

A COMPARISON OF
SATISFACTION, SERVICES CHARACTERISTICS AND OUTCOMES IN THE
FULL SERVICE PARTNERSHIP PROGRAMS RELATIVE TO USUAL CARE

Petris Report # 2010-1

Nicholas C. Petris Center on Health Care Markets and Consumer Welfare
School of Public Health
University of California, Berkeley

May 2010

Timothy T. Brown, PhD, Associate Director of Research, Petris Center
Sun-Soon Choi, MS, Statistician, Petris Center
Jessica Chung, PhD, Statistician, Petris Center
Mistique Felton, MPH, Senior Research Associate, Petris Center
Richard M. Scheffler, PhD, Director of the Petris Center/Distinguished
Professor of Health Economics & Public Policy

Table of Contents
(Petris Report # 2010-1)

Acknowledgements 3

Summary 4

Chapter 1: Introduction 6

Chapter 2: Data and Methods 10

Chapter 3: Consumer Outcomes 12

Chapter 4: Consumer Experience with Public Mental Health Services..... 14

Technical Appendix..... 16

References..... 27

Acknowledgements

This project was jointly funded by the California Department of Mental Health (Contract No: 08-78106-000) and the California HealthCare Foundation (Award No. 04-1616). We would like to thank the following individuals for providing comments on earlier versions of this report:

Neal Adams, MD, MPH, *Associate Medical Director, United Behavioral Health*

Susan L. Ettner, PhD, *Professor in the Division of General Internal Medicine and Health Services Research in the UCLA Department of Medicine and in the Department of Health Services in the UCLA School of Public Health*

David Salkever, PhD, *Professor Emeritus, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health.*

Any remaining errors or omissions are the sole responsibility of the authors.

Summary

In November 2004, Californians approved the ballot measure Proposition 63 (which became the Mental Health Services Act) to expand public mental health funding and services. This report focuses on one subcomponent of the MHSA, the Full Service Partnership (FSP), which is part of the Community Services and Supports component.

In general, FSP programs are designed to provide an expanded range of services and supports within a recovery framework with programs tailored to meet the specific needs of children, transitional age youth, adults, and older adults who have been unserved or underserved. For the adult and older adult populations, which are the focus of this study, this included a focus on serving individuals most severely impacted by their mental health needs. Because of this, a simple comparison of the outcomes between those in FSP programs and those receiving usual care in the public mental health system would be misleading as it would not be an “apples-to-apples” comparison. Individuals enrolled in FSP programs are different in important ways from those receiving usual care and it is not possible, without using appropriate methodology, to determine if differences in outcomes between those in FSP programs and those receiving usual care are due to FSP participation or due to other ways in which these groups differ.

The best way to obtain an apples-to-apple comparison is through a randomized controlled trial where individuals are randomized to treatment and control groups. The purpose of randomization is to make the characteristics of individuals in the treatment group and the characteristics of individuals in the control group essentially the same such that the two groups only vary with respect to the treatment. When this is true, the difference in outcomes between the treatment and control group can be legitimately interpreted as the effect of the treatment (in this case FSP participation). Unfortunately, such a trial is prohibitively expensive and would take many years to complete.

However, it is possible to statistically mimic a randomized controlled trial in order to come close to an apples-to-apples comparison. Instead of assigning consumers to treatment and control groups using random numbers, we statistically assign them to treatment and control groups using data that are related to the likelihood that they enter an FSP program, but not related to the end-points we are studying. Similar to randomization, this approach allows us to infer the extent to which any differences in outcomes are due to being enrolled in an FSP program rather than being due to some other reason. For example, individuals may be more likely to receive public mental health services if they live closer to the location where such services are provided. However, how close individuals live to the location providing services would not be expected to affect the outcome of using such services, other things equal. In this case, distance from the service location serves to “randomize” people who receive or not receive services. Since individuals who live at various distances from the service location are unlikely to differ in important ways, distance serves a function similar to randomization.

By attempting to statistically mimic randomized controlled trials, we have determined the outcomes of participation in FSP programs for adults and older adults on one overall global rating scale, four consumer outcomes, and three service characteristics, in each case comparing FSP programs to usual care in the California public mental health system. The table below summarizes our findings which are all based on data from a multi-year survey of consumers receiving services in the California public mental health system

Table 1. Summary of Evaluation Findings: FSP vs. Usual Care

Measure	Evaluation Finding
Global Rating	
General satisfaction	FSP participants are 27% more satisfied (or higher)
Outcomes of Services	
Outcomes of services	FSP participants perceive 30% better outcomes*
Functioning	FSP participants perceive 27% better functioning
Social connectedness	FSP participants perceive 59% less connectedness [†]
Arrests	FSP participants are 56% less likely to be arrested [†]
Characteristics of Services	
Access to services	FSP is not different from usual care
Quality and appropriateness	FSP participants perceive 28% higher quality (or more)
Participation in treatment planning	FSP is not different from usual care

*This is not an average treatment effect and only applies to some FSP participants.

[†]See chapter 3 for a detailed explanation of this finding.

[†]More precisely, FSP participants are less likely to be arrested by 56 *percentage points*.

In summary, compared to usual care, FSP programs produce higher general satisfaction, higher functioning, fewer arrests, better outcomes of services in many cases, but do not seem to improve social connectedness according to the social connectedness measure used (see chapter 3 for a detailed discussion). Access to care and participation in treatment planning does not vary between FSP programs and usual care, but FSP participants experience higher quality services compared to those receiving usual care. At this time, it appears that the FSP program for adults and older adults is highly successful.

Chapter 1: Introduction

In November 2004, Californians approved the ballot measure Proposition 63 (which became the Mental Health Services Act) to expand public mental health funding and services. This report focuses on one subcomponent of MHSA, the Full Service Partnership (FSP), which is part of Community Services and Supports (CSS) component of MHSA. The CSS component provides funding for direct services and supports to people with a serious mental illness (SMI) or a serious emotional disturbance (SED).

Full Service Partnerships, according to the California Code of Regulations (Title 9, § 3620, 2010), may include the following services for adults:

Full Service Partnership Service Category.

(a) The County shall develop and operate programs to provide services under the Full Service Partnership Service Category. The services to be provided for each client with whom the County has a full service partnership agreement may include the Full Spectrum of Community Services necessary to attain the goals identified in the Individual Services and Supports Plan (ISSP). The services to be provided may also include services the County, in collaboration with the client, and when appropriate the client's family, believe are necessary to address unforeseen circumstances in the client's life that could be, but have not yet been included in the ISSP.

(1) The Full Spectrum of Community Services consists of the following:

(A) Mental health services and supports including, but not limited to:

- (i) Mental health treatment, including alternative and culturally specific treatments.*
- (ii) Peer support.*
- (iii) Supportive services to assist the client, and when appropriate the client's family, in obtaining and maintaining employment, housing, and/or education.*
- (iv) Wellness centers.*
- (v) Alternative treatment and culturally specific treatment approaches.*
- (vi) Personal service coordination/case management to assist the client, and when appropriate the client's family, to access needed medical, educational, social, vocational rehabilitative and/or other community services.*
- (vii) Needs assessment.*
- (viii) ISSP development.*
- (ix) Crisis intervention/stabilization services.*
- (x) Family education services.*

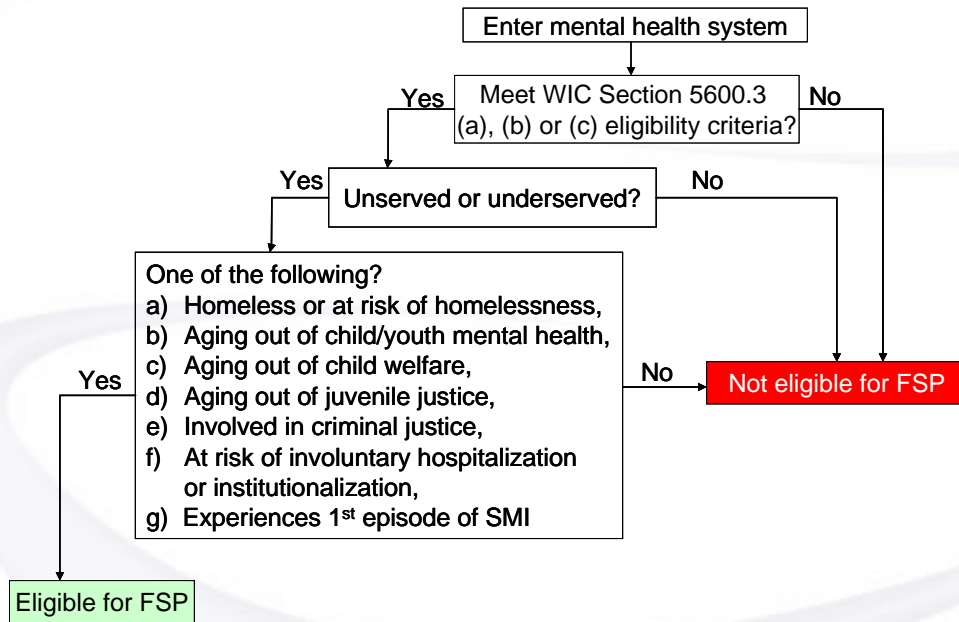
(B) Non-mental health services and supports including, but not limited to:

- (i) Food.*
- (ii) Clothing.*
- (iii) Housing, including, but not limited to, rent subsidies, housing vouchers, house payments, residence in a drug/alcohol rehabilitation program, and transitional and temporary housing.*
- (iv) Cost of health care treatment.*
- (v) Cost of treatment of co-occurring conditions, such as substance abuse.*
- (vi) Respite care.*

(C) Wrap-around services to children in accordance with WIC Section 18250 et. seq.

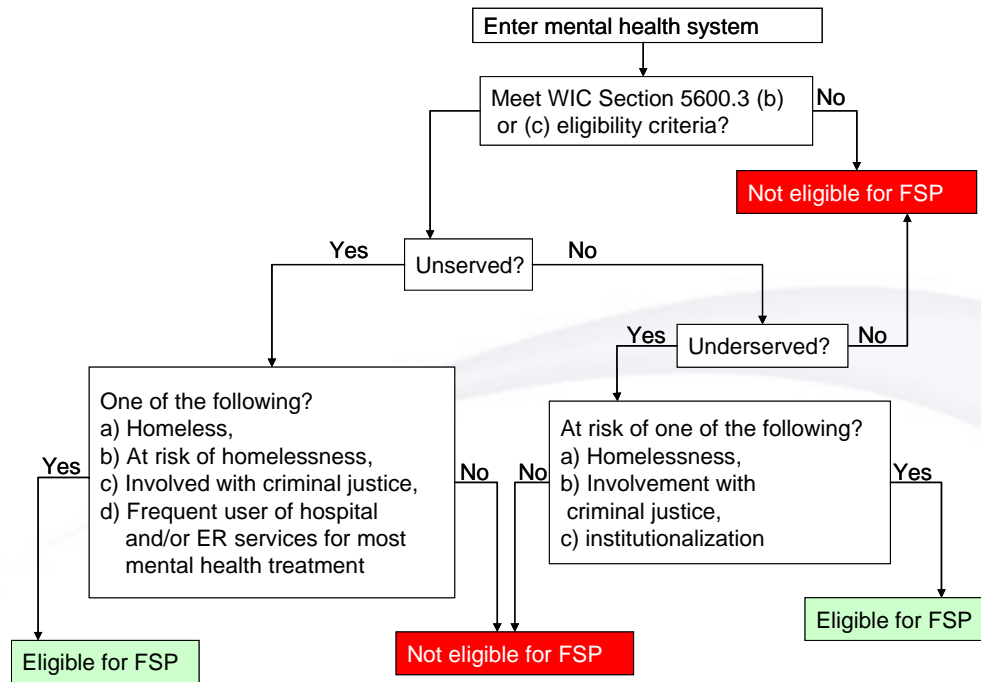
Figures 1.1, 1.2 and 1.3 show the criteria for admission into FSPs for transition age youth, adults and older adults respectively. First, a person must meet the eligibility criteria for mental health services as defined in WIC Section 5600.3 (a), (b) or (c). Next, an individual must be unserved or underserved. Unserved is defined as someone with an SMI or SED who is not receiving mental health services. People who have only had emergency or crisis-oriented contact and/or services are considered unserved. The definition of underserved is extremely broad, including anyone with an SMI or SED who does not receive services to support their wellness, recovery or resilience (California code of regulations, Title 9, § 3200.300, 2010). The last criteria that participants must meet varies by age group but can include: homelessness, at risk of homelessness, involvement or at risk of involvement with the criminal legal system, at risk of institutionalization, frequent users hospitals and/or emergency room treatment for mental health care, or for transition age youth, aging out of the child and youth mental health system, child welfare system or juvenile legal system (California code of regulations, Title 9, § 3620.05, 2010)

Figure 1.1: FSP Criteria for Transition Age Youth



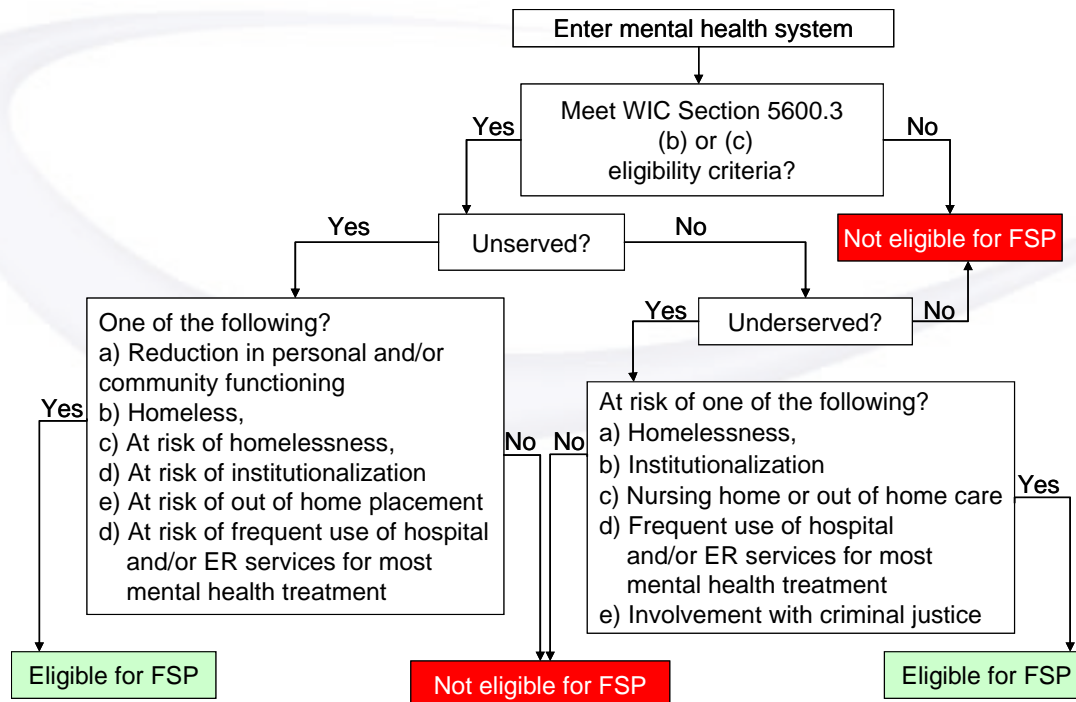
Notes: Petris Center Analysis of the California Code of Regulations, Title 9, Section 3620.05. FSP=Full Service Partnership.

Figure 1.2: FSP Criteria for Adults



Notes: Petris Center Analysis of the California Code of Regulations, Title 9, Section 3620.05. FSP=Full Service Partnership.

Figure 1.3: FSP Criteria for Older Adults



Notes: Petris Center Analysis of the California Code of Regulations, Title 9, Section 3620.05. FSP=Full Service Partnership.

Satisfaction, Services Characteristics and Outcomes

In this report we answer the following questions:

- (1) Do consumers participating in the Full Service Partnership Program have higher satisfaction with services received compared to those receiving usual care?
- (2) How do consumers participating in the Full Service Partnership Program rate the characteristics of the services they receive compared to those receiving usual care?
- (3) Do consumers participating in the Full Service Partnership Program have better outcomes as compared with consumers receiving usual care?

Answers to these questions will enable policymakers to determine the effectiveness of the FSP program.

Chapter 2: Data and Methods

The data for this analysis come from the Consumer Perception Survey (CPS) for adults and older adults (California Department of Mental Health, 2008a). Note that adults and older adults are evaluated as a single group in order to maximize sample size.

Data

We analyzed measures of consumer outcomes, characteristics of public mental health services, and satisfaction constructed from data from the Mental Health Statistics Improvement Program (MHSIP) portion of the CPS. Consumer outcomes are measured using the following domains: perception of outcomes of services, perception of functioning, and perception of social connectedness. Characteristics of the services received by consumers are measured by the following domains: perception of quality and appropriateness, perception of participation in treatment and planning, and perception of access. Finally, the extent to which consumers are satisfied with public mental health services, including whether they would choose public mental health services if they had other choices and would recommend public mental health services to their friends, is measured by the following domain: general satisfaction (California Department of Mental Health, 2009a, National Association of State Mental Health Program Directors Research Institute, 2009).

In addition to the outcomes listed above, we also use information from the Quality of Life portion of the CPS to measure criminal justice involvement. This is self-reported information on whether a consumer has been arrested in the previous month, and if so, how many times they have been arrested (California Department of Mental Health, 2008a).

Other data used in our analysis include the age, sex, and race of consumers which is available from the CPS. In addition we include the psychiatric diagnosis of each consumer (schizophrenia, bipolar disorder, depression, anxiety and related disorders, personality disorder, substance abuse (including alcohol abuse), unable to diagnose, and other disorders) from the Client and Service Information System (CSI). Information about whether each consumer was enrolled in a Full Service Partnership (FSP) program came from the Data Collection and Reporting System (DCR). Finally, all analyses include indicators of the county within which each consumer receives services.

Sampling

The Consumer Perception Survey is fielded for a period of approximately two weeks twice per year (beginning in May and November) and attempts to capture all individuals receiving public mental health services during the two-week window (California Department of Mental Health, 2008b). In addition, according to the Department of Mental Health, all individuals participating in Full Services Partnerships are required to participate in the survey (California Department of Mental Health, 2009b). The surveys are available in the following languages: English, Spanish, Chinese (Traditional), Hmong, Russian, Tagalog, and Vietnamese (California Department of Mental Health, 2008a).

This approach yields an approximate random sample of individuals being served in the California public mental health system, with the exception of individuals who are in acute hospitals, psychiatric health facilities, being treated for a crisis (stabilization, residential, and intervention), in jail or a jail hospital setting, or in long-term care institutional placements (e.g., state hospitals, Institute for Mental Disease) (California Department of Mental Health, 2008b). Therefore, the findings of this report do not apply to those in long-term care institutional placements. However, to the extent that receiving services in settings that are

temporary rather than permanent (acute hospitals, psychiatric health facilities, crisis treatment settings (stabilization, residential, and intervention), jails, jail hospitals) and are randomly distributed across the public mental health population and across time, the exclusion of individuals in these settings will have little effect on the findings presented in this report.

We use CPS data from May 2005 to May 2008 (7 waves). Not all California counties are represented in the data. During the period this analysis was conducted, data on participation in Full Service Partnerships was only available for 43 of California's 58 counties: Butte, Calaveras, Colusa, Contra Costa, Del Norte, El Dorado, Fresno, Glenn, Humboldt, Imperial, Inyo, Kern, Lake, Los Angeles, Madera, Mariposa, Merced, Mono, Napa, Nevada, Orange, Plumas, Sacramento, San Benito, San Bernardino, San Diego, San Francisco, San Joaquin, San Luis Obispo, San Mateo, Santa Clara, Santa Cruz, Shasta, Sierra, Siskiyou, Stanislaus, Trinity, Tulare, Tuolumne, Ventura, Yolo, and Sutter/Yuba. The extent to which consumers and FSP programs in the remaining 15 counties are substantially the same as the average consumer and the average FSP program in the 43 counties included in the analysis is the extent to which the findings of this analysis will apply to those 15 counties on which no data was available.

Statistical Methods

The goal of this project is to compare the effects of receiving FSP services on measures of outcomes, service characteristics, and satisfaction, when all other consumer characteristics, county characteristics, and time characteristics are held constant. In order to accomplish this goal we use quasi-experimental methods. Quasi-experimental methods are experimental in that they compare a treatment group (consumers enrolled in an FSP program) to a control group (consumers not enrolled in an FSP). The "quasi" aspect in quasi-experimental refers to the fact that individual are not randomized into treatment and control groups, which is the case in true experiments. The purpose of randomization in experiments is to ensure that consumer characteristics are the same in the treatment and control groups such that the only difference between the groups is whether they receive treatment. Using a true experimental design would make it possible to infer that any differences in the outcomes, services characteristics, and satisfaction measures of individuals participating in FSP programs relative to individuals participating in usual care in the public mental health system are due to being enrolled in FSP programs and not due to some other reason.

As randomization is not possible in the current context, we apply statistical methods that allow us to make similar inferences without actual randomization. Instead of assigning consumers to treatment and control groups using random numbers, we statistically assign them to treatment and control groups using data that are related to the likelihood that they enter an FSP program, but are not related to the end-points we are studying (outcomes, service characteristics, and satisfaction). Similar to randomization, this approach allows us to infer that any differences in outcomes, services characteristics, and satisfaction measures are due to being enrolled in an FSP program and are not due to some other reason. See Section A.2.1 of the Technical Appendix for the specific details of our methodology.

Chapter 3: Consumer Outcomes

Consumer outcomes are the fundamental measures of success for any mental health program. We measured consumer outcomes using data from the California Perception Survey (CPS) for the combined group of adults and older adults in four areas: perception of outcomes of services, perception of functioning, perception of social connectedness, and the probability of being arrested during the previous month. These outcomes relate to two of the goals of the Mental Health Services Act: prolonged suffering and incarceration. All outcomes are based on data reported by consumers. Each finding is from a statistical model that controls for age, sex, race, psychiatric diagnosis, time in the public mental health system, county of residence, and FSP participation. Other individual characteristics are controlled for using specific statistical techniques.¹

Perception of Outcomes of Services

This domain seeks to determine overall life functioning. It is derived from consumer evaluation of the following eight statements and focuses on problem solving, self-control, crisis management, social effectiveness, housing, and psychiatric symptoms.

As a direct result of the services I received:

1. I deal more effectively with daily problems.
2. I am better able to control my life.
3. I am better able to deal with crisis.
4. I am getting along better with my family.
5. I do better in social situations.
6. I do better in school and/or work.
7. My housing situation has improved.
8. My symptoms are not bothering me as much.

Each question is evaluated according to a five-point scale where “strongly disagree” is equal to one and “strongly agree” is equal to five. The total score was determined by summing the numerical answers to each question and dividing by the total number of questions answered which yields a number that also ranges from one to five. Scores were only calculated for observations where at least two-thirds of the items were responded to.

Our analysis indicates that individuals participating in an FSP have outcomes that are 30% better than those receiving usual care. However, the above figure only applies to some consumers. Other FSP consumers show variation in the outcomes achieved. See section A2.3.1 of the Technical Appendix for additional details.

Perception of Functioning

This domain is similar to the domain above, but is more inward focused. It is derived from consumer evaluation of the following five statements.

As a direct result of the services I received

1. I do things that are more meaningful to me.
2. I am better able to take care of my needs.
3. I am better able to handle things when they go wrong.

¹ See section A2.1 of the Technical Appendix for details of the statistical analysis.

4. I am better able to do things that I want to do.
5. My symptoms are not bothering me as much.

The total score for the domain is determined in the same way as for the perception of outcomes of services above.

Our analysis finds that the effect of FSP programs on perceptions of functioning is 37% higher relative to those receiving usual care. This is an average treatment effect. An average treatment effect is the overall average additional effect that a person would receive from participating in an FSP program relative to participation in usual care. See section A2.3.2 of the Technical Appendix for details of the statistical analysis.

Perception of Social Connectedness

This domain focuses on satisfaction with friends, social support, and belonging and is derived from consumer evaluation of the following four statements.

As a direct result of the services I received:

1. I am happy with the friendships I have.
2. I have people with whom I can do enjoyable things.
3. I feel I belong in my community.
4. In a crisis, I would have the support I need from family or friends.

The total score for the domain is determined in the same way as for the perception of outcomes of services above.

Our analysis indicates that individuals participating in an FSP program have outcomes that are at least 59% worse than those receiving usual care. This is an average treatment effect. The reason for this finding lies in understanding the individual questions. Using the same methods to analyze each component of this measure as were used to analyze the overall measure, we find no difference between FSP participants and those receiving usual care for questions one and two. We find a large negative difference for FSP participants relative to those receiving usual care for questions three and four. This finding should be studied in more detail in future work. A potential reason for this finding with respect to question three is that the increased interaction with the community that many FSP enrollees experience may be result in new experiences that cause them to feel less belonging in the community (possibly due to stigma). The finding for question four suggests that FSP participants may understand the question to be referring to a mental health crisis, a situation for which they would have been taught to seek professional care, care which family and friends are not ordinarily qualified to provide. See section A2.3.3 of the Technical Appendix for details of the statistical analysis.

Arrests

This is simply a count of how many times a consumer has been arrested in the last month. The question asked is as follows:

1. In the past MONTH, how many times have you been arrested for any crimes?
where the answers are “none”, “1 arrest”, “2 arrests”, “3 arrests”, “4 or more arrests”.

Our analysis indicates that individuals participating in an FSP program are approximately 56 percentage points less likely to be arrested relative to those receiving usual care in the public mental health system. This is an average treatment effect. See section A2.3.4 of the Technical Appendix for details of the statistical analysis.

Chapter 4: Consumer Experience with Public Mental Health Services

The relative experiences of consumers in one program as compared to another are indicators of how quality may vary between programs. Programs can vary in terms of the breadth of services and supports offered as well as in terms of the quality and availability of those interventions.

We measured the experiences of consumers of public mental health services in FSP programs relative to usual care in the public mental health system using data from the Consumer Perception Survey (CPS) for adults and older adults in four areas: perception of quality and appropriateness, perception of participation in treatment planning, perception of access, and general satisfaction. All outcomes are based on data reported by consumers. Each finding is from a statistical model that controls for age, sex, race, psychiatric diagnosis, time in the public mental health system, county of residence, and FSP participation. Other individual characteristics are controlled for using specific statistical techniques. See section A2.4 of the Technical Appendix for details of the statistical analysis.

Perception of Quality and Appropriateness

This domain seeks to determine quality from a recovery-oriented perspective. It is derived from consumer evaluation of the following nine statements and focuses on the way in which staff treats consumers with respect to encouragement, cultural background, treatment information, privacy, and openness to criticism.

1. Staff here believe that I can grow, change and recover.
2. Staff encouraged me to take responsibility for how I live my life.
3. Staff were sensitive to my cultural/ethnic background.
4. Staff helped me obtain the information needed so that I could take charge of managing my illness.
5. Staff told me what side effects to watch for.
6. I was encouraged to use consumer-run programs (support groups, drop-in centers, crisis phone line, etc.).
7. I was given information about my rights.
8. Staff respected my wishes about who is, and is not to be given information about my treatment.
9. I felt free to complain.

Each question is evaluated according to a five-point scale where “strongly disagree” is equal to one and “strongly agree” is equal to five. The total score was determined by summing the numerical answers to each question and dividing by the total number of questions answered which yields a number that also ranges from one to five. Scores were only calculated for observations where at least two-thirds of the items were responded to.

This domain focuses on characteristics that are relatively objective which make this a strong measure of overall consumer experience. Our analysis indicates that FSP participants receive 28% higher quality care than those receiving usual care and in some cases the quality difference is even higher than 28%. See section A2.4.1 of the Technical Appendix for details of the statistical analysis.

Perception of Participation in Treatment Planning

This domain is short and straightforward and seeks to determine the level to which consumers participate in their treatment planning, a fundamental aspect of recovery-oriented treatment. It is derived from consumer evaluation of the following two statements.

1. I, not staff, decided my treatment goals.
2. I felt comfortable asking questions about my treatment and medication.

The total score for the domain is determined in the same way as for the perception of quality and appropriateness above.

This domain focuses on characteristics that are relatively objective and make this a relatively strong measure of consumer experience with respect to the role they play in directing the services they need to support their recovery. Our analysis indicates that there is no statistical difference in the perceptions of participation in treatment planning between individuals participating in an FSP program and those participating in usual care. See section A2.4.2 of the Technical Appendix for details of the statistical analysis.

Perception of Access to Services

This domain seeks to measure access in terms of location, time and frequency of appointments, responsiveness, and adequacy of services. It is derived from consumer evaluation of the following six statements.

1. The location of services was convenient.
2. Staff were willing to see me as often as I felt it was necessary.
3. Staff returned my calls within 24 hours.
4. Services were available at times that were good for me.
5. I was able to get all the services I thought I needed.
6. I was able to see a psychiatrist when I wanted to.

The total score for the domain is determined in the same way as for the perception of quality and appropriateness above.

This domain focuses on characteristics that are relatively objective. Our analysis indicates that there is no statistical difference in the perceptions of access to care between individuals participating in an FSP program and those participating in usual care. See section A2.4.3 of the Technical Appendix for details of the statistical analysis.

General Satisfaction

This domain seeks to measure overall satisfaction. It is derived from consumer evaluation of the following three statements.

1. I like the services that I received here.
2. If I had other choices, I would still get services from this agency.
3. I would recommend this agency to a friend or family member.

The total score for the domain is determined in the same way as for the perception of quality and appropriateness above.

This domain focuses on characteristics of access that are completely subjective and depend on the extent and depth of consumers' knowledge of alternative sources of care. Our analysis indicates that individuals participating in an FSP have 27% higher satisfaction relative to those receiving usual care with some FSP participants experiencing even higher satisfaction. See section A2.4.4 of the Technical Appendix for details of the statistical analysis.

Technical Appendix

This appendix presents the statistical methodology used in this report and the detailed results of each of the analyses presented in chapters three and four.

A2.1 Econometric Methods

We used a quasi-experimental approach including a treatment group of consumers participating in Full Service Partnership (FSP) programs and a control/comparison group of consumers not participating in an FSP program. Both groups were receiving treatment within the California public mental health system. Since randomization to treatment and control groups was not possible, we used the instrumental variable (IV) technique to statistically assign consumers to treatment and control groups. This method is designed to correct bias arising from measurement error, omitted variables, and endogeneity that may affect the estimated relationship between FSP participation and the various dependent variables of interest. Using this method, we attempt to estimate the average treatment effect (ATE) of participating in an FSP program on each of these dependent variables of interest. The ATE estimates the overall average additional effect that a person would receive from participating in an FSP program relative to participation in usual care.

In our analyses, we primarily use the two-stage limited information maximum likelihood (LIML) estimator. The LIML estimator provides finite-sample bias reduction which allows us to achieve more reliable estimates in an instrumental variables context (Angrist and Pischke, 2009). The two stages are estimated simultaneously in order to obtain valid standard errors (Baum *et al.*, 2007). Our instruments must meet the following assumptions in order to be valid:

Assumption One: The instrument set Z is correlated with participation in FSP:

$$\text{Cov}(Z, FSP) \neq 0$$

Assumption Two: The instrument set Z is uncorrelated with the error term u in the second stage: $\text{Cov}(Z, u) = 0$

We test assumption one using the criteria developed by Stock and Yogo (2005) for LIML estimators (note that the critical values when using LIML are generally lower than when using two-stage least squares). In order to be valid, the set of instruments must be sufficiently strong.

We test assumption two using Hansen's J test to evaluate the overidentifying restrictions of our equation below (Hayashi, 2000). The logic of the overidentification test is that if at least one of the identifying instruments is valid then the test can evaluate whether the other instrument(s) are invalid (correlated with the error term in the second stage).

In each of the analyses below, we use as our assumed valid instrument the fact that due to the various ways in which the public mental health system is organized (i.e., holidays, administrative procedures, and other reasons), there are consistently more consumers whose date of entry into the public mental health system, as defined by the date that they first receive public mental health services according the Client and Services Information System, is on an odd day rather than on an even day. Table A1 shows this fact annually over the 4-year period represented by our data. This same phenomenon also appears at the county level in our data with only rare exceptions for some years within some counties.

We assume that this phenomenon results in a near random draw of consumers with respect to their consumer characteristics, including the severity of their psychiatric condition(s). If more consumers enter on an odd day than an even day, then the less likely more severe cases are to stand out on such days. If this is the case,

consumers with more severe psychiatric conditions who enter on odd days would be mathematically less likely to enter an FSP program because they are less likely to be identified as candidates. The reverse would be true for consumers who entered on an even day. This assumption is supported by a positive and statistically significant correlation in our first-stage regressions between entering the public mental health system on an even day and eventual entrance into an FSP program.

Finally, we reasonably assume that whether a consumer enters the public mental health system on an odd or even day has no impact on their perceptions of outcomes from services received or their perceptions of the characteristics of such services. We refer to this instrument as odd/even entry.

Table A1: Date of First Mental Health Service: Odd/Even Days

Odd/Even Days	Year			
	2005	2006	2007	2008
<i>Odd Days</i>	18,139	19,652	17,254	9,770
<i>Even Days</i>	14,194	15,723	14,336	8,118

Source: combined data from the CSI and the CPS.

An additional instrument includes an index of the type of help a consumer may have received in filling out their Consumer Perception Survey (CPS). We hypothesize that consumers who ask questions or otherwise indicate that they need help in understanding and answering survey questions are more likely to come to the attention of the staff and thus enter an FSP, assuming they otherwise meet the qualifications. The index gives higher values to those who have more decision making power or influence regarding whether a consumer enters an FSP program. The index is constructed as follows. Family members and other consumers (*e.g.*, friends) receive a zero. Family and friends give extremely valuable support to consumers, but are not decision makers regarding entry into an FSP program. Needing no help to answer the survey receives a one. Receiving help from another mental health consumer or a volunteer/mental health advocate receives a two. Receiving help from a staff member receives a three. Receiving help from a clinician receives a four. Finally, receiving help from a professional interviewer receives a five. We refer to this instrument as the help index. While obtaining help on a survey is associated with a probability of entering in FSP, we do not expect that the receipt of such help will be related to treatment outcomes or treatment characteristics. We test this hypothesis using an overidentification test that also includes our assumed valid instrument of odd/even entry defined above.

A third instrument is the season of a consumer's birth. Over 100 studies have examined the relationship between mental illness and season of birth (Castrogiovanni *et al.*, 1998). While the causes for this relationship have not been fully explained, the relationship between mental illness and season of birth does exist. Those with more severe unipolar depression are more likely to be born in the spring (*e.g.*, Fountoulakis *et al.*, 2007). Those with non-deficit schizophrenia are more likely to be born in the winter (*e.g.*, Kirkpatrick *et al.*, 1998). Those with more severe bipolar disorder are more likely to be born in the spring and the autumn (*e.g.*, Rihmer, 1980). All of the above suggests that those born in June, July, August, and September will be consumers who have less severe forms of these mental illnesses. Thus our instrument is not being born in these months.

Because such consumers not born during these months are overrepresented among consumers, such consumers will also be more likely to enter an FSP program, other things equal. Season of birth has been suggested or used as an instrumental variable in multiple studies in other areas of economics (Angrist and Krueger, 1991, 1992, 1995, 2001; Neal and Johnson, 1996; Staiger and Stock, 1997; Levin and Plug, 1999; Gelbach, 2002; Chamberlain and Imbens, 2004; Honoré and Hu, 2004; Skirbekk, Kohler, and Prskawetz, 2004; Chesher, 2005; Cruz and Moreira, 2005; Imbens and Rosenbaum, 2005; Chernozhukov and Hansen, 2006; Lefgren and McIntyre, 2006; Dufour and Taamouti, 2007; Leigh and Ryan, 2008). While season of birth may also correlate with the socioeconomic status of one's parents (Buckles and Hungerman, 2008), there is no

evidence that season of birth correlates with treatment outcomes or treatment quality in the public mental health system. We evaluate this instrument using an overidentification test that also includes our assumed valid instrument. We refer to this instrument as season of birth.

A fourth instrument is a number assigned to each consumer in the California public mental health system that is based on the time at which the consumer entered the public mental health system and then divided by 1×10^6 . All consumers within a given county who entered on the same date will have the same number assigned to them, but numbers may vary by county. As noted above, while there are specific criteria for choosing consumers to participate in FSP programs, not all consumers who met these criteria were granted a space in the FSP program since spaces were limited over the time period we are examining. The logic of the instrument is as follows. On any given date, consumers enter the public mental health system. Some of these consumers are eligible for FSP participation, but will only be able to participate to the extent that spaces are available. Over time, more and more spaces become available as the implementation of FSPs expands such that consumers who enter the public mental health system at a later date have a greater probability of entering the FSP program, given that they are eligible for FSP participation. All consumer characteristics of consumers are thus random with respect to the date they entered the public mental health system. We test this instrument using an overidentification test that also includes our assumed valid instrument. We refer to this instrument as order of entry.

In addition to the above tests, a test of endogeneity is also performed to determine whether the potentially endogenous FSP variable is actually endogenous (Hayashi, 2000). All equations are estimated with robust standard errors to correct for the effects of heteroscedasticity on the estimation of standard errors.

Because the dependent variable is constrained by the structure of the Consumer Perception Survey (CPS) to vary from 1 to 5 and because some consumers may have desired to choose values less than 1 or greater than 5, we consider the dependent variable to be censored from both the left and the right. In order to account for this censoring we also estimate instrumental variable Tobit models. Such models allow us to appropriately scale the parameter on FSP participation and are only used for this reason. Using LIML instrumental variables alone yields results that are valid for testing the strength and exogeneity of instrumental variables and to test the endogeneity of FSP participation, but may yield parameters of FSP participation that are too large due to the lack of accounting for censoring. All instrumental Tobit models are estimated simultaneously and have robust standard errors.

Finally, to test whether the estimated effects of participation in an FSP are constant across consumers or vary across consumers, we perform a test of essential heterogeneity (Heckman *et al.*, 2006; Basu *et al.*, 2007). The instrumental variable technique will provide an estimate of the ATE of participation in an FSP program in two cases: (1) there is no unobserved heterogeneity (all consumers obtain the same benefit from FSP participation conditional on the covariates), or (2) there is unobserved heterogeneity (consumers with the same covariate value respond differently to the treatment), but FSP participation is *not* based on expected individual differences in treatment outcomes.

There is a third case: the situation where there is unobserved heterogeneity and consumers *are* selected into treatment based on expected individual differences in treatment outcomes. When this is true then there exists what is known as essential heterogeneity. If essential heterogeneity is present, the marginal effect of participation in an FSP program using the instrumental variable model *does not* represent the ATE.

However, in the current study, consumers are not selected into FSP programs based on expected individual differences in treatment outcomes, but rather based on specific regulations from the California Department of Mental Health as listed in chapter one of this report. Therefore we do not make assumptions about whether essential heterogeneity will exist or not in any given analysis, but simply test to determine whether it is present.

In the case where essential heterogeneity exists and we are unable to estimate the ATE, our estimates represent the weighted average of the local average treatment effects estimated one at a time (Angrist and Pischke, 2009, p. 173-175). This assumes that the theoretical criteria for a local average treatment effect (LATE) are met (Angrist and Pischke, 2009, p. 152-155). The LATE is the average outcome for consumers whose FSP status changes when the value of the instrument changes. When multiple instruments are used, the instrumental variable method produces a weighted average of the outcomes received by consumers whose consumer FSP status changes when the value of each instrument changes. In other words, it is the average effect for some consumers.

All analyses were performed using Stata 10.1. The following Stata routines were used: *regress*, *ivreg2*, *probit*, and *ivtobit*.

A2.2 Data Issues

Matching Across Datasets

We discovered a number of apparent data errors when matching data from the CPS with data on participation in FSP programs from the Data Collection and Reporting System (DCR) using the Client and Services Information System (CSI) number. Our matching assumptions are as follows: (1) any observations with missing county identification or missing CSI number were deleted; (2) any duplicated records with regard to county identification, CSI number, survey date, assessment date, date of partnership status change, and key event tracking (KET) status were deleted; (3) if a KET for discontinuation-interruption of FSP participation was not filed, it is assumed that the consumer is still in the FSP system; (4) any records with incorrect or inconsistent entry/exit sequences were excluded; (5) when a KET for either reestablishment or discontinuation-interruption was duplicated consecutively with the different date of partnership status change, the observation from only the first instance was kept and all subsequent multiple entries or exits were deleted (*e.g.*, if a KET for reestablishment was followed by another KET for reestablishment without a KET for discontinuation-interruption, or if a KET for discontinuation-interruption is followed by another KET for discontinuation-interruption without a KET for reestablishment, the second duplicated record was deleted); (6) if a record started with a KET for reestablishment without a KET for discontinuation-interruption, this record was also deleted due to the incorrect sequence (*e.g.*, if a partnership was established and reestablished without a KET for discontinuation-interruption, the observation was deleted); (7) if age was missing in the CPS dataset, but not missing in the DCR dataset, the value was replaced and adjusted with the value found in the DCR dataset based on the survey date; (8) if age existed but varied between the CPS and DCR, the observation was deleted; (9) if gender varied between the CPS and DCR, the observation was deleted; and (9) if a CSI number started with zeros or letters, the leading zeros or letters were deleted before merging with other datasets.

The treatment group consists of consumers who stayed in the FSP program for at least one month from the survey date. Thus, consumers who exited the FSP program and reentered were also included as long as the total duration was at least one month. The control group consists of consumers who meet one of the following criteria in the counties with FSP programs on the survey date: (1) consumers who have been in the FSP program for less than one month on the survey date, (2) consumers who have never been in the FSP program in counties where an FSP program is available, or (3) consumers who had exited the FSP program on the survey date (past FSP participants). We assumed that FSP treatment of less than one-month would not be adequately reflected in outcomes. If this assumption is incorrect, then our findings will be conservative: we will underestimate the true effect of FSP participation.

Finally, the specific measures used (for all but arrests) were calculated by recording the “not applicable” response as missing values and by excluding respondents when more than one third of the items were missing in

a given domain per the procedure recommended by the Department of Mental Health (California Department of Mental Health, 2009a; National Association of State Mental Health Program Directors Research Institute, 2009). Note that this procedure would not exclude a consumer from another domain where they were not missing more than one third of the items.

Psychiatric Diagnoses

Each time a consumer accesses public mental health services, a psychiatric diagnosis should be recorded in the CSI dataset for that consumer. As a result, our preliminary data analysis of principal and secondary diagnoses for both DSM-IV Axis I and Axis II shows that it is possible for each consumer to have multiple psychiatric diagnoses spanning the years of 2000-2008. To determine the appropriate way in which to code and analyze psychiatric diagnoses taken from the CSI, we consulted with Neal Adams, MD, MPH, former Director of Special Projects for the California Institute for Mental Health and former Medical Director for Adult Services with the California Department of Mental Health.

Based on this consultation, we made three simplifying assumptions regarding psychiatric diagnoses. First, we grouped psychiatric diagnoses into eight distinct categories: schizophrenia, bipolar disorder, depression, anxiety and related disorders, personality disorder, substance abuse (including alcohol abuse), unable to diagnose, and other disorders. Second, a consumer with a history of a given disorder at any time within their admission records in the CSI dataset was coded as “1” for that particular disorder. This approach captures each diagnosis without discarding potentially valuable and relevant information. Third, we collapsed principal and secondary diagnoses into one group to limit the number of covariates in the model to a reasonable size.

A2.3 Results: Consumer Outcomes

In this section we present the results pertaining to the perception of outcomes of services, perception of functioning, perception of social connectedness, and arrests. The instruments used are somewhat different across the analyses in order to satisfy the strict statistical criteria required.

A2.3.1 Perception of Outcomes of Services

We estimated the following equations,

$$OUTCOME = \beta_0 + \beta_1 \hat{FSP} + \beta_2 DEMOG + \beta_3 DIAG + \beta_4 COUNTY + \beta_5 TIME + \varepsilon \quad (1)$$

$$\Pr(FSP=1) = \alpha_0 + \alpha_1 IV + \alpha_2 DEMOG + \alpha_3 DIAG + \alpha_4 COUNTY + \alpha_5 TIME + \eta \quad (2)$$

where the two equations are estimated simultaneously.

OUTCOME is perception of service outcomes. *FSP* is equal to one if a consumer participated in an FSP program and zero otherwise. The variable \hat{FSP} refers to the predictions from equation (2). *DEMOG* contains a vector of variables for sex (male, female), race/ethnicity (white, black, Asian/Pacific Islander, Hispanic, other race), age (18-25, 26-34, 35-44, 45-54, 55-64, 65-74, 75 or older). *DIAG* is a vector of psychiatric diagnosis variables (schizophrenia, bipolar disorder, depression, anxiety and related disorders, personality disorder, substance abuse (including alcohol abuse), unable to diagnose, and other disorders). *COUNTY* contains a vector of county-level fixed effects. Note that counties differ in many ways (*e.g.*, weather, built environment, structure of policing, culture, *etc.*). Including county-level fixed effects allows us to control for all aspects of counties that do not vary over the time period covered by our data. *TIME* contains a vector of year fixed effects (2005, 2006, 2007, 2008). Finally, *IV* contains a vector of instrumental variables (odd/even entry, help index, and

season of birth). The reference group contains consumers who are not participating in an FSP program, male, aged 18-24, white, have a diagnosis of major depression, and were surveyed during the year 2005. The symbols ε and η refer to the error terms.

The analytic sample contained 73,923 observations including 1,417 observations of individuals participating in FSP programs. The joint strength of the instruments (odd/even entry, help index, season of birth) is such that maximal size bias using LIML is less than 10% (Kleibergen-Paap rk Wald F -statistic: 21.99; critical value: 6.46). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instruments are uncorrelated with the error term. Testing the help index against odd/even entry yields a Hansen's J statistic of 1.83, $p = 0.18$. The Hansen's J statistic when all three instruments are included is 2.63, $p = 0.27$. A test of the exogeneity of participation in an FSP program is rejected (χ^2 : 18.97; $p < 0.01$) indicating that the instruments are necessary.

Instrumental variables Tobit (IV-Tobit) estimates show a marginal effect of participation in FSP relative to usual care in the public mental health system that is 1.17 points higher (on a scale from 1 to 5) where the overall average value is 3.85. This is an approximate 30% improvement.

Essential heterogeneity was tested for using for a predicted propensity score (probability) from a probit model containing the three instruments and all other exogenous variables. A model of the marginal treatment effect on the perception of outcomes of services was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). Including all of the polynomials resulted in the square of the propensity score being statistically significant at the 5% level with a negative sign and the cube of the propensity score being statistically significant at the 5% level with a positive sign. This indicates that there is a significant variation in the effect of FSP treatment for this outcome measure. Thus our estimate does not represent the ATE. Consumer subgroups will vary in the outcomes received. The above estimate of improvement represents the weighted average of the outcomes received by consumers whose consumer FSP status changes when the values of each instrument changes.

A2.3.2 Perception of Functioning

This analysis is identical to that of the perception of outcomes of services in A2.3.1, only the dependent variable used is different. The analytic sample contained 30,167 observations including 1,244 observations of consumers participating in FSP programs. This is lower than the number of observations used in A2.3.1 as this domain was implemented beginning in May of 2007.

The joint strength of the instruments (odd/even entry, help index, season of birth) is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 18.27; critical value: 6.46). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instruments are uncorrelated with the error term. Testing the help index against odd/even entry yields a Hansen's J statistic of 2.35, $p = 0.13$. The Hansen's J statistic when all three instruments are included is 3.45, $p = 0.18$. A test of the exogeneity of participation in an FSP program is rejected (χ^2 : 11.22; $p < 0.01$) indicating that the instruments are necessary.

Instrumental variables Tobit estimates show a marginal effect of participation in FSP relative to usual care in the public mental health system that is 1.05 points higher (on a scale from 1 to 5) where the overall average value is 3.84. This is an approximate 27% improvement.

Essential heterogeneity was tested for using for a propensity score (probability) predicted from a probit model containing the three instruments and all other exogenous variables. A model of the marginal treatment effect on the perception of functioning was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). None of the polynomial terms was found to be statistically significant at the 5% level. This indicates the improvement presented above represents the ATE.

A2.3.3 Perception of Social Connectedness

This analysis is identical to that of the perception of outcomes of services in A2.3.1, only the dependent variable (perception of social connectedness) and instrument set are different. The analytic sample contained 31,700 observations including 1,473 observations of consumers participating in FSP programs. This is lower than the number of observations used in A2.3.1 as this domain was implemented beginning in May of 2007. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.2 reflect differences in the number of consumers evaluating each of the questions in the relevant domains and differences in data availability for the instruments used.

We used the following instruments: odd/even entry and order of entry. The strength of the instruments is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 13.61; critical value: 8.68. Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instrument is uncorrelated with the error term (Hansen's J statistic: 0.03, $p = 0.87$). A test of the exogeneity of participation in an FSP is rejected (χ^2 : 7.26; $p < 0.01$) indicating that the instruments are necessary.

Instrumental variables Tobit estimates show that the marginal effect of participation in FSP relative to usual care in the public mental health system that is 2.26 points lower (on a scale from 1 to 5) where the overall average value is 3.83. This is an approximate 59% reduction.

In order to determine the possible reasons for this reduction, the four components making up this measure were analyzed separately. We determined that there was no statistical difference between FSP participants and consumers receiving usual care for the first two questions: (1) I am happy with the friendships I have, and (2) I have people with whom I can do enjoyable things. In contrast, we found that FSP participants were statistically less likely than consumers receiving usual care to agree as strongly to the third and fourth statements that make up the measure: (3) I feel I belong in my community, and (4) In a crisis, I would have the support I need from family or friends. We speculate that consumers in FSP may perceive more stigma from the community due to increased interaction with the community. In addition we speculate that consumers in FSP may interpret the term "crisis" in question four to refer to mental health crises, in which case they would have been instructed to use the public mental health services available to them rather than look for the support they need from family and friends.

Essential heterogeneity was tested for using for a propensity score (probability) predicted from a probit model containing the two instruments and all other exogenous variables. A model of the marginal treatment effect on the perception of social connectedness was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). None of the polynomial terms was found to be statistically significant at the 5% level. This indicates the decline presented above represents the ATE.

A2.3.4 Arrests

Arrests are analyzed differently from the above outcomes because of the very nature of arrests. The great majority consumers will have zero arrests in any given period. Because of this, heteroscedasticity is highly likely to occur in any models that we use. One of the estimators used in all other analyses in this report, the Tobit estimator, yields inconsistent estimates in the presence of heteroscedasticity, which can result in bias to both the estimated parameters and standard errors (Greene, 2003). While this is not a problem for the other analyses in this report which use instrumental variables Tobit estimators, as the measures used in these other analyses appear to have well-behaved distributional properties, this is not the case for arrests. Because of this problem, we only use two-stage least squares estimates (2SLS) with a correction for heteroscedasticity. The parameter estimates of 2SLS are consistent in the presence of heteroscedasticity and standard errors can be adjusted to account for heteroscedasticity using standard techniques (Wooldridge, 2006). We also estimate LIML versions of this model which yields results virtually identical to 2SLS while allowing for more precise estimates when using instrumental variables (Angrist and Pischke, 2009).

An additional issue is that the measure of arrests is censored (top-coded) at 4 arrests. This means that using 2SLS or LIML to determine the number of arrests experienced by consumers will likely yield biased estimates. Therefore, we only estimate the probability of experiencing an arrest between consumers enrolled in an FSP program relative to those receiving usual care.

In order to statistically allocate consumers between the treatment (FSP) and the control (non-FSP) groups, we use as instruments odd/even entry and order of entry. Otherwise our estimation approach is similar to A.2.3.1 above.

The analytic sample contained 77,504 observations including 1,657 observations of consumers participating in FSP programs. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.1 reflect differences in the number of consumers evaluating this question relative to the questions used in A2.3.1 and differences in the availability of data for the instruments used. The joint strength of the instruments is such that maximal size bias using LIML is approximately equal to 10% (Kleibergen-Paap rk Wald F -statistic: 8.46; critical value: 8.68). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instrument is uncorrelated with the error term (Hansen's J statistic: 0.52, $p = 0.47$). A test of the exogeneity of participation in an FSP program is rejected (χ^2 : 5.21; $p < 0.05$) indicating that the instruments are necessary. The model shows that consumers participating in FSP programs are 56 percentage points less likely to be arrested in any given month relative to those receiving usual care in the public mental health system.

Essential heterogeneity was tested for using for a propensity score (probability) predicted from a probit model containing the two instruments and all other exogenous variables. A model of the marginal treatment effect on the probability of being arrested was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). None of the polynomial terms was found to be statistically significant at the 5% level. This indicates the improvement presented above represents the ATE.

A2.4 Results: Consumer Experience with Public mental health services

A2.4.1 Perception of Quality and Appropriateness

With the exception of the dependent variable (perception of quality and appropriateness) and the instruments used (odd/even entry and order of entry), this analysis is identical to that of the perception of outcomes of services in A2.3.1. The analytic sample contained 80,202 observations including 1,738 observations of consumers participating in FSP programs. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.1 reflect differences in the number of consumers evaluating each of the questions in the relevant domains and in the availability of data for the instruments used.

The joint strength of the instruments is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 11.12; critical value: 8.68). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instrument is uncorrelated with the error term (Hansen's J statistic: 1.25, $p = 0.26$). A test of the exogeneity of participation in an FSP could not be rejected (χ^2 : 69.37; $p < 0.01$) indicating instruments are necessary.

Instrumental variables Tobit estimates show that the marginal effect of participation in FSP relative to usual care in the public mental health system that is 1.11 points higher (on a scale from 1 to 5) where the overall average value is 4.03. This is an approximate 28% improvement.

Essential heterogeneity was tested for using for a propensity score (probability) predicted from a probit model containing the two instruments and all other exogenous variables. A model of the marginal treatment effect of the perception of outcomes of services was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). None of the added terms were statistically significant at the 5% level except for the squared propensity score which was positive. Thus our estimate does not represent the ATE. Consumer subgroups will vary in the outcomes received. However, our results indicate that these other subgroups will experience larger positive outcomes.

A2.4.2 Perception of Participation in Treatment Planning

With the exception of the dependent variable (perception of participation in treatment planning) and the instruments used (odd/even entry, help index, season of birth), this analysis is identical to that of the perception of outcomes of services in A2.3.1. The analytic sample contained 73,581 observations including 1,425 observations of consumers participating in FSP programs. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.1 reflect differences in the number of consumers evaluating each of the questions in the relevant domains and differences in the availability of data for each instrument used.

The joint strength of the instruments is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 20.10; critical value: 6.46). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instruments are uncorrelated with the error term. Testing the help index against odd/even entry yield a Hansen's J statistic of 0.01, $p = 0.97$. The Hansen's J statistic when all three instruments are included is 3.18, $p = 0.20$. A test of the exogeneity of participation in an FSP program is not rejected (χ^2 : 2.14; $p < 0.14$) indicating that the instruments are unnecessary. The same estimation without the instruments shows that perceptions of participation in treatment planning do not vary between those in FSP programs relative to those participating in usual care.

A2.4.3 Perception of Access

This analysis is identical to that of the perception of outcomes of services in A2.3.1, only the dependent variable used was different. The analytic sample contained 77,190 observations including 1,480 observations of consumers participating in FSP programs. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.1 reflect differences in the number of consumers evaluating each of the questions in the relevant domains and differences in the availability of data for each instrument used.

The joint strength of the instruments is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 21.71; critical value: 6.46). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instruments are uncorrelated with the error term. Testing the help index against odd/even entry yield a Hansen's J statistic of 0.49, $p = 0.48$. The Hansen's J statistic when all three instruments are included is 1.61, $p = 0.45$. A test of the exogeneity of participation in an FSP program is not rejected (χ^2 : 2.14; $p < 0.14$) indicating that the instruments are unnecessary. The same estimation without the instruments shows that perceptions of access to care do not vary between those in FSP programs relative to those participating in usual care.

A2.5 Results: Consumer Satisfaction

A2.5.1 General Satisfaction

This analysis is identical to that of the perception of outcomes of services in A2.3.1 with the difference being the dependent variable (general satisfaction) and that the instruments employed are odd/even entry and order of entry. The analytic sample contained 81,565 observations including 1,756 observations of consumers participating in FSP programs. Differences in the number of observations of consumers analyzed here as compared to the number of observations of consumers analyzed in A2.3.1 reflect differences in the number of consumers evaluating each of the questions in the relevant domains and differences in the availability of data for each instrument used.

The joint strength of the instruments is such that maximal size bias is less than 10% (Kleibergen-Paap rk Wald F -statistic: 10.45; critical value: 8.69). Note that the critical values for LIML are much smaller than for two-stage least squares (Stock and Yogo, 2005). The overidentification test fails to reject the null hypothesis that the over-identifying instrument is uncorrelated with the error term (Hansen's J statistic: 1.61, $p = 0.20$). A test of the exogeneity of participation in an FSP program is rejected (χ^2 : 48.51; $p < 0.01$) indicating that the instruments are necessary.

Instrumental variables Tobit estimates show a marginal effect of participation in FSP relative to usual care in the public mental health system that is 1.09 points higher (on a scale from 1 to 5) where the overall average value is 4.06. This is 27% higher for consumers in an FSP program relative to those in usual care.

Essential heterogeneity was tested for using for a propensity score (probability) predicted from a probit model containing the two instruments and all other exogenous variables. A model of the marginal treatment effect of the perception of general satisfaction was estimated that included all exogenous variables, the propensity score, and interactions of the propensity score with all exogenous variables. We then sequentially added the squared propensity score, the cubed propensity score, and the propensity score to the fourth power. Each equation was estimated using ordinary least squares (OLS) where standard errors were computed using bootstrapping methods (500 replications). Only the squared term was statistically significant at the 5% level. The parameter of the squared term was positive. Thus our estimate does not represent the ATE. Consumer subgroups will vary in the outcomes received. However, our results indicate that these other subgroups will experience greater satisfaction.

References

- Angrist J, Krueger A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*, 106, 979-1014.
- Angrist J, Krueger A. (1992). The Effect of Age at School Entry on Educational Attainment: An Application on Instrumental Variables with Moments from Two Samples. *Journal of the American Statistical Association*, 87, 328-336.
- Angrist J, Krueger A. (1995). Split Sample Instrumental Variables Estimates for the Return to Schooling. *Journal of Business & Economic Statistics*, 13, 225-235.
- Angrist J, Krueger A. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives*, 15, 69-85.
- Angrist JD, Pischke JS. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.
- Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. (2007). Use of Instrumental Variables in the Presence of Heterogeneity and Self-Selection: An Application to Treatments of Breast Cancer Patients, *Health Economics*, 16, 1133-1157.
- Baum, C.F., Schaffer, M.E., and Stillman, S. 2007. ivreg2: Stata Module for Extended Instrumental Variables/2SLS, GMM and AC/HAC, LIML and k-class regression. <http://ideas.repec.org/c/boc/bocode/s425401.html> [October 31, 2009].
- Buckles K, Hungerman DM. (2008). Season of Birth and Later Outcomes: Old Questions, New Answers. NBER Working Paper 14573.
- California Code of Regulations. (2010). Title 9.Rehabilitative and Development Services. Division 1. Department of Mental Health. Chapter 14. Mental Health Services Act Section. Section § 3200.300 Available from <http://ccr.oal.ca.gov/linkedslice/default.asp?SP=CCR-1000&Action=Welcome>.
- California Code of Regulations. (2010). Title 9.Rehabilitative and Development Services. Division 1. Department of Mental Health. Chapter 14. Mental Health Services Act Section. Section § 3620. Available from <http://ccr.oal.ca.gov/linkedslice/default.asp?SP=CCR-1000&Action=Welcome>.
- California Code of Regulations. (2010). Title 9.Rehabilitative and Development Services. Division 1. Department of Mental Health. Chapter 14. Mental Health Services Act Section. Section § 3620.05 Available from <http://ccr.oal.ca.gov/linkedslice/default.asp?SP=CCR-1000&Action=Welcome>.
- California Department of Mental Health. (2008a). Performance Outcomes and Quality Improvement (POQI): Consumer Perception Surveys. Available at: http://www.dmh.ca.gov/POQI/Consumer_Perception_Surveys.asp [Accessed February 3, 2010].
- California Department of Mental Health. (2008b). *Performance Outcomes Data Collection & Submission Training Manual, Revised May 2008*. Available at: http://www.dmh.ca.gov/POQI/docs/Performance_Outcomes_Manual_FINAL.pdf [Accessed August 14, 2009].
- California Department of Mental Health. (2009a). MHSIP Domains and Scoring Instructions. Performance Outcomes and Quality Improvement. Available at: http://www.dmh.ca.gov/POQI/docs/MHSIP_DomainScoringInstructions_REVISED02062008.pdf [Accessed August 14, 2009].
- California Department of Mental Health. (2009b). Personal communication with staff.
- Castrogiovanni, P, Ipachino, S, Pacchierotti, C, Pieraccini F. 1998. Season of Birth in Psychiatry: a Review. *Neuropsychobiology*, 37, 175-181.
- Chamberlain G, Imbens G. (2004). Random Effect Estimators with Many Instrumental Variables. *Econometrica*, 72, 295-306.
- Chernozhukov V, Hansen C. (2006). Instrumental Quantile Regression Inference for Structural and Treatment Effect Models. *Journal of Econometrics*, 132, 491-525.
- Chesher A. (2005). Nonparametric Identification under Discrete Variation. *Econometrica*, 73, 1525-1550.
- Cruz L, Moreira M. (2005). On the Validity of Econometric Techniques with Weak Instruments. *The Journal of Human Resources*, XL, 393-410.
- Dufour JM, Taamouti M. (2007). Further Results on Projection-Based Inference in IV Regressions with Weak, Collinear or Missing Instruments. *Journal of Econometrics*, 139, 133-153.

- Gelbach J. (2002). Public Schooling for Young Children and Maternal Labor Supply. *American Economic Review*, 92, 307-322.
- Greene WH. (2003). *Econometric Analysis*, 5th Edition, New Jersey: Prentice Hall, p. 768-769.
- Fountoulakis KN, Iacovides A, Karamouzis M, Kaprinis GS, Ierodiakonou C. (2007). Season of birth, clinical manifestations and Dexamethasone Suppression Test in unipolar major depression. *Annals of General Psychiatry*, 6, 20.
- Hayashi, F. (2000). *Econometrics*. 2000. Princeton: Princeton University Press.
- Heckman JJ, Urzua S, Vytlacil E. (2006). Understanding Instrumental Variables in Models with Essential Heterogeneity, *Review of Economics and Statistics*, 88(3), 389-432.
- Honoré B, Hu L. (2004). On the Performance of Some Robust Instrumental Variables Estimators. *Journal of Business & Economic Statistics*, 22, 30-39.
- Imbens G, Rosenbaum P. (2005). Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education. *Journal of the Royal Statistical Society*, 168, 109-126.
- Kirkpatrick B, Ram R, Amador XF, Buchanan RW, McGlashan T, Tohen M, Bromet E. (1998). Summer Birth and the Deficit Syndrome of Schizophrenia. *American Journal of Psychiatry*, 155, 1221-1226.
- Lefgren L, McIntyre F. (2006). The Relationship between Women's Education and Marriage Outcomes. *Journal of Labor Economics*, 24, 787-830.
- Leigh A, Ryan C. (2008). Estimating the Returns to Education Using Different Natural Experiment Techniques. *Economics of Education Review*, 27, 149-160.
- Levin J, Plug E. (1999). Instrument Education and the Returns to Schooling in the Netherlands. *Labour Economics*, 6, 521-534.
- National Association of State Mental Health Program Directors Research Institute. (2009). MHSIP Consumer Surveys: Scoring Protocol. Available at: http://www.nri-inc.org/projects/SDICC/TA/Ganju.Smith_2.pdf [Accessed August 17, 2009].
- Neal, D, Johnson WR. (1996). The Role of Premarket Factors in Black-White Wage Differences. *Journal of Political Economy*, 104, 869-895.
- Rihmer Z. (1980). Season of Birth and Season of Hospital Admission in Bipolar Depressed Female Patients. *Psychiatry Research*, 3, 247-251.
- Skirbekk V, Kohler HP, Prskawetz A. (2004). Birth Month, School Graduation, and the Timing of Births and Marriages. *Demography*, 41, 547-568.
- Staiger D, Stock J. (1997). Instrumental Variables with Weak Instruments. *Econometrica*, 65, 557-586.
- Stock JH, Yogo M. (2005) Testing for Weak Instruments in Linear IV Regression, in *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas J. Rothenberg* (Eds.) Andrews DWK, Stock JH, Cambridge University Press, Cambridge, pp. 80-108.
- Wooldridge JM. (2006). *Introductory Econometrics*, 3rd Edition, Mason: Thomson South-Western, p. 535.