4670	0.0025	0.4867	0.0033
$p_t^c \| p_t^c > 0$	41.2218	48.2893	0.7120	45.8167	0.7933
Sessions $\| c_t \neq 0$	6.0731	5.5664	0.0283	6.0731	5.7258
Any Rx	0.0752	0.0748	0.0002	0.0751	0.0002
if $c_{t-1} = 1$	0.7136	0.7431	0.0019	0.7462	0.0021
if $r_{t-1} = 1$	0.7483	0.7470	0.0014	0.7514	0.0015
if $M_{t-1} = 5$	0.0224	0.0277	0.0002	0.0288	0.0002
if $M_{t-1} = 4$	0.0500	0.0505	0.0002	0.0506	0.0003
if $M_{t-1} = 3$	0.1073	0.1011	0.0003	0.0982	0.0003
if $M_{t-1} = 2$	0.2956	0.2879	0.0009	0.2948	0.0010
if $M_{t-1} = 1$	0.5099	0.4851	0.0018	0.4958	0.0020
Share with $p_t^r = 0$	163.7217	174.9151	0.9212	0.1127	0.0009
$p_t^c = 0 \| p_t^r > 0$	21.1150	20.8113	0.0224	174.8384	1.0508
Employment					
PT	0.1696	0.1694	0.0002	0.1691	0.0002
if $PT_{t-1} = 1$	0.8974	0.8975	0.0007	0.8854	0.0008
if $FT_{t-1} = 1$	0.0085	0.0041	0.0001	0.0087	0.0001
if $M_{t-1} = 5$	0.1696	0.1699	0.0003	0.1693	0.0003
if $M_{t-1} = 4$	0.1740	0.1732	0.0003	0.1733	0.0003
if $M_{t-1} = 3$	0.1785	0.1780	0.0003	0.1774	0.0004
if $M_{t-1} = 2$	0.1327	0.1341	0.0007	0.1344	0.0009
if $M_{t-1} = 1$	0.0819	0.0814	0.0013	0.0818	0.0017
Mean: W_t^1	21.1150	20.8113	0.0224	21.5742	0.0396
SD: W_t^1	17.7599	17.0623	0.0451	20.2790	0.1046
FT	0.5918	0.5944	0.0002	0.5952	0.0003
if $PT_{t-1} = 1$	0.0377	0.0139	0.0002	0.0275	0.0004
if $FT_{t-1} = 1$	0.9572	0.9562	0.0002	0.9525	0.0003
if $M_{t-1} = 5$	0.6591	0.6601	0.0003	0.6610	0.0004
if $M_{t-1} = 4$	0.6349	0.6374	0.0003	0.6381	0.0003
if $M_{t-1} = 3$	0.5281	0.5312	0.0004	0.5316	0.0004
if $M_{t-1} = 2$	0.3282	0.3381	0.0008	0.3389	0.0009
if $M_{t-1} = 1$	0.1168	0.1108	0.0013	0.1124	0.0012
Mean: W_t^0	24.3149	24.3675	0.0098	24.9099	0.0177
SD: W_t^0	15.6820	16.0961	0.0148	17.9126	0.0365
Mental Health					
$MH_t = 5$	0.3781	0.3785	0.0000	0.3785	0.0000
$MH_t = 4$	0.3029	0.3034	0.0000	0.3034	0.0000
$MH_t = 3$	0.2447	0.2446	0.0000	0.2446	0.0000
$MH_t = 2$	0.0586	0.0578	0.0000	0.0578	0.0000
$MH_t = 1$	0.0157	0.0157	0.0000	0.0157	0.0000

Notes: The simulated data are constructed by sampling from the joint error distribution, permanent and time-varying unobserved heterogeneity distributions, therapy treatment effect distribution, and estimated parameter covariance matrix 50 times for each individual, then forward simulating **just one period from the observed state space in the estimation data**. All moments are then calculated over all four simulation period; however, there is no dynamic updating. For example, assume a simulated individual enters has $M_1 = 4$ and the model simulates $M_2 = 5$. When we simulate M_3 for this individual in period 2, we do *not* necessarily set entering M_2 to 5, the previously simulated value. Rather, we assume that M_2 is equal to whatever it takes in the estimation data. We do the same with all other variables that update dynamically in the model: treatment, employment, and experience.

Table A.XV: Model Predictions by Permanent Unobserved Type

	k=1		k=2		k=3	
Variable	Mean	S.E.	Mean	S.E	Mean	S.E
Treatment						
Therapy Ever	0.0319	0.0021	0.1027	0.0013	0.0397	0.0005
Any Therapy per. t	0.0104	0.0008	0.0454	0.0007	0.0128	0.0002
Ave. Sessions	5.0736	0.1438	5.4511	0.0400	5.5949	0.0399
Share with $p_t^c = 0$	0.4565	0.0700	0.6750	0.0039	0.2437	0.0037
$p_t \vert p_t^c > 0$	155.4274	19.5547	31.9191	0.9465	48.6226	0.7625
Rx Ever	0.1152	0.0037	0.2203	0.0013	0.1029	0.0007
Any Rx period t	0.0582	0.0022	0.1402	0.0010	0.0510	0.0004
Share with $p_t^r = 0$	0.0429	0.0040	0.1741	0.0018	0.0558	0.0010
$p_t \vert p_t^r > 0$	123.5846	5.5661	180.7791	1.7798	173.7966	1.4645
Employment						
PT	0.2312	0.0048	0.0769	0.0010	0.1954	0.0005
Mean: W_t^1	44.8332	0.4246	10.2312	0.0478	21.1276	0.0382
SD: W_t^1	34.4512	0.5687	5.9378	0.1263	17.6586	0.0571
FT	0.5925	0.0053	0.1294	0.0010	0.7479	0.0006
Mean: W_t^0	50.0380	0.2901	11.9045	0.0578	24.8911	0.0143
SD: W_t^0	32.6552	0.3076	6.7203	0.1075	16.5698	0.0189
Mental Health						
$MH_t = 5$	0.4161	0.0026	0.2765	0.0009	0.4090	0.0004
$MH_t = 4$	0.3093	0.0013	0.2662	0.0005	0.3185	0.0005
$MH_t = 3$	0.2332	0.0021	0.3105	0.0007	0.2354	0.0004
$MH_t = 2$	0.0352	0.0010	0.1093	0.0005	0.0328	0.0002
$MH_t = 1$	0.0061	0.0004	0.0375	0.0003	0.0042	0.0001
$j_t = 1$	0.4014	0.0021	0.3255	0.0010	0.4023	0.0009
$j_t = 2$	0.2475	0.0023	0.3456	0.0016	0.2410	0.0013
$j_t = 3$	0.3511	0.0018	0.3288	0.0014	0.3567	0.0013
Ω_0						
PT_0	0.2307	0.0052	0.0493	0.0011	0.2024	0.0004
FT_0	0.5035	0.0055	0.0972	0.0014	0.7730	0.0005
$MH_0 = 5$	0.4628	0.0050	0.3125	0.0008	0.4498	0.0003
$MH_0 = 4$	0.2818	0.0029	0.2298	0.0006	0.3042	0.0002
$MH_0 = 3$	0.2010	0.0036	0.2875	0.0007	0.2054	0.0003
$MH_0 = 2$	0.0434	0.0016	0.1202	0.0004	0.0358	0.0001
$MH_0 = 1$	0.0110	0.0008	0.0500	0.0002	0.0048	0.0001
c_0	0.0183	0.0013	0.0306	0.0003	0.0080	0.0001
r_0	0.0598	0.0025	0.1300	0.0007	0.0439	0.0003
female	0.4441	0.0053	0.7135	0.0009	0.4909	0.0003
age_0	40.1480	0.0809	40.5768	0.0184	40.3214	0.0054
$year_0$	8.3486	0.0525	9.1491	0.0095	8.4367	0.0038
ave. msa	0.8259	0.0036	0.8048	0.0008	0.8211	0.0003
ave. pub. ins.	0.0954	0.0037	0.3798	0.0010	0.0732	0.0003
ave. priv. ins.	0.6697	0.0049	0.3564	0.0012	0.7718	0.0004
hs education	0.4726	0.0031	0.4872	0.0006	0.5417	0.0002
college education	0.4151	0.0042	0.1451	0.0008	0.2851	0.0003
nonwhite	0.1955	0.0034	0.2546	0.0008	0.2180	0.0003
ave. married	0.6719	0.0047	0.5958	0.0012	0.6615	0.0004
ave. hh inc.	16,630.29	177.40	13,966.70	42.81	14,702.72	13.58
share in northeast	0.1769	0.0045	0.1566	0.0007	0.1603	0.0003
share in midwest	0.1883	0.0038	0.1664	0.0009	0.2121	0.0003
share in south	0.3339	0.0047	0.4012	0.0009	0.3683	0.0003
share in west	0.3009	0.0048	0.2758	0.0008	0.2593	0.0003
Share of Population	0.0399	0.0005	0.2400	0.0005	0.7202	0.0007

Notes: The simulated data are constructed using the process described in Section 6.2. All moments are calculated over all four simulation periods.

Table A.XVI: Terminal Value Function Parameter Estimates

		K=1, J=1		K=3, J=3	
Variable	Param.	Est.	S.E	Est.	S.E
$(5 - M_{T+1})$	$\chi_{0,0}$	-1.9677	1.5884	-3.0227	1.8126
$\times$ age$_t$	$\chi_{0,1}$	0.8398	0.7949	1.5693	0.8926
$(5 - M_{T+1})^2$	$\chi_{1,0}$	-0.0221	0.6848	-0.0044	0.5975
$\times$ age$_t$	$\chi_{1,1}$	-0.4169	0.3406	-0.4577	0.2930
K_{T+1}	$\chi_{2,0}$	1.9965	0.8708	0.8718	0.9039
$\times$ age$_t$	$\chi_{2,1}$	-0.2626	0.2114	-0.2709	0.2365
c_T	χ_3	0.7938	0.1216	0.7620	0.1107
r_T	χ_4	0.4764	0.1433	0.4757	0.1438
PT_T	χ_5	1.9556	0.0994	1.6347	0.1012
FT_T	χ_6	2.4342	0.1357	2.0727	0.1356

Notes: The terminal value function is discussed in Section 4.2.4. We use Sample D from Appendix Table A.I to estimate the structural model. We present estimates for two models: one that includes (K=3, J=3) and one that does not include (K=1, J=1) unobserved heterogeneity.

Table A.XVII: Mental Health Transitions and Time-Varying Unobserved Heterogeneity

		Any	Any	M_{t+1} after	
	Share	Therapy	Rx	No Treat.	Any Treat.
j_t=1					
$M_t = 1$	0.005	0.138	0.265	3.809	4.097
$M_t = 2$	0.029	0.045	0.146	4.410	4.590
$M_t = 3$	0.193	0.020	0.075	4.821	4.874
$M_t = 4$	0.277	0.014	0.057	4.915	4.938
$M_t = 5$	0.496	0.011	0.043	4.907	4.935
j_t=2					
$M_t = 1$	0.032	0.176	0.346	1.393	1.884
$M_t = 2$	0.107	0.077	0.226	2.268	2.577
$M_t = 3$	0.336	0.027	0.101	2.809	2.967
$M_t = 4$	0.273	0.019	0.071	2.980	3.123
$M_t = 5$	0.253	0.018	0.063	2.962	3.107
j_t=3					
$M_t = 1$	0.005	0.095	0.198	2.713	2.975
$M_t = 2$	0.034	0.040	0.135	3.246	3.462
$M_t = 3$	0.257	0.021	0.083	3.724	3.889
$M_t = 4$	0.362	0.015	0.059	3.907	4.062
$M_t = 5$	0.342	0.013	0.052	3.886	4.038

Notes: The simulated data are constructed using the process described in Section 6.2. All moments are calculated over all four simulation periods. "Share" above corresponds to the share of individuals of time-varying type j that enter the period in each health state. As is stated in Table A.XI, we estimate that approximately 38 percent, 27 percent, and 35 percent of individuals are of types 1, 2, and 3, respectively, in each time period. "Any Treatment" above refers to either therapy or antidepressant treatment in period t.

A.IV.1 Robustness and Limitations Modeling treatment choices requires many decisions. We have experimented, estimating models with a number of alternative assumptions. In general, our results remain robust to these assumptions; all models yield similar unobserved types,

Table A.XVIII: Reducing the Disutility of therapy

	Full Sample			Sick Sample		
Disutility Reduction	Therapy Use	Sessions $\mid > 0$	Rx Use	Therapy Use	Sessions $\mid > 0$	Rx Use
0%	0.0210	5.3944	0.0727	0.0445	5.4496	0.1378
2.5%	0.0230	5.6627	0.0722	0.0495	5.8098	0.1464
5%	0.0253	5.6664	0.0739	0.0510	5.8063	0.1393
7.5%	0.0273	5.5745	0.0749	0.0570	5.6987	0.1433
10%	0.0335	5.6242	0.0753	0.0684	5.6733	0.1477
12.5%	0.0375	5.9735	0.0740	0.0726	6.1757	0.1442
15%	0.0430	5.9759	0.0789	0.0833	6.0438	0.1527
17.5%	0.0485	5.8528	0.0797	0.0929	6.0589	0.1557
20%	0.0568	6.0438	0.0817	0.1106	6.2445	0.1616
22.5%	0.0631	6.1210	0.0838	0.1134	6.0295	0.1646
25%	0.0731	6.1323	0.0934	0.1302	6.2253	0.1761
27.5%	0.0857	6.2393	0.0991	0.1486	6.4624	0.1907
30%	0.0969	6.0713	0.1039	0.1593	6.3929	0.1981

Notes: The simulated data are constructed using the process described in Section 6.2. All moments are calculated over all four simulation periods. This table shows the population share using treatment in a survey period, for different percentage reductions in the disutility of therapy. Each row of the table represents a separate simulation with a different level of therapy disutility. The Sick Sample is limited to individuals who enter the first period with initial mental health less than 4 (i.e., worse than "good").

parameter estimates, and model fit comparisons. We discuss a subset of alternative assumptions in more detail here. Parameter estimates and model fit comparisons are available upon request.

First, there is some concern that persistently healthy individuals never consider mental health treatment and, thus, our model overstates distaste for treatment. To address this concern, we reestimate the model while excluding individuals who report "excellent" mental health in every period (13.4 percent of the sample). The disutility of treatment indeed falls slightly, as average treatment rates are higher in this sample. The disutility of mental health becomes more quadratic (i.e., agents really want to avoid very low mental health states) which is likely the product of individuals with poor mental health comprising a larger share of the sample and therefore having more influence on the likelihood function. We see no real changes in unobserved type probabilities and model fit. Second, in an effort to validate the chosen interpretation of our results (i.e., that our results relate mainly to depressive conditions and related treatment) we reestimate the model while excluding those whose only reported mental health condition is *not* a SAD condition (about 2 percent of the sample), as described in Section 3.2. Estimation with this sample yields no meaningful differences. Third, rather than using the strategy described in Appendix Section A.III.2 to match therapy sessions to model periods, we ignore all therapy dynamics and simply assume that all sessions attended are decided upon in the interview period that we observe them occurring in. Again, no meaningful differences are observed in estimates, model fit, or unobserved type probabilities.

We then turn to alternative assumptions on therapy treatment effects. Our fourth robustness exercise reestimates the model without therapy treatment effect heterogeneity, assuming therapy is equally productive (and equal to the sample mean, which is taken from the clinical literature) for all individuals. Model fit and unobserved type probabilities are similar. The likelihood function value falls some, as one parameter has been removed from the model. Parameter estimates are virtually identical, except the disutility of poor mental health grows slightly, while the disutility of therapy shrinks. This result is expected. There is a strong empirical relationship between poor mental health and therapy use—worse health implies more use—which identifies the disutility of poor mental health; agreeing to take on the cost of going to therapy while in poor mental health tells us that individuals do not like being in poor mental health. If the benefits to treatment are less salient (i.e., if treatment effects are heterogeneous), then the fact that treatment rises when health declines tells a less clear signal about preferences for better health.

Next, recall that we use average treatment effects from the clinical literature to inform the mean treatment effect for 12 therapy sessions and the uniform treatment effect for antidepressant use in our model. Our next set of robustness tests considers alternative assumptions; in particular, we assume that the true average effects (i) are half those reported in the clinical literature, (ii) are double those reported in the clinical literature, and (iii) for antidepressants, are a function of lagged mental health. This last exercise is motivated by the meta-analysis of Fournier et al. (2010), which finds that the effectiveness of antidepressants is increasing in illness severity, and Elkin et al. (1989), which finds that therapy is no more effective than antidepressants for severely depressed patients.[53] These alternative specifications have a notable impact on the disutility of mental illness, but little else. When treatment effects are small (large), the disutility of poor mental health increases (decreases).[54] Moreover, when treatment effects are halved, the share of individuals in the sickest time-varying unobserved heterogeneity group grows. These differences reflect the empirical challenge discussed in Section 5. Namely, the raw data suggest that those with the worst mental health are the most likely to consume treatment (see Appendix Table A.IV), yet poor mental health is persistent, even with treatment (see Table 2). This pattern is difficult to rationalize with positive treatment effects; thus, when large treatment effects are imposed on the model, persistence in poor mental health observed in the data is rationalized with

[53]We can find no consistent evidence to suggest that therapy effectiveness is either increasing or decreasing in illness severity. Fournier et al. (2010) reports antidepressant effect sizes of 0.11 for mild to moderate depression, 0.17 for severe depression, and 0.47 for very severe depression; the latter has a 95 percent confidence interval from 0.22 to 0.71, making it somewhat consistent with Elkin et al. (1989). In our robustness analysis, we assume that those entering a period with self-reported mental health of "poor" are, as Fournier et al. (2010) describes, "very severely depressed", meaning antidepressants have an effect size of 0.47. Similarly, we assume treatment effects for those with fair, good, very good, and excellent mental health are 0.17, 0.11, 0.09, and 0.09, respectively.

[54]When antidepressant treatment effects are heterogeneous, results look similar to when both treatment effects are halved because (i) antidepressants are more popular than therapy and (ii) most individuals have relatively good mental health, meaning the alternative specification represents a decline in antidepressant effectiveness.

low marginal utility from better mental health, producing a weaker mental health treatment gradient. Despite this, and the fact that the likelihood function value improves with lower treatment effects, there is no improvement in model fit. Both Heckman (1991) and Deaton (2009) discuss why treatment effects estimated in controlled settings are often larger than the treatment effects that are actually realized in clinical settings. Estimated treatment effects for medical care may be too large if, for example, patients included in RCTs have few co-morbidities or physicians drawn from research hospitals are more skilled than the average physician.

Another issue of concern is that we do not model the supply of therapists, which could mean that the high utility costs of therapy usage that we estimate really reflects difficulties finding a therapist. Since we do not model the endogenous supply of therapists, we cannot use our model to examine the impacts of counterfactual policies that would affect treatment usage through shifts in the supply of therapy. For example, it would be a mistake to use our model to examine the impact of incentives to offer insurance to low-income patients. On the other hand, we are able to make some progress on the question of whether omitting supply from the model biases our estimates or otherwise drives results. In general, evidence suggests this is not the case. Our strongest evidence comes from our companion working paper Cronin et al. (2020). As mentioned in the manuscript (see footnote 23), in the early stages of this project, we attempted to estimate therapy and antidepressant treatment effects outside our structural model by instrumenting for treatment with several supply-related variables. In Appendix Table A.VI of the working paper, you can see first-stage results for privately insured individuals when we use number of psychiatrists in one's county as an instrument. This instrument has very little predictive power, even less so when uninsured and publicly insured individuals are included. As part of this analysis, we also attempted to use various combinations of the following variables as instruments: number of community mental health centers, number of social workers, number of general practitioners (all in levels and per-capita), as well as indicators for the passage of mental health parity laws (which effectively makes more providers "in network" for a larger share of the population). None of these variables seem to predict usage and, thus, are unlikely to be key impediments to treatment use.[55] Related to this, it is useful to point out that the majority of the patients we observe are adults suffering from mild symptoms that virtually any licensed therapy can treat. If our focus were on a more specialized condition, it would be more likely that large utility costs we estimate reflected the unavailability of treatment.

Several of our assumptions cannot be tested, so our results should be interpreted with this

[55]The analysis just described used variation at the county level. We estimated 2SLS models with county and time fixed effects. All supply variables came from the Area Health Resource File. Importantly, one can only observe county of residence in the MEPS data in a Census Research Data Center, which is one reason these measures of supply do not enter the structural model.

in mind. For example, as discussed in Appendix Section A.III.3, we assume that individuals have rational expectations and full information, which implies that when making treatment decisions, they understand average treatment effects as reported in the medical literature and act accordingly. A possibility is that individuals make treatment decisions with incorrect expectations, which would bias our estimates. This hypothesis could be tested with subjective data on expected treatment effects, which we leave to future work. Additionally, our model assumes patients have full agency in their medical decision-making. While this assumption is common in the dynamic structural literature (see, e.g., Chan et al., 2015), it is likely that doctors advise patients on these decision and that different doctors may advise different things. For example, whether the first treatment recommended to patients is antidepressants or talk therapy is likely influenced by whether the health professional is a primary care physician (who can prescribe antidepressants or refer the patient to a specialist), a psychologist (who only offers talk therapy), or a psychiatrist (who can prescribe/conduct both treatments, but may prefer antidepressants for financial reasons). As patients ultimately select both their doctor and treatment, the agency we've assigned patients in our model seems appropriate. That said, we recognize that because doctor recommendations are unmodeled and may influence decision-making, estimated patient preferences are likely influenced and should be interpreted as such. We also remind readers that all simulations are conducted in a partial equilibrium setting. This limitation is most notable in Section 6.4, where we conceive of a new medical treatment that could, in theory, have strong employment effects. Clearly, both medical care and labor markets are likely to respond to this counterfactual. Finally, we are careful not to simulate long-run effects. Our model is estimated using just two years of data and the permanent unobserved types revealed in estimation are strong determinants of mental health, treatment, and labor force participation. It is certainly possible that these types are more flexible over a longer time horizon.

References

Ben-Akiva, Moshe E, and Steven R Lerman. 1985. *Discrete choice analysis: theory and application to travel demand.* Vol. 9. MIT press.

Chan, T.Y., B.H. Hamilton, and N.W. Papageorge. 2015. "Health, risky behaviour and the value of medical innovation for infectious disease." *The Review of Economic Studies* 83 (4): 1465–1510.

Cronin, C.J., M.P. Forsstrom, and N.W. Papageorge. 2020. "What good are treatment effects without treatment? Mental health and the reluctance to use talk therapy." *NBER, 27711.*

Deaton, A. 2009. "Instruments of Development: Randomisation in the Tropics, and the Search for the Elusive Keys to Economic Development." In *Proc. of the British Academy,* 162:123–160.

Eckstein, Zvi, Michael Keane, and Osnat Lifshitz. 2019. "Career and family decisions: Cohorts born 1935–1975." *Econometrica* 87 (1): 217–253.

Elkin, Irene, Tracie Shea, John Watkins, et al. 1989. "NIMH Treatment of Depression Collaborative Research Program." *Archives of General Psychiatry* 46:971–982.

Fournier, Jay C, Robert J DeRubeis, Steven D Hollon, Sona Dimidjian, Jay D Amsterdam, Richard C Shelton, and Jan Fawcett. 2010. "Antidepressant drug effects and depression severity: a patient-level meta-analysis." *Jama* 303 (1): 47–53.

Hans, Eva, and Wolfgang Hiller. 2013. "Effectiveness of and Dropout From Outpatient Cognitive Behavioral Therapy for Adult Unipolar Depression: A Meta-Analysis of Nonrandomized Effectiveness Studies." *Journal of Consulting and Clinical Psychology* 81 (1): 75–88.

Heckman, J., and B. Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52 (2): 271–320.

Heckman, James J. 1991. "Randomization and social policy evaluation." *National Bureau of Economic Research, Working Paper No. 107.*

Kasahara, Hiroyuki, and Katsumi Shimotsu. 2009. "Nonparametric identification of finite mixture models of dynamic discrete choices." *Econometrica* 77 (1): 135–175.

Keane, M.P., and K.I. Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy* 105 (3): 473–522.

Magnac, Thierry, and David Thesmar. 2002. "Identifying dynamic discrete decision processes." *Econometrica* 70 (2): 801–816.

Mroz, Thomas A. 1999. "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome." *Journal of Econometrics* 92 (2): 233 –274.

Peter G. Peterson Foundation. 2020. *Medical Care Prices.* https://www.pgpf.org/chart-archive/0254_medical_care_prices. (Accessed: 05-04-2020).

Pigott, H.E, A.M. Leventhal, G.S. Alter, and J.J. Boren. 2010. "Efficacy and Effectiveness of Antidepressants: Current Status of Research." *Psychotherapy and Psychosomatics* 79:267–279.

Stewart, Rebecca E, and Dianne L Chambless. 2009. "Cognitive–behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies." *Journal of consulting and clinical psychology* 77 (4): 595.

Swift, J., and R. Greenberg. 2012. "Premature discontinuation in adult psychotherapy: A meta-analysis." *Journal of consulting and clinical psychology* 80 (4): 547.

Teachman, B., D. Drabick, R. Hershenberg, D. Vivian, B. Wolfe, and M. Goldfried. 2012. "Bridging the gap between clinical research and clinical practice: introduction to the special section." *Psychotherapy* 49 (2): 97–100.

Wamhoff, Steve, and Matthew Gardner. 2019. *Who Pays Taxes in America in 2019?* Technical report. Institute on Taxation and Economic Policy.

Wooldridge, J.M. 2005. "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity." *J. of Applied Econometrics* 20 (1): 39–54.

Yang, Z., D.B. Gilleskie, and E.C. Norton. 2009. "Health insurance, medical care, and health outcomes: A model of elderly health dynamics." *Journal of Human Resources* 44 (1): 47–114.