EXPERIMENTAL ANALYSIS OF NEIGHBORHOOD EFFECTS ON YOUTH

Jeffrey R. Kling and Jeffrey B. Liebman *

May 2004

* Princeton University and NBER, Harvard University and NBER.

We thank the U.S. Department of Housing and Urban Development, the National Institute of Child Health and Development (R01-HD40404), the National Institute of Mental Health (R01-HD40444), the National Science Foundation (SBE-9876337 and BCS-0091854), the Robert Wood Johnson Foundation, the Russell Sage Foundation, the Smith Richardson Foundation, the MacArthur Foundation, the W.T. Grant Foundation, and the Spencer Foundation for funding the interim MTO evaluation and our research. Additional support was provided by grants to Princeton University from the Robert Wood Johnson Foundation and from NICHD (5P30-HD32030 for the Office of Population Research), and by the Princeton Industrial Relations Section, the Bendheim-Thomas Center for Research on Child Wellbeing, the Princeton Center for Health and Wellbeing, and the National Bureau of Economic Research.

We are grateful to Todd Richardson and Mark Shroder of HUD, to Judie Feins, Barbara Goodson, Robin Jacob, Stephen Kennedy, and Larry Orr of Abt Associates, to our collaborators Jeanne Brooks-Gunn, Alessandra Del Conte Dickovick, Greg Duncan, Lawrence Katz, Tama Leventhal, Jens Ludwig, and Lisa Sanbonmatsu, to our research assistants Erin Metcalf and Ken Fortson, and to numerous colleagues for their suggestions. We thank Christopher Jencks for detailed comments on an earlier draft.

EXPERIMENTAL ANALYSIS OF NEIGHBORHOOD EFFECTS ON YOUTH

Jeffrey R. Kling and Jeffrey B. Liebman

May 2004

ABSTRACT

We examine the effects of moving out of high-poverty neighborhoods on the outcomes of teenage youth, a population often seen as most at risk from the adverse effects of such neighborhoods. The randomized design of the Moving To Opportunity demonstration allows us to compare groups of youth, initially similar and living in high-poverty public housing. An "experimental" group was offered vouchers valid only in a low-poverty neighborhood; a "Section 8" group was offered traditional vouchers without geographic restriction; and a control group was not offered vouchers.

We study outcomes in four domains: education, risky behavior, mental health, and physical health. Females in the experimental group experienced improvements in education and mental health and were less likely to engage in risky behaviors. Females in the traditional voucher group experienced improvements in mental health. Males in both treatment groups were more likely than controls to engage in risky behaviors and to experience physical health problems. We adopt a multiple-testing framework to account for the large number of estimates considered. We show that the overall effects on females in the experimental group and the effects on mental health for females in both treatment groups were least likely to be due to sampling variation.

Families with female children and families with male children moved to similar neighborhoods, suggesting that their outcomes differ not because of exposure to different types of neighborhoods but because male and female youth respond to their environments in different ways.

Keywords: neighborhood effects; social experiment; multiple testing

JEL classifications: H43, I18, J18

Jeffrey R. Kling
Department of Economics
and Woodrow Wilson School
Princeton University
Princeton, NJ 08544
and NBER
kling@princeton.edu

Jeffrey B. Liebman
Kennedy School of Government
Harvard University
Cambridge, MA 02138
and NBER
jeffrey_liebman@harvard.edu

I. Introduction

Youth who grow up in disadvantaged neighborhoods fare substantially worse than those who grow up with more affluent neighbors on a wide variety of health and socioeconomic outcomes. A fundamental question in the design of appropriate education, health, and social policies for low income families and communities is the extent to which these correlations reflect the causal impacts of neighborhoods as opposed to family and individual attributes that are not directly affected by the residential environment. This paper uses data from a randomized housing mobility experiment to estimate the causal effects on teenage youth of moving out of high-poverty neighborhoods.

There are a variety of theoretical views about the potential effects of neighborhoods on youth.[1] One school of thought argues that disadvantaged neighbors and neighborhoods have adverse causal effects on adolescent development through exposure to violence and poor peer influences, absence of appropriate adult role models, and lack of school, community, and health care resources. Indeed, teenage youth are often seen as the age group most susceptible to the adverse influences of disadvantaged neighborhoods (Ellen and Turner, 1997). An alternative view is that neighborhoods have only limited effects on youth outcomes since a wide variety of peers and role models are available in all neighborhoods, and even those in the poorest areas can find peers who stay out of trouble. Under this view, family influences and individual human capital investment are often seen as determinative of child outcomes, and some scholars have theorized that most behavioral patterns result from a combination of genetic factors and early childhood development which have already been formed by the time children enter school (Bouchard, 1997; Rowe, 1994; Scarr, 1992). Still another view, based on "relative deprivation" models, is that children in low-income families fare better in low-income neighborhoods than in high-income neighborhoods. Living with high socioeconomic status (SES) neighbors could provoke resentment among poor children, and the children may face discrimination from being a racial or ethnic minority. Moreover, competition with high-SES peers in school could lead to low grades, low class rank, and low self-esteem, potentially translating into social isolation, depression, anxiety, and/or delinquency (Wood, 1989; Marsh and Parker, 1984; Collins, 1996). In these models, low-income children have limited ability to tap into the resources of the high-

income community and may in fact be able to obtain more resources in a low-income neighborhood where they are able to build stronger social ties. Finally, the particular children in high-income neighborhoods with whom poor children associate could be those with the lowest incomes and could be engaging in more risky and delinquent behaviors than the peers that the poor children would have spent time with if they had remained in a low-income neighborhood.

Each of these views is potentially plausible. But the evidence to date distinguishing among these views is essentially indeterminate. Differences in youth outcomes by type of neighborhood reflect, not only the effects of living in a given neighborhood, but also the systematic sorting of families across neighborhoods. Because the sorting process involves many unobserved characteristics, estimating the true impact of neighborhoods is extremely difficult in the absence of a genuinely exogenous source of variation in residential location.

This paper uses data from a randomized housing mobility experiment to minimize such difficulties. The youth studied in this paper were all initially living in high-poverty public housing projects, and all came from families that applied for a randomized housing voucher demonstration known as Moving to Opportunity (MTO), administered by the U.S. Department of Housing and Urban Development (HUD). From 1994-97, 4600 households enrolled in MTO in Baltimore, Boston, Chicago, Los Angeles, and New York. Applicants were assigned by lottery to one of three groups:

- *Experimental group.* Received a Section 8 certificate or voucher[2] plus assignment of a counselor for housing search assistance. The voucher was initially valid only in Census tracts with a 1990 poverty rate below 10 percent. The geographic restriction lasted for one year, after which the families were free to use the voucher to rent in any location.
- *Section 8 group.* Received a regular tenant-based Section 8 certificate or voucher with no location restriction and no special counseling.
- *Control group.* Received no tenant-based assistance but remained eligible for current project-based housing.

---

[1] Mayer and Jencks (1989) and Jencks and Mayer (1990) provide the classic delineation of possible mechanisms. Brock and Durlauf (2001), Duncan and Raudenbush (2001), Ellen and Turner (1997), Leventhal and Brooks-Gunn (2000), and Sampson, Morenoff, and Gannon-Rowley (2002) summarize the more recent literature.

[2] Section 8 tenants typically pay about 30 percent of their income in rent. The assistance payment makes up the difference between the tenant contribution and an area-wide threshold established by HUD and set at the fortieth percentile of area rents. Residents of public housing also pay about 30 percent of their income in rent.

Data collected during 2002 allow us to assess the experimental impacts of residential location on the educational achievement, risky behavior, mental health, and physical health of youth. At the time of random assignment (conducted throughout 1994-97) these youth were ages 8-16. In 2002 they were ages 15-20 and many had been living in substantially different neighborhoods for 4 to 7 years.

We present results of the experiment on specific outcomes (such as test scores) to give a concrete sense of the measures and their magnitudes. We also present summary measures of mean effect sizes within domains (such as educational outcomes for females) in order to aggregate information. Since we have multiple summary measures, we follow an approach to inference that controls the probability of at least one rejection of a true null hypothesis (the familywise error rate). Systematic attention to the multiple testing problem is rare in economics, and we view the approach set out in this paper as one that could be widely applied.

In addition to providing an assessment of the overall causal impacts of neighborhoods on youth development in the context of a specific housing policy evaluation, we present evidence on the mechanisms that produced the results. We examine experimental impacts on a wide range of "mediating" variables, such as exposure to neighborhood violence, presence of adult role models, and access to medical care.

This paper offers several major enhancements relative to previous MTO research, which consisted of small pilot studies at individual MTO sites and studied outcomes shortly after random assignment.[3] In particular, this paper features standardization of data collection across the five MTO cities, significantly larger sample sizes, and extended exposure time to the new neighborhoods. Moreover, we analyze questions asked directly to youth themselves, including previously unexplored topics such as school drop out rates, substance use, and adolescent mental health. In other recent papers, we examine adult health and economic self-sufficiency (Kling, Liebman, Katz, and Sanbonmatsu 2004), youth criminal behavior (Kling, Ludwig, and Katz, 2004), and nonlinearities of neighborhood effects (Liebman, Katz, and Kling 2003). Some of this work appears in less technical form in a report from HUD (Orr et al 2003).

This paper is organized into ten sections. Section II describes our conceptual framework and analytical methods. Section III discusses the data and provides descriptive statistics.

[3] The earlier research is summarized by Goering and Feins (2003). In section IX, we compare our results to those from the earlier studies.

Sections IV-VII provide results for education, risky behavior, mental health, and physical health, respectively. Section VIII summarizes these primary results and applies our methodology for inference in the presence of multiple outcomes. Section IX presents evidence on potential mechanisms producing the results. Section X concludes.

II. Conceptual Framework and Analytical Methods

Neighborhood effects can arise from social interactions or from other features of the neighborhood environment. Several recent survey articles on the economic analysis of social interactions have emphasized that progress on theoretical models and conditions for econometric identification have outpaced empirical work. Brock and Durlauf (2001) note that "a decisive empirical demonstration has yet to be made" and that "the long-run success of the interactions-based approach in economics depends on a clear demonstration of its empirical salience over a range of contexts." Moffitt (2001) concludes that "theory has run considerably ahead of empirical testing, the development of policy interventions that work through social interactions, and the evaluation of such interventions." Manski (2000) reports that "the empirical literature has not shown much progress" in studying social interactions and sees "a compelling need to enrich the data that researchers bring to bear." While there have been some important empirical findings subsequent to these surveys,[4] there remains significant uncertainty about the contexts in which social interactions are likely to be economically important.

This uncertainty is perhaps greatest in the neighborhood effects literature. Numerous non-experimental studies document strong associations between neighborhood characteristics and individual outcomes. However, these associations appear to be much weaker or nonexistent in the studies with the most credible identification strategies. Reviewing the early non-experimental literature, Mayer and Jencks (1989) conclude "the more we learn about a given outcome, the smaller the effects of mean SES usually look." More recently, Ellen and Turner (1997) report that "some recent studies that have done the most careful job of controlling for unobserved family characteristics … find no independent neighborhood effects, casting doubt on the robustness of results from other studies." Finally, recent quasi-experimental studies (Jacob,

[4] Notable recent contributions include: Glaeser, Sacerdote, and Scheinkman (2003) who identify social interactions based on the ratio between coefficients estimated using aggregate and individual level data; Sacerdote (2001), Hoxby (2001), Angrist and Lang (2002), Kremer and Levy (2003), Zimmerman (2003), Duncan et al (2003) and

2004; Oreopoulos, 2003) find little or no effect of living in high-poverty housing projects on child outcomes. Thus, a threshold issue for the current paper is whether a large randomized experiment that moves people out of some of the most distressed U.S. neighborhoods demonstrates the existence of substantively important effects of neighborhoods on individuals.

The MTO experiment provides a convincing mechanism for eliminating selection bias and therefore for answering the threshold question of whether people's residential and social environments affect their outcomes. However, the experiment does not provide a direct method for distinguishing between different types of social interactions or between social interactions and other features of the neighborhood environment (Moffitt, 2001). The design of the experiment deliberately avoided clustering movers in their new neighborhoods. In most cases, therefore, the experiment will not have had large effects on the receiving neighborhoods, and will not produce detectable feedback effects via endogenous social interactions.[5] In theory, the experiment could nonetheless supply useful information about endogenous interactions if we could observe, for example, the impact of peer educational attainment on the educational attainment of MTO youth. In practice, however, MTO changed a large bundle of neighborhood characteristics for families that moved, including not only peer outcomes for the variable under study but also the exogenous characteristics of peers and the institutional environment. Therefore, it will not in general be possible to disentangle the separate influences of each. In some cases, however, the pattern of experimental impacts on mediating variables may suggest which mechanisms are most consistent with the data, as discussed in section IX.

Intent-to-treat estimation. Our basic empirical approach will be to separately compare each treatment group to the control group on a wide range of measures. We use the term "treatment" groups to refer collectively to the experimental and Section 8 groups. We illustrate our approach to estimation in a simple regression framework. Let D be an indicator variable for use of an MTO housing voucher to move ("treatment compliance"). Let Z, with coefficient π_1,

---

Gould et al (2004) who study peer effects in educational settings; and Duflo and Saez (2003) who study how information provision alters the behavior of coworkers who were not given the information.

[5] See Manski (1993) on the distinctions between endogenous and exogenous social interactions. There are three circumstances in which MTO could have detectable effects on receiving neighborhoods. First, even if having a single MTO family in a Census tract has very little impact on the Census tract as a whole, it might impact the families who live very close to the MTO family. However, the MTO research design did not collect information on neighbors and even if it had it would be difficult to analyze such data since there is not an obvious counterfactual for what the neighbors' outcomes would have been. Second, if MTO youth had a high propensity to commit crimes, introducing even a single family into each neighborhood could have significant effects. Third, it is possible that a single disruptive child could impact an entire classroom and produce detectable effects.

be an indicator variable for being eligible for an MTO housing voucher subsidy ("treatment group assignment"). Let ε_1 be the other determinants of subsidy use, which is modeled in equation (1).[6]

$$\textbf{(1)} \qquad D = \beta_1 + Z\pi_1 + \varepsilon_1$$

The "intent-to-treat effect" (ITT) is captured by the ordinary least squares (OLS) estimate of the coefficient π_2 in a regression of the outcome, Y, on an indicator for assignment, Z, to a treatment or control group as in equation (2).[7]

$$\textbf{(2)} \qquad Y = \beta_2 + Z\pi_2 + \varepsilon_2$$

Treatment-on-treated estimation. A complementary parameter of interest is the effect on individuals who use an MTO housing voucher to move; we refer to these individuals as compliers, using terminology of Angrist, Imbens, and Rubin (1996).[8] Under several assumptions (treatment group assignment is random; control group members are prohibited from receiving program subsidy assistance; the effect on outcomes of treatment assignment works entirely through making a subsidized move through the program), we can use treatment assignment as an instrumental variable to estimate the parameter commonly known as "the effect of the treatment on the treated" (TOT).[9] One estimate of TOT is π_2/π_1, or ITT divided by the proportion

[6] Note that program subsides are not offered to control group members, so D equals zero when Z equals zero. Equation (1) simply establishes notation for the first stage of the two stage least squares estimation of equation (3), where the probability of subsidy use conditional on offer is π_1.

[7] The intent-to-treat effect measures the overall expected value of the intervention on the entire treatment group, including both those who leased an apartment through MTO and those who did not. ITT is often the parameter most directly relevant for forecasting the effects of public policies, because the costs and benefits of a policy depend both on the policy's compliance rate and on its effect on those who comply.

[8] This parameter can be useful for extrapolating to circumstances in which the compliance rate differs (for example if housing market conditions make it much easier or more difficult for families to lease-up), but outcomes conditional on compliance are expected to be similar to those experienced by MTO compliers. Our analysis focuses on families who were caused to move into neighborhoods that tend to have few families receiving public rental assistance. There may well be additional "general equilibrium" effects not captured in our analysis if the subsidized movers were to become a large enough presence to affect aggregate neighborhood characteristics (Manski 1993; Garfinkel, Manski, and Michalopoulos 1992; Heckman 2001). See Miguel and Kremer (2004) for a recent example of empirical estimation of treatment effect externalities.

[9] The assumptions required for TOT say that those in the treatment groups who did not accept the treatment offer had no average treatment effect in comparison to those in the control group who would not have accepted the treatment if it had been offered to them. For the experimental group, this assumption implies that the later outcomes of households who met with a housing mobility counselor were not affected by the counselor if that household did not make a subsidized move through the MTO program. For both treatment groups, this assumption implies that the experience of housing search induced by assignment to a treatment group did not affect later outcomes if that

receiving the treatment.[10] This is numerically identical to a two stage least squares (2SLS) regression of Y on D with Z used as an instrumental variable for D, as in equation (3).

(3) $$Y = \beta_3 + D\gamma_3 + \varepsilon_3$$

Regression models. There are two treatment groups in this application, and we use separate regressions for experimental-control and Section 8-control estimates. We fully interact our models by gender; joint estimation of the female and male results allows us to directly implement Huber-White standard errors with family-level clustering to account for sibling correlation in outcomes. Let G be an indicator for female gender, and let the subscript *g* refer to being female (g=1) or male (g=0). The subscript *k* refers to the *k*th outcome. For the estimates reported in this paper, we include adjustment for baseline covariates X.[11] The regression-adjusted intent-to-treat estimates for the *k*th outcome, π_{gk}, are calculated using equation (4), where the point estimates are numerically identical to separate models by gender.

(4) $$Y_k = (1-G)(X\beta_{0k} + Z\pi_{0k}) + G(X\beta_{1k} + Z\pi_{1k}) + \upsilon_k = W\theta_k + \upsilon_k$$

For TOT, we use the parameter γ_{gk} from a 2SLS regression with treatment indicators used as instrumental variables, as shown in equation (5).

(5) $$Y_k = (1-G)(X\lambda_{0k} + D\gamma_{0k}) + G(X\lambda_{1k} + D\gamma_{1k}) + \zeta_k$$

---

household did not make a subsidized program move. We doubt that this assumption is strictly true, but for those who did not make a subsidized move through the program we also believe that the effects of mobility counselors (who mainly provided housing advice and not general social services) and of housing search on youth outcomes are likely to be orders of magnitude smaller than the effects of moving to a new residential location. In this sense, we interpret TOT as a useful approximation. The assumptions for TOT also require that control group members were not affected by the experience of "losing the lottery." We similarly view this as not literally true but as a reasonable approximation. In qualitative research with MTO families, we found they have been "lottery losers" numerous times in life and that some control group adults do not even recall enrolling in the demonstration. Moreover, while the parents in the control group enrolled and then were notified that they would not receive a program voucher, many control group youth were not involved at all in this process and were completely unaware of their families' participation.

[10] Inflation by the proportion in the treatment group who actually received the treatment was introduced in the program evaluation context by Bloom (1984); see Heckman, LaLonde, and Smith (1999) for a comprehensive discussion of alternative parameters of interest in the evaluation of social programs.

[11] Covariates such as age and youth behavior at the time of random assignment improve the precision of the estimates by reducing residual variation in the regression. Including these baseline covariates does not change the treatment effect coefficients unless the covariates happen to differ between groups due to small-sample variability.

The 2SLS estimate of γ_{gk} in (5) uses the information in X to obtain additional statistical precision, and is asymptotically equivalent to the unadjusted indirect least squares estimate of TOT ($\gamma_3 = \pi_2/\pi_1$) in (3).

Control complier mean. The mean of the outcome for those accepting the treatment, which we refer to as the treatment complier mean (TCM), is directly identified by the data. Katz, Kling, and Liebman (2001) show how to identify the implied mean outcome for those in the control group who would have accepted the treatment if it had been offered to them -- the control complier mean (CCM) -- under the assumptions needed to estimate the TOT effect. The regression-adjusted TOT estimate is used, implicitly evaluating the CCM at the mean level of X for treatment compliers. CCM is defined in equation (6) for each gender *g* and outcome *k*.

$$\textbf{(6)} \qquad CCM_{gk} = E[Y_k \mid D = 1, G = g] - \gamma_{gk}$$

We use the CCM to provide a base rate against which the relative magnitude of the treatment-on-treated effect can be assessed.[12]

Summary measures. In order to make summary statements about each outcome domain (education, risky behavior, mental health, physical health), we construct summary measures. The building blocks of the summary measures are standardized treatment effect sizes. Equations (7) - (9) define the mean effect size, τ_g, for a set of K outcomes for a single gender, based on the treatment effect estimates and the control group standard deviations.[13]

$$\textbf{(7)} \qquad \sigma^2_{gk} = Var(Y_k \mid G = g, Z = 0)$$

[12] Simple measures of relative change such as TCM/CCM are sensitive to whether one studies an outcome or its absence (i.e. one minus the outcome). The percentage change in the odds ratio ((((TCM/(1-TCM))/(CCM/(1-CCM)))-1)*100) is a measure of relative change that is independent of the way in which the outcome is defined. For binary outcomes, sampling variation can produce negative estimates of the CCM. Our analytic method uses data on the fraction of youth in treatment group families that did not comply (treatment never-takers) and the outcome prevalence for these youth, and assumes that the same fraction in the control group had exactly the same outcome prevalence. In any one sample this method may produce a negative estimate of a CCM if a particular realization of the treatment never-taker mean is higher than the realization of the control never-taker mean for that particular sample even though the method is unbiased in repeated sampling. In our results, we report a CCM of zero when the CCM estimate for a binary outcome is negative.

[13] Normalizing by the standard deviation of the control group gives the effect size relative to the counterfactual of no treatment, a metric known as Glass's delta (Glass, McGaw, and Smith 1981). This measure has the advantage that effects for our two treatment groups relative to the control group are expressed in the same metric. Other choices could have been made; see Rosenthal (2000) for a discussion. Use of one common alternative, the pooled variance of the treatment and control groups, changes our results only trivially.

(8) $$\tau_{gk} = \frac{\pi_{gk}}{\sigma_{gk}}$$

(9) $$\tau_{g} = \frac{1}{K}\sum_{k=1}^{K}\tau_{gk}$$

In order to calculate the sample variance of τ_g in equation (9), we need to account for covariance in the estimates τ_{gk}. We obtain this covariance matrix using the seemingly unrelated regression system shown in equation (10). Point estimates for each outcome are identical to those obtained using equation (4). Let I_K be a K by K identity matrix and let W be defined as in (4).

(10) $$Y = (I_K \otimes W)\theta + \upsilon \qquad Y = (Y_1^{'},\ldots,Y_K^{'})'$$

The estimated mean effect size is simply the average of the estimated coefficients π_{gk}, elements of θ in (10) normalized by σ_{gk} from (7). We calculate a point estimate, standard error, and p-value for τ_g based on the π_{gk} jointly estimated in (10). Note that this method treats σ_{gk} as known under the assumption that its sampling variance does not affect the results. To examine the sensitivity of our results to this assumption, we also use delta-method calculations for the standard error of τ_g based on a seemingly unrelated regression system that has been expanded to incorporate the covariance between π_{gk} and σ_{gk}. Let ν_k be the deviation from the gender-specific mean of Y_k for members of the control group.

(11) $$\begin{pmatrix} Y \\ \nu^2 \end{pmatrix} = \begin{pmatrix} I_K \otimes W & 0 \\ 0 & I_K \otimes \begin{pmatrix} (1-G) & G \end{pmatrix} \end{pmatrix}\delta + \zeta \qquad \nu = (\nu_1^{'},\ldots,\nu_K^{'})'$$

This model allows us to estimate π_{gk} and σ_{gk} (as elements of the coefficient vector δ) along with their covariances -- which are used in the delta-method standard error calculations.[14]

We use the mean effect size as our summary measure for several reasons. Foremost is that we wish to detect a global pattern of effects that are generally beneficial or generally adverse for a group. The test of the mean effect size has its roots in the biostatistics tradition of global

[14] Note that when all individuals have data on all outcomes and there is no regression adjustment of the treatment effects, the computation of mean effect size would be identical to constructing an index at the individual level, where each outcome is divided by its standard deviation and the individual's index value is the mean of the standardized outcomes. When an individual is missing data on an outcome however, the other non-missing outcomes implicitly are given more weight when the index is based on a simple average of non-missing standardized

significance testing for multiple endpoints in clinical trials (O'Brien 1984; Logan and Tamhane 2003) and in a parallel but separate literature on meta-analysis (Hedges and Olkin 1985). Some common alternative procedures, such as calculating an F-test of the joint significance of multiple outcomes, are nondirectional and have less power to detect an alternative hypothesis that all effects go in the same direction (Tamhane and Logan 2003).[15] When there is no a priori reason to assign different weights to different outcomes in the decision problem, using the mean effect size provides a simple way of aggregating disparate outcomes on a common metric. It is also directly scalable, so that if we assume the effect size is zero for some fraction of the population (such as the noncompliers), then the effect size for the remainder is a simple scalar multiple (such as the reciprocal of the compliance rate times the original estimate).

Multiple testing. We report results on many outcomes in this paper. If we were to perform separate hypothesis testing for each outcome, the probability that we would wrongly reject a true null hypothesis for at least one outcome would be much greater than the significance level used for each test. When examining the statistical significance of treatment effects, we therefore discuss both the significance of coefficients viewed in isolation (known as per-comparison significance) and as a member of a family of hypotheses (known as familywise significance). An example of a family would be the set of null hypotheses of no treatment effects for each of K outcomes.

To fix the familywise error rate (FWE), we can set critical values for test statistics such that the probability is less than .05 that at least one of the tests in a family would exceed the critical value under the joint null hypothesis of no effects. More generally, we can calculate an adjusted p-value that is the smallest familywise significance level at which one rejects the null hypothesis that there is no effect for all members of the family given the observed effects.

The simplest form of these adjustments are Bonferroni adjustments, in which the per-comparison p-values are multiplied by the number of estimates in the family.[16] The testing

---

outcomes. The formulation described above based on the mean of estimated effects is a more direct summary of the estimates for each outcome.

[15] Another possibility would have been to have used a multivariate test for one-sided alternatives (e.g. Follman 1996), but we made an ex-ante decision to retain the ability to detect effects that were consistently beneficial or consistently adverse in our multiple testing analysis just as we decided to use two-sided t-tests for specific outcomes.

[16] The Bonferroni bound result is the following. Let $P(H_i)$ be the probability that the null hypothesis H_i is falsely rejected -- that is, the significance level of the test of H_i. Let P(Union of all H_i) be the familywise Type I error, or the probability that at least one H_i is falsely rejected. P(Union of all H_i) <= Sum of all $P(H_i)$. See Savin (1984) for a detailed review. The intuition behind this statement is that the worst-case scenario is when there is no overlap between the false-rejection regions and therefore the total area in which we falsely reject at least one hypothesis

procedure can be made more powerful using sequential testing, where hypotheses are removed from the family of nulls after they have been rejected. The Bonferroni-Holm adjustment is a sequential rejection procedure that multiplies a per-comparison p-value, p, by the number of raw p-values in the family greater than or equal to p (Holm 1979). Formally, let p_j be the observed per-comparison p-values for J estimates in a family, ordered such that the subscript j indicates descending significance (i.e. p_1 is the smallest p-value). Let P_j be a random variable for the p-value of the per-comparison significance test under the null hypothesis H_j. The definition of the Bonferroni-Holm adjusted p-value p^a_j for H_j is given in equation (12).[17]

(12) $p^a_j = \Pr(\min(P_j, ..., P_J) < p_j \mid H_j, ..., H_J \text{ are true})$

Since Bonferroni adjustments (including Bonferroni-Holm adjustments) become increasingly conservative (have lower power) as the correlation in the test statistics increases, we also calculate bootstrap estimates of adjusted p-values incorporating correlation in test statistics based on the free step-down resampling algorithm of Westfall and Young (1993). Our implementation of the Westfall-Young approach is described in Appendix A.

Mediating factors. We aim not only to provide evidence on the causal effects of neighborhoods, but also to shed some light on the mechanisms that produce these effects. The MTO treatments may affect youth outcomes through changes in both neighborhood and family environments, which we collectively refer to as mediators. By providing evidence on the experimental impacts on mediators, we can narrow the set of mechanisms that are consistent with our data. We test hypotheses regarding how neighborhood and family processes affect youth in two steps. First, we estimate experimental treatment effects for the mediating factors

---

equals the sum of all of the individual areas. In other words, the equality holds when the intersection of all H_i is the empty set. This implies that a bound on the familywise error is the significance level times the number of estimates. For example, assume two t-statistics are independent, with absolute values of 1 and 1.96. Using a significance level of .05, the Bonferroni bound is .05*2 = .10. The exact probability of at least one of the two being falsely rejected is 1-.95^2 = .0975. The probability that both null hypotheses are falsely rejected is .0025; since this intersection is not empty, the exact probability is slightly smaller than the bound. The Bonferroni adjusted p-value associated with the t-statistic of 1.96 is .10.

[17] This procedure is clearly more powerful than the standard Bonferroni procedure since the p-value adjustment involves multiplying the per-comparison p-value by a number that is less than or equal to the multiplier used in the standard Bonferroni procedure. To see that the Bonferroni-Holm procedure controls the familywise error rate, consider first the test of H_1 (the hypothesis with the lowest per-comparison p-value). If H_1 is true, we either fail to reject H_1 in which case we also fail to reject the rest of the H_j or we accept H_1. For this first test, the Bonferroni-Holm adjustment is the same as the standard Bonferroni adjustment which we know controls the familywise error rate. If H_1 is false and we reject it, we move on to consider H_2. There are now J-1 remaining hypotheses, and for

(separately by gender) through which we theorize that neighborhoods affect youth outcomes.[18] Then we compare the patterns of experimental effects for the mediators and the associated outcomes with our theories for how mediators should affect outcomes. To the extent that there is no average treatment effect on a mediating factor, we interpret this as evidence against the importance of that factor -- in that a theory involving this factor would need to be complex, with positive effects for one group offsetting negative effects for some other group.[19]

III. Data

In 2002 we collaborated with HUD and Abt Associates to collect survey data from one adult and up to two randomly selected children in each MTO household. The surveys, described more fully in Orr *et al*. (2003), covered a range of topics including housing, neighborhoods, health, child behavior, education, social interactions, employment, and public assistance receipt. The survey data encompass the basic areas of youth well-being: cognitive/academic, emotional/mental, social, and physical development (Brooks-Gunn, 1990; Hauser et al., 1997; McCormick and Brooks-Gunn, 1989; Meisels and Shonkoff, 1990).

The data used in this paper focus on about 1800 youth whose ages were 15-20 as of December 31, 2001. Older youth were not interviewed in this study, since they were expected to be more difficult to locate and to have lived in new neighborhoods (if they moved through the program) for less time before starting their own households. We selected age 15 as the lower

---

this remaining set the standard Bonferroni bound result ensures that multiplying the per-comparison p-values by J-1 is sufficient to control the familywise error rate. The argument continues until we fail to reject a hypothesis.

[18] Many mediators are household-level variables, and households may include both female and male youth. The identification for effects that differ by gender comes from the fact that many households have only one youth present and others have two of the same gender, and the outcome for that household can then be associated with a specific gender. Specifically, when Y in equation (4) is a household-level variable, we use household averages of the right hand side variables (and the household sum of sample weights) for individual youth. Estimating the effects in an aggregated model allows the treatment effect for females to be estimated conditional on having the effect for males held constant, making use of the information from households where there are two children of opposite gender.

[19] We fully recognize that our examination of the logical pattern of results on mediators and outcomes does not enable us to definitively accept theories about particular mechanisms or to distinguish among multiple mechanisms, especially when multiple mechanisms and the outcome change in the predicted direction for all treatment subgroups. This is an inherent limitation of a research design that does not experimentally manipulate particular mechanisms but instead assigns a treatment that may simultaneously affect bundles of mechanisms within different subgroups. Nevertheless, by providing causal estimates of the impacts of the experimental treatment on mediators, the MTO experiment provides a much more definitive test for the presence or absence of mediators than is possible with non-experimental data when the variation in mediators is endogenously determined.

limit for this study since many of our outcomes (e.g. dropping out of school, substance use) have quite low prevalence below this age.

Response rates. Data are used from an interview with the youth and an interview with an adult from the youth's household (usually the mother). The data were collected in two phases. In our main phase, we attempted to collect data from 10,931 children ages 5-20 and from adults from the 4248 households randomly assigned to MTO as of December 31, 1997. This data collection effort extended from December 2001 to July 2002, and we completed data collection with 78 percent of the sample. We refer to this as our Initial Response Rate (IRR). Among all individuals without completed surveys at that time, we drew a 3-in-10 subsample. The purpose of the subsampling was to concentrate our remaining resources on finding hard-to-locate families in a way that would minimize the potential for non-response bias in our analyses. Between July 2002 and September 2002, we completed surveys with 49 percent of the subsample. We refer to this as our Subsample Response Rate (SRR). Since individuals assigned to the subsample are representative of all nonrespondents from the initial phase, we combine them to report an overall Effective Response Rate; ERR = IRR+(1-IRR)*SRR. For the study overall, the ERR is therefore about 89%. The overall ERR for this sample of youth is about 88%, with 90% for females and 86% for males.[20] Interviewers were not informed about the random assignment group of the respondent, though in some cases they may have been able to infer it. Nearly all interviews were conducted in person.[21]

Descriptive statistics. The families of the youth in our study enrolled in the demonstration from 1994-1997, when the youth were 8-16 years old. At the time of enrollment, the head of household completed a baseline survey that included information about the family as well as a limited amount of information about each child. The families in Baltimore and Chicago were almost entirely non-Hispanic African-American; there was a nearly equal mixture of African-Americans and Hispanics in Boston, Los Angeles, and New York.

[20] The differences in response rates between randomly assigned groups is relatively small -- 87% for the experimental and control groups, and 90% for the Section 8 group. For our sample of youth ages 15-20, the number of respondents at each site is small so we focus our analysis on the pooled results for all five sites. Effective response rates did vary by site, ranging from 83% in Boston to 95% in Chicago. We believe that two factors were primarily responsible for variation in response rates by site: differences in the extent of Hope VI demolition and other displacement of people from public housing in each city during this period and differences in abilities of interviewers hired in the different cities to locate sample members.

[21] Thirty-eight interviews of people who had moved to remote areas were conducted by telephone.

Means of the baseline characteristics of the sample with whom we completed a youth survey are shown separately by gender and random assignment group in Table 1. Missing values for baseline covariates were imputed at the mean, based on gender, site, and age at random assignment. There are some differences in the distribution of baseline characteristics between groups, with three of the treatment group means significantly different from the control mean across 32 comparisons, providing a rationale beyond residual variance reduction for regression adjustment in this application.[22] The X variables for equations (4) and (5) use the covariates shown in Table 1, as well as additional ones discussed in the table notes. In this table and throughout the paper, all statistical estimates are computed using sample weights, described in the notes to Table 1.

The addresses at which MTO youth resided were geocoded and matched to Census data. At the time of random assignment, 94 percent of youth resided in tracts with poverty rates of 36 percent or higher (according to the 1990 Census). The distribution of poverty rates (as measured by the 2000 Census) for the 2002 locations of the youth are shown in Table 2. A remarkable feature of Table 2 is that the middle three rows -- describing the poverty rates of the control group and of the subsets of the experimental and Section 8 groups that did not make a subsidized move through MTO -- are very similar, implying that decisions to move without a voucher were similar for those in the control group who would and would not have moved if offered a voucher.

The fraction making subsidized moves (the treatment compliance rate) was 43 percent in the experimental group and 55 percent in the Section 8 group. 70 percent of the experimental compliers and 41 percent of Section 8 compliers were living in a tract with a poverty rate of less than 24 percent, as opposed to 21 percent of those in the control group. In these data, 4-7 years after random assignment, very few treatment compliers were in tracts with poverty rates greater than 48 percent, whereas 32 percent of control group members were in these very high poverty

[22] A summary indicator of the joint significance of differences in the X variables between groups is the p-value on an omnibus F-test of the null hypothesis that all the differences between treatment and control group means in the X variables are zero. This is computed from the seemingly unrelated regression system in equation (10), but where Y is a stacked vector of the 23 baseline covariates and W contains only a constant and a treatment group indicator. Using the entire sample of youth whom we attempted to interview, the experimental-control contrast has a p-value of .19 for females and .41 for males. For the Section 8-control contrast, this p-value is .74 for females and .31 for males. We interpret these results as consistent with successful random assignment of treatment status. Using the sample of people with whom we successfully completed interviews in the initial phase and those whom we attempted to interview in the subsampling phase we obtain p-values of .11, .08, .49, and .01, respectively. Thus the random 3-in-10 draws which allocated people into the subsample introduced some differences in the distribution of

tracts. Finally, the mobility outcomes are quite different for compliers in the two treatment groups. In particular, 31 percent of those in the experimental group (initially restricted to low poverty areas) were living in neighborhoods with poverty rates of 0 to 12 percent compared with only 8 percent of Section 8 group compliers.[23] Thus, even though a significant period of time had passed and some experimental group families had made second or third moves back to higher poverty neighborhoods, the experiment still did manage to produce large differences in neighborhood characteristics across the three experimental groups.[24]

IV. Education Outcomes

Background. A large literature links neighborhood characteristics with educational outcomes.[25] Associations have been documented between affluent neighbors and adolescent school completion, educational attainment, and self-reported grades (Aaronson, 1998; Brooks-Gunn et al., 1993; Connell and Halpern-Felsher, 1997; Dornbusch, Ritter, and Steinberg, 1991; Duncan *et al.*, 1994).[26] Crane (1991) found higher dropout rates in neighborhoods that had a very low percentage of professional workers and interpreted these findings as evidence of the importance of adult role models. Some of the most striking results on the effects of neighborhoods come from a study of the Gautreaux program in Chicago, comparing families who were placed in central city locations versus those placed in suburban locations. Rosenbaum (1995) found that children in suburban locations had higher satisfaction with teachers and better attitudes about schools, and that high school dropout rates were much lower for suburban

---

X variables across groups. Finally, the p-values for the sample with whom we actually completed interviews were .02, .14, .70, and .09, respectively.

[23] The main reason why only 31 percent of experimental group compliers were living in neighborhoods with poverty rates below 12 percent is that the neighborhoods to which they initially moved experienced rising poverty. Thus, measured by the 1990 Census, 95 percent of initial moves were to low-poverty neighborhoods (<12 percent). Measured by the 2000 Census, 51 percent of the initial moves were to low-poverty neighborhoods.

[24] 66 percent of control group members and about 64 percent of treatment non-compliers had moved between random assignment and 2002. Part of the reason for these high mobility rates is that federal and local efforts to improve distressed public housing such as the HOPE VI program demolished many units that MTO families were living in (and often provided Section 8 vouchers to displaced residents). In 2002, 20 percent of control group households reported they were using a Section 8 vouchers (obtained through a non-MTO channel). These vouchers were issued later than the vouchers received through the MTO program, so control group exposure to new neighborhoods was less than that for treatment compliers.

[25] There is also a literature focused specifically on classmate effects. See Henderson, Mieszkowski, and Sauvageau (1978), Arnott and Rowse (1987), and Hoxby (2001). In addition, Benabou (1993) and Fernandez (2002) model links between neighborhood choice, education, and human capital.

[26] In contrast, Solon, Page, and Duncan (2000) show that correlations between neighboring children in their subsequent educational attainment is small once family background is controlled for suggesting only a limited role for neighborhood factors.

children -- 5 percent compared with 20 percent among those in city neighborhoods. Given the very small samples and low response rates, however, the Gautreaux evidence is suggestive at best.

In a more recent study of children moving out of public housing in Chicago due to Hope VI demolitions, Jacob (2003) found no effect on children's test scores, and found only small changes in neighborhood circumstances despite departure from public housing. Currie and Yelowitz (2000) found that children in public housing projects were less likely to be held back in school than children in similarly poor families without access to public housing and speculate that this resulted from public housing providing better living conditions than these families would have had in the absence of the public housing. Overall, the previous empirical evidence is inconclusive.

Measures. Our primary educational outcomes are continuation of schooling and test scores. We asked parents whether their children had graduated from high school, were still in school, or had dropped out. We asked youth whether they were in school, on summer vacation from school, working during the past week, or none of the above. We measured youth achievement using Woodcock-Johnson reading and math tests, administered during our interviews.

Results. The results for education outcomes are presented in Table 3. The first outcome we consider is whether youth are enrolled in school or have graduated from high school. In column 1, we show that 77 percent of control group females had graduated from high school or were still in school at the time of the survey. Our ITT estimate in column 2 is that an additional 6 percent of the experimental group had graduated or were still in school, with a standard error of .037 and an associated p-value of .10. Assuming there was no program effect for those who did not make a subsidized move through the program, our TOT estimate in column 3 is that there was a 12 percentage point increase for the experimental group compliers. The control complier mean (CCM) in column 4 is estimated to be .71. Thus, the TOT estimate reflects a 16 percent increase in graduation/retention, or (equivalently) a 41 percent decline in the dropout rate. The odds of graduation/retention for control compliers are .710/(1-.710) = 2.45. The odds for experimental compliers are .827/(1-.827) = 4.78, implying a relative odds ratio of 1.95. That is, odds of graduating or staying in school are nearly twice as high for experimental complier females as for females in the control group who would have complied if offered a voucher. This

illustrates that the estimated magnitude of some our estimates can be large even when not statistically significant at the five percent level. For both treatment groups, the effect on the outcome of graduating or still being in school is positive for females and negative for males. In column 9, we find that the male - female difference in TOT effects is -.21, with a p-value of .07.

For the Section 8 group, the effects on the rate of graduating or continuing in school have the same pattern. The 10.5 percentage point difference in effects between genders is not statistically significant. The results for the outcome of being in school or working are not significant for either treatment or gender.

The other main education outcomes we examine are achievement test scores for math and reading. The scores are reported in the metric of z-scores.[27] The reading and math scores for both the experimental and Section 8 groups are insignificantly different from those of the control group. In specifications examining various cutpoints of the score distribution (e.g. 25th, 50th, 75th percentiles, not shown in the table) we also find no significant effects.[28] However, the true effects on complier youth ages 15-20 would have had to have been quite large -- between .4 and .7 of standard deviation (depending on the test and sample) -- in order to be detected 80% of the time with 95% confidence. In comparison, Krueger (1999) finds that class size reductions lead to an increase in test scores of .25 standard deviations, Rockoff (2003) finds that moving up one standard deviation in teacher quality raises test scores by 0.1 standard deviations, and Kane (2003) reports that between the end of fifth grade and the end of sixth grade test scores rise by an average of .30 standard deviations in reading and .25 in math. As discussed in Appendix B, we detected systematic differences in scores on tests administered by particular interviewers, and the results presented in Table 3 are adjusted for potential interviewer effects; the unadjusted results discussed in Appendix B show somewhat larger treatment effects for experimental group females, but none are statistically significant.

V. Risky Behavior Outcomes.

Background. Because a change in residential neighborhood could alter the peers, community norms, and economic opportunities to which an adolescent is exposed, the MTO

[27] The z-score reported here simply takes the scaled score for the test, subtracts the sample mean, and divides by the sample standard deviation. The scaled score ("W score") means for reading and math are 513.7 and 515.4 and the standard deviations are 21.3 and 18.0 respectively.

[28] See Bitler, Gelbach, and Hoynes (2003) on the importance of studying quantile treatment effects.

intervention may have affected the prevalence of risky behaviors among adolescent sample members. Recent economic research has emphasized that teenage risky behaviors are responsive to price and regulatory incentives and to the economic environment. However, Gruber (2001) concludes that these economic factors alone cannot account for the dramatic short-term shifts in behaviors that have been observed, and that analysis of peer effects is the highest priority for future research since peer multipliers might be able to explain how relatively small changes in the environment can spread rapidly through the population.

The nonexperimental empirical literature reveals mixed results on the importance of peer and neighborhood mechanisms in affecting risky behaviors. Studies have found an association between own and peer group smoking and drinking behavior (Krosnick and Judd, 1982; Norton, Lindrooth, and Ennett, 1998) In addition, Case and Katz (1991) found a strong relationship between own and peer group illegal drug use in the Boston Youth Survey. In contrast, Esbensen and Huizinga (1990) found that the level of neighborhood disorganization did not affect neighborhood-level prevalence or frequency of drug use. Studies of a sample of young black women in Chicago found some relationship between pregnancy risk and low neighborhood socioeconomic status (Hogan and Kitagawa 1985) and evidence that this risk was related to lower contraceptive use (Hogan, Astone, and Kitagawa 1985). Billy, Brewster, and Grady (1994) document an association between community characteristics and adolescent sexual activity. In addition, the proportion of managerial workers in a census tract has been shown to be associated with teen childbearing (Crane 1991, Brooks-Gunn *et al*. 1993). However, Evans, Oates, and Schwab (1992) show that peer group effects on teenage pregnancy disappear when metropolitan area characteristics are used to instrument for peer group characteristics.

Measures. We asked youths about smoking, alcohol use, marijuana use, and sexual activity, using items drawn from the National Longitudinal Survey of Youth 1997. We suspect that these self-reported data understate true levels of delinquency and risky behavior, but we believe the differences between groups are still informative.

Results. The results for substance use in the past 30 days are shown for marijuana, cigarettes, and alcohol in Table 4. We found higher substance use for both experimental and Section 8 group males relative to the control group for all three substances, with p-values ranging from .08 to less than .001. The magnitudes of these estimated effects are quite large. For example, the Control Complier Mean is .069 for males in the experimental group having had

alcohol in the past 30 days, and the treatment-on-treated effect was estimated to be about .16 – a more than tripling of alcohol use among compliers. The relative changes were even larger for marijuana use and for smoking.

For females, we found decreases in substance use in both treatment groups relative to the control group -- although the differences were smaller and not always statistically significant, with p-values ranging from .14 to .04. As the treatment effects go in opposite directions by gender, the male-female differences in the treatment effects are larger than the separate effects for each gender. The evidence is quite strong that there was substantial treatment effect heterogeneity for this outcome, with beneficial effects for females and adverse effects for males.

The results for whether the youth were ever pregnant or had gotten someone pregnant are also shown in Table 4. No differences were statistically significant for either gender.

A puzzling feature of these risky behavior results is that control complier means for males are much lower than those for females. In order to provide some context for the prevalence of these risky behaviors relative to a national sample, we compared outcome means for MTO and those for a national sample of youth asked identical questions in the National Longitudinal Survey of Youth, 1997 (NLSY97). We calculated unadjusted prevalences for NLSY97 15-20 year olds weighted to represent the U.S. population. To focus on those in the NLSY97 who are demographically similar to the MTO sample, we also estimated regression models of prevalence based on both youth and adult background characteristics in NLSY97 data, and computed adjusted estimates by evaluating these models at the means of the characteristics of the MTO sample.[29]

Comparing the national estimates to the MTO estimates, we see different patterns for males and females. For females, the MTO control group looks quite similar to the adjusted NLSY97 estimates, while the two treatment groups have lower rates of marijuana use, smoking,

[29] The estimates are shown in Appendix Table A1, and the regression adjustment is described in detail in the table notes. Comparing columns 4 and 5 of Table A1, we see that the demographic adjustments alter the NLSY estimates in predictable ways. Reported rates of teen smoking and marijuana use do not vary much by race and ethnicity in national estimates (Substance Abuse and Mental Health Services Administration, 2002), and we find that the adjusted and unadjusted estimates are similar. In contrast, white youth report using alcohol at a significantly higher rate than minority youth (particularly black youth), and we see that the demographically adjusted estimates are lower than the unadjusted estimates. Teen pregnancy rates are higher among black and Hispanic youth and, as expected, the demographic adjustments result in higher estimates. The set of variables available in the NLSY presumably allow us to go only part of the way towards making the national sample comparable to the MTO sample. Therefore, we might expect to see lower rates of risky behaviors in the MTO control group than in the adjusted NSLY

and alcohol use. This pattern is consistent with beneficial effects from MTO program provision. For males, the treatment groups have rates of risky behaviors that are similar to the adjusted NLSY97 estimates, whereas the control group has lower rates. There are two possible interpretations of this male result pattern. One is that, absent the intervention, male youth in MTO families are less likely to engage in risky behaviors than would be predicted by their demographic characteristics. Under this interpretation, the intervention increased risky behavior among these males. The other possibility is that due to random sampling variability, the control group happened to consist disproportionately of male youth who had positive outcomes on these variables during this time period. Under this interpretation, the significant treatment effects for male youth would be unlikely to reoccur if we reran the experiment.[30]

VI. Mental Health Outcomes.

Background. The pediatric epidemiology literature suggests that living in a high-poverty urban setting is associated with a wide variety of adverse health outcomes for children. Mental health may be particularly sensitive to the neighborhood environment. Poor urban children are extremely likely to witness violence both on the streets and at home (Augustyn et al 1995, Taylor et al 1994), and this exposure can have long-term consequences including behavioral and psychological problems (Groves *et al.* 1993, Famularo et al 1996). Early MTO research demonstrated that treatment group families were significantly less likely to have their children be attacked, robbed, or threatened or to have heard gunfire or seen drugs in their neighborhoods than those in the control group (Katz, Kling, and Liebman, 2001); thus decreased fear and reduced exposure to violence is one plausible channel through which the MTO intervention might have affected youth mental health.

Research has also documented associations among maternal mental health, parenting behavior, and child mental health (Downey and Coyne, 1990; Beardslee, Versage, and Gladstone, 1998; Weissman et al 1997; Hammen and Brennan, 2003; McLeod and Shanahan, 1993). Since the MTO intervention appears to have improved maternal mental health (Kling,

---

estimates in cases in which the NLSY adjustments lowered prevalence and higher rates in the MTO control group than in the adjusted NSLY estimates in cases in which the NLSY adjustments raised the prevalence.

[30] Given the relatively high fraction of control group males having participated in gifted classes at baseline, shown in Table 1, we further investigated the hypothesis about having an unusual draw of control group males by dropping all youth in gifted classes in panels B and D of Table A1. We observed almost no change in the results. There were

Liebman, Katz, and Sanbonmatsu 2004), maternal mental health represents a second channel through which the intervention might have affected youth mental health. Other possible mechanisms include changes in lifestyle (for example more time spent exercising outside), peers, social isolation, and access to medical care.

Measures. To assess mental health, we used a series of items adapted from the National Comorbidity Survey Replication - Adolescent Supplement that focus on non-specific psychological distress, major depressive episodes, and generalized anxiety disorder. Our distress measure, developed by Kessler et al (2002) for the National Health Interview Survey, is commonly scored by summing the scale scores of the items, with the total ranging from 0-24; our results are reported as z-scores, scaling by the standard deviation.[31] Our measures of serious depression and anxiety involve a series of screening questions about the duration and intensity of the feelings and the presence of related symptoms during the worst period in life.[32] These psychological conditions were chosen because they are sufficiently common in the population to ensure that their minimum detectable effects were reasonable with our sample sizes. Since these serious events are still relatively rare, we examined the prevalence of any event, rather than focusing on a limited time such as the past year. Some of the reported events occurred prior to random assignment; we did not attempt to date them precisely, but instead rely on the

---

similarly no important changes in other results in the analyses reported in this paper when youth in gifted classes at baseline are omitted.

[31] This measure is commonly known as the K6 and is based on a six-item Likert scale measuring how much of the time during the past 30 days the youth felt: "so depressed nothing could cheer you up," "nervous," "restless or fidgety," "hopeless," "everything was an effort," or "worthless." The sample mean is 5.0 and the standard deviation is 4.7.

[32] A youth is considered to have had a Major Depressive Episode during his or her lifetime if he or she met the following five conditions. A. The youth experienced a period in which for most of the day he or she felt one of the following: sad, empty or depressed, very discouraged or hopeless about how things were going in his or her life, or loss of interest and boredom with most things usually enjoyed like work, hobbies, and personal relationships. B. Either felt this way most of the day almost every day for a period of two weeks or longer, or for a period of three days or longer and had a year or more in life when felt this way just about every month for several days or longer. C. During times when mood was most severe and frequent, the feelings usually lasted not less than 3 hours a day. D. These feelings were either more than mild, sometimes felt so bad that nothing could cheer him or her up, or sometimes felt so bad that he or she could not carry out daily activities. E. These feelings were accompanied by changes in sleeping, eating, energy, his or her ability to keep mind on things, feeling badly about his or herself, or other problems.

A youth was considered to have had Generalized Anxiety Disorder during his or her lifetime if the following four criteria were met: A. The youth reported there was a period when he or she was either worrying a lot more about things than other people with the same problems, much more nervous or anxious than most people with the same problems, or anxious or worried most days. B. The youth reported being worried about nothing in particular, everything, or more than one specific thing. C. The youth sometimes or often either found it hard to stop the worries or anxiety or could not think about anything else no matter how hard he or she tried. D. The period of being anxious, nervous, or worried lasted at least one month.

assumption that these were few in number (since prevalence of reported events is extremely low at young ages) and that the prevalence of these early events was similar on average between the randomly assigned groups.

Results. The results in Table 5 show improvement for treatment group females relative to the control group on most measures of mental health. P-values were less than .03 for four of the six estimates for females (with depression for the experimental group and distress for the Section 8 group not being significant). The magnitudes of the significant changes were large. Psychological distress as measured by the K6 fell by about one-half of a standard deviation for experimental group compliers. In comparison to the control group, the relative odds of generalized anxiety disorder were about 70 percent lower among compliers in the experimental group and 80 percent lower in the Section 8 group.

In general, the mental health of male youth in the control group was characterized by substantially less distress, depression, and anxiety than that of control group females. All treatment effects for males are small in magnitude and statistically insignificant.

VII. Physical Health Outcomes.

Background. There are four main channels through which the MTO intervention might be expected to affect the physical health of teenage participants. First, the intervention directly changes the physical environment and this might affect conditions such as asthma and injuries.[33] Second, the intervention might change the community norms, peers, and adult role models, which, in turn, might affect health behaviors, such as exercise and diet.[34] Third, the intervention might affect the resources available to the youths, including family income and access to health care.[35] Fourth, the intervention might change the social environment, such as the amount of crime and violence to which the youth are exposed. In addition, it is possible that impacts of the

[33] Children in urban areas are more likely to suffer from asthma (Weiss *et al.* 1992), possibly due to crowding (Weitzman *et al.* 1990), poor air quality (Thurston 1997), stress (Wright 1998), and exposure to allergens from cockroaches, mites, cats, and cigarette smoke (Gelber et al 1993). Accidents are a leading cause of death for youth, and urban youth, particularly those living in poor housing conditions, have higher rates of injuries and accidents (Scharfstein and Sandel 1998, Qunilan 1996).

[34] While research has shown that children in families receiving housing assistance were much less likely to be under-nourished than a comparison group on the public housing waiting list (Meyers et al 1995), obesity and poor nutrition are more common among urban youth than in the population as a whole (Anand et al., 1999). Lee and Cubbin (2002) and Richter et al (2000) document associations between neighborhood characteristics and cardiovascular health behaviors (diet, physical activity, smoking) among youth.

intervention on mental health could affect physical health. For example, research has linked depression to obesity in adolescents (Goodman and Whitaker, 2002).

Measures. Our analysis of youth health is based on self-reported information. We asked questions drawn mainly from the National Health Interview Survey about general health status, injuries, asthma attacks, height, and weight. Self-reported health is strongly related to life expectancy among adults (see Idler and Kasl 1995). While less is known about the predictive power of self-reported health in children, Case, Lubotsky, and Paxson (2002) and Currie and Stabile (2003) find that reported poor health correlates strongly with children's chronic conditions, bed days, and hospitalization episodes. For asthma attacks, our measure follows the standard practice of combining attacks requiring medical attention with other episodes of wheezing or whistling in the chest (Pearce et al, 1998). The purpose of including the other episodes is to ameliorate potential confounding with health care access. We asked for details of any injuries, accidents, or poisonings that required medical attention or were serious enough to limit activities during the previous twelve months, and we focused our analysis on non-sports injuries. We had hypothesized that non-sports injuries might decrease due to a reduction in dangerous external factors in treatment group neighborhoods, but that safety increases might be offset by greater sports participation and lead to no change or an increase in sports injuries. To assess obesity, we collected self-reported height and weight and calculated the body mass index for each individual. Other studies that have collected self-reports and measurements indicate that older adolescents slightly over-report height and under-report weight and that the correlation between self-reported and direct measures is around .9 (Goodman et al. 2000, Brener et al. 2003). Our measure of obesity is body mass index greater than the 95th percentile of the national norms for the youth's age and gender. In interpreting this measure, it is worth noting that national norms for height and weight are benchmarked to 1988-94 data and do not reflect the distribution in 2003, when the population appears to have been heavier.[36]

Results. In Table 6, the prevalence of fair or poor health in the control group was low for females (.10) and especially low for males (.05). There were no significant effects at

[35] Currie and Reagan (2003) find that distance to hospital has a significant effects on the utilization of preventive care among central-city black children.

[36] For health status, asthma attacks, and injuries, we also collected parental reports. However, we present results here only for the youth self-reports, under the presumption that the self-reports will be more accurate. The agreement of youth and parents was only moderate, with kappa statistics of .23 for fair/poor health, .49 for asthma attack in the past year, and .45 for injuries (sports or non-sports) in the past year.

conventional levels for either treatment group on the prevalence of self-reported health being fair or poor for either gender. The effects on asthma attacks were also not significantly different from zero.

Table 6 shows that there was a large and highly significant increase in the prevalence of serious non-sports injuries for males in both the experimental and Section 8 groups relative to the control group.[37] The control group mean is lower for males (.06) than for females (.12) even though in most populations it is males who have more injuries, and the estimated control complier mean for males is zero. We take this as evidence that this particular control group realized very few injuries in the past 12 months, but that such an outcome could be less likely to be true in a different time period (or if we had drawn a second control group).

There were no significant effects on obesity (having body mass index above the 95th percentile of national norms) for either treatment group or gender. Kling, Liebman, Katz, and Sanbonmatsu (2004) found decreases in obesity in the treatment groups among MTO adults and also some increases in exercise and healthy eating. In results not shown in the table, there were significant reductions for females in the experimental group relative to the control group in the mean of BMI percentiles even though the fraction in the upper tail was not reduced.

It is worth emphasizing that for all of these physical health outcomes, the prevalences are low enough that we cannot detect even fairly large relative increases with precision. For example, in order to detect a true effect on obesity 80 percent of the time at 95 percent confidence, the minimum detectable effect size for compliers is about the same as the control group mean.

VIII. Summary of Effects on Primary Outcomes

The preceding sections answer questions about specific outcomes. This section aggregates the information in order to draw summary conclusions about the effects of the MTO

[37] Estimates not shown in the table based on self-reports of any injury (sports and non-sports) requiring medical attention show a control mean of .11 for both genders, insignificant treatment effects for females, and significant intent-to-treat effects of .06 for experimental group and .12 for Section 8 group males. Parental reports of any injuries requiring medical attention, also not shown in the table, had the same sign pattern of effects, but none were statistically significant. Decomposing the injuries (both requiring medical attention or other serious injuries) into sports, non-sports not involving other youth (e.g. stepping on broken glass), or non-sports involving other youth (e.g. fights), we find the following results for males. Sports: control mean .094, E-C ITT .005, S-C ITT .042. Non-sports without other youth: control mean .050, E-C ITT .052*, S-C ITT .055*. Non-sports with other youth: control mean .015, E-C ITT .032*, S-C ITT .023. For the outcome of being "jumped" in the past 12 months, the results were: control mean .181, E-C ITT .002, S-C ITT -.013.

intervention on youth. These summary measures are based on intent-to-treat estimates. Given that compliance rates were around fifty percent, the equivalent TOT estimates would be roughly twice as large.

Summary measures. Table 7 contains summary measures for each domain. The summary measures are τ_g, the mean of the standardized treatment effect sizes, within the domain, as described in Section II. For example, $\tau^{education} = (\tau^{inschool} + \tau^{notidle} + \tau^{readscore} + \tau^{mathscore})/4$.[38] The first column indicates that the mean effect size for the education outcomes for females was .129 for the experimental versus control group contrast. This implies that the average of the effect sizes of the intent-to-treat estimates for the four education outcomes in Table 4 was more than one-tenth of a standard deviation. Although none of the effects on specific individual outcomes had p-values less than .10, the summary measure in aggregating information across outcomes had a p-value of .051.[39] The beneficial effects for females and the opposite-signed effects for males led to a male-female difference of -.150, with a p-value of .067. Other mean effect sizes for education estimates in row 1 were small in absolute value (between .06 and .02) and statistically insignificant.

In order to interpret positively signed coefficients in Table 7 as beneficial effects, the signs of specific outcome effects for risky behavior, mental health, and delinquency were reversed. Thus, the effect size of .137 in column 1 of row 2 indicates that the mean effect size for absence of risky behaviors (which we view as beneficial) was more than one tenth of a standard deviation for the experimental versus control contrast among females, with a p-value of less than .01. While the effect size for the Section 8 versus control contrast was smaller for girls, the adverse effect size for males (higher substance use in the voucher groups) were quite large, so that the difference in mean effect sizes by gender was more than three tenths of a standard deviation (in columns 5 and 6) for both the experimental vs. control and Section 8 vs. control contrasts, with p-values less than .001.

For the mental health summary measures in row 3, the beneficial absence of mental health problems was statistically significant for females in both voucher groups relative to the control group, with mean effect sizes of about .18. The effects for males were insignificant. The

[38] Although we weight each outcome equally, if a decision maker had a reason to value one outcome more than another in his or her loss function, it could be appropriate to apply other weights.

summary-measure effect sizes for physical health show no significant effects for females and significant adverse effects for males of more than one tenth of a standard deviation. For the overall summary measure, averaging across all 15 outcomes, the results show beneficial effects for females and adverse effects for males. The adverse effects for males are largely driven by the results on substance use and injuries for which the control group males appeared to have surprisingly low control group means, as discussed previously.

In results based on pooled samples of females and males, not shown in the table, the summary measures indicate small and insignificant average effects for all domains except mental health. Roughly speaking, beneficial effects for females tend to be offset by adverse effects for males. In the mental health domain, the beneficial effects for females are large enough to generate an overall average that is significantly beneficial even though the effects for males are adverse.

Internal validity. Nearly all of the results discussed in this paper are based on self-reported measures. Because some participants and interviewers were aware of treatment status, it is possible that some survey responses reflected what the participants or interviewers thought the investigators wanted to hear rather than the truth. However, for social desirability bias to be consistent with the results, it would have to be very complex – positive bias for female substance use and mental health, negligible for female physical health, and negative for males – and the available evidence is not consistent with a broad, systematic effect of this sort.[40] Also, in related MTO research studying youth arrests (Kling, Ludwig, and Katz 2004) and adult earnings (Orr et al. 2003), self-reported and administrative data have generated similar results.

In order to investigate the potential effects of survey attrition on the results, we computed treatment effect estimates under various assumptions about the values of data missing due to

---

[39] In contrast, the nondirectional F-test on the four education outcomes had a p-value of .36, consistent with the lower power of this test to detect effects that are consistently of the same sign and magnitude across outcomes within a domain.

[40] On measures where one might expect a strong social desirability bias, such as obesity, poor health, dropping out of high school, or being idle (not working or in school), there are not significant treatment effects. Moreover, using the same type of demographic adjustments as in Table A1, we find that the MTO treatment groups are within a couple of percentage points of similar youth in the NLSY97 on these measures, wheras social desirability bias might predict that they would report significantly more desirable behavior. A lack of systematic social desirability bias between the treatment and control groups is consistent with a low level of awareness among youth about treatment status from a housing voucher lottery that their parents participated in when they were ages 8-16 and how it affected their residential location 4-7 years later when they were 15-20 years old. To the extent that outcomes like risky behavior are under-reported by a constant factor (say, two-thirds of the time) in all groups, the lower prevalence in

survey attrition. The methods and results of this investigation are presented in Appendix C. While worst case assumptions about missing data (e.g. everyone with missing data smoked in the past month) can change the results a great deal, the sign of summary measure estimates do not change under less extreme assumptions about missing data. We also found that our results were not sensitive to alternative econometric models.[41] The inclusion of covariates in our regression estimation made negligible differences in most estimates.[42]

External validity. One consideration in generalizing these results to other contexts is whether these results are relatively consistent across the five MTO sites. We do not find systematic evidence of site differences in treatment effects.[43] The sign pattern of effects across sites was fairly consistent within gender for the domains with significant effects.[44] Another consideration for external validity is the extent to which the findings in this study generalize to other age groups. However, the evidence of beneficial effects for females and of zero or

---

self-reported data does reduce the statistical power to detect treatment effects, but does not bias their direction or result in the appearance of treatment effects when the true effects are zero.

[41] Two advantages of calculating mean effect sizes using a linear regression, as we do, are that we can combine continuous and binary outcomes in a simple manner and that the estimates can be directly rescaled between intent-to-treat and treatment-on-treated magnitudes. For binary outcomes, a one percentage point effect has a larger effect size when the base prevalence (and the standard deviation) is smaller, making it a measure of relative change. For domains consisting entirely of dichotomous outcomes, such as risky behavior and physical health, we also investigated alternate mean effect size measures based on probit (average of z-scores) and logit (average of log odds) models of intent-to-treat effects. Since the scale of measurement differs across the methods, the magnitudes of the estimates necessarily differ. However, the sign pattern of effects is the same across methods, and the relative magnitudes of female and male effects are also similar. Inference about statistical significance is very similar across the alternate measures, as reflected in the p-values for risky behavior and physical health mean effect sizes presented in Table A2.

[42] Using a seemingly unrelated regression in equation 11, but replacing ν^2 with Y, we tested the difference between our regression-adjusted and unadjusted estimates. There were 60 estimates for two genders, two treatment groups, and the 15 outcomes in Tables 3-6. We found four differences with p-values less than .15: E-C female graduated HS or still in school, E-C male graduated HS or still in school, E-C male working or in school, S-C male ever gotten someone pregnant. In these four cases, the regression adjustment lead to more beneficial treatment effects.

[43] F-tests on four site interactions with treatment status show one p-value less than .05 for 30 tests on females and males in the experimental group for the outcomes in Tables 3-6, and two p-values less than .05 for 30 tests on females and males in the Section 8 group. The sample sizes within each site-gender-treatment group are small and the power of this test to detect site differences is low.

[44] This analysis was based upon equations (4) and (7)-(10), with Z consisting of five treatment-by-site indicators instead of a single treatment indicator. For each gender, we examined the signs of 10 estimates corresponding to the five site estimates for the experimental group and the five for the Section 8 group. For the overall summary measure, 9 of 10 estimates were beneficial for females, and 8 of 10 were adverse for males. For risky behavior, 8 of 10 were beneficial for females and 10 of 10 were adverse for males. For mental health, 10 of 10 were beneficial for females. For physical health, 8 of 10 were adverse for males.

negative effects for males is concentrated in the outcomes such as risky behavior and psychological problems that that have substantial prevalence only for older teenage youth.[45]

Multiple testing. Although the use of summary measures reduces the number of simultaneously presented statistical tests, we should still expect five percent (1 or 2) of the 30 estimates presented in Table 7 to have t-statistics of 1.96 or higher even if the true effects were all zero. While the statistical inference discussed above is perfectly valid when looking at each comparison on its own, for a general assessment of the full range of outcomes it is important to consider the number of tests being performed.[46]

One way to assess the chance of false significances within a set of estimates is to calculate a familywise error rate, which is the probability of rejecting at least one true null hypothesis within the set. In order to calculate a familywise error rate, one needs to specify the members of the family before examining the data. In our case, prior to seeing the data, we planned to estimate effects for females and males pooled together, and for females and males separately. The estimates highlighting the differences between females and males were added to the analysis to further explore this apparently important pattern in the data, but were not pre-specified. So for our analysis of familywise significance, we use summary measures for the five rows in Table 7 (four domain plus an overall measure), three pre-specified subgroups (pooled, female, male), and two treatment groups, for a total of thirty estimates. Per-comparison and familywise adjusted p-values are shown in Table 8, with rows for the eleven domain-gender-contrast results that had per-comparison p-values less than .05 in Table 7. Although not shown

[45] We focus on ages 15-20 in this paper because many of our outcomes (dropping out of school, substance use, sexual activity, depression, and generalized anxiety) have very low prevalence for younger ages. However, we conducted reading and math tests and collected self-report data on psychological distress and general health status from younger youth. Of the four measures for which we have data across age groups, only distress had a significant treatment effect (for experimental group females) among respondents aged 15-20. Analysis of these data for ages 10-14 finds no significant treatment effects for either gender, and no significant differences between genders. On the basis of this evidence, we do not find any evidence to support one of our original hypotheses, which was that the treatment effects would be larger for younger individuals, since these individuals would have spent a larger fraction of their lives in new surroundings than older treatment group members.

[46] To be clear about the circumstances in which adjustments for multiple testing might be applicable, consider the following hypothetical example. If a principal at an all-girls Catholic school wanted to know what the effect on risky behavior outcomes was of moving out of high-poverty neighborhoods for female youth in the experimental group, it would be perfectly reasonable to look only at the effect size of .137 and the p-value of .01. However, if a social scientist wanted to search over all the results presented in this study and discuss only those with t-statistics greater that 1.96, then he or she should acknowledge that the possibility is quite high that one or more of the results (e.g. among 30 in Table 7) may have a t-statistic that large in our sample even if the true effect were zero. That is, the chance of at least one "false significance" can be quite high with a large number of tests.

in the table, nineteen other estimates (for a total of thirty in the family) are used in the calculations of familywise significance.

The first three columns of Table 8 show per-comparison p-values using different estimation techniques. The first column contains p-values based on asymptotic standard errors assuming negligible variation from normalization of the effect sizes by the standard deviation σ_{gk}, and is based on equation (10) and the coefficients and standard errors in Table 7. The second column contains p-values that use the delta method approach described in equation (11) to directly incorporate sampling variability of σ_{gk} in the asymptotic standard error calculation. The third column calculates p-values using bootstrap estimation, as discussed in Appendix A. The results indicate that the per-comparison p-values are similar for the three methods.[47] Thus, for simplicity, the standard errors in Tables 3-7 are asymptotic standard errors, analogous to those in column 1. However, we believe that the bootstrap estimates are slightly more accurate in that they account for the finite sample properties of our estimators (Horowitz 2001).[48]

Columns four through six of Table 8 contain familywise adjusted p-values. Columns 4 and 5 use the Bonferroni-Holm method discussed in Section II, applied to columns 1 and 3 respectively. Our preferred approach is the Westfall-Young method in column 6 that incorporates correlation between estimates within the family, as outlined in Appendix A. Since we prefer the bootstrap among the per-comparison methods, we have modified the standard Westfall-Young procedure to incorporate these bootstrap per-comparison estimates.[49] The Bonferroni-adjusted values in column 5 are much simpler to compute and turn out to be similar to those from the Westfall-Young procedure, due to the relatively low correlation between effect size estimates across domains in our application.

For the overall summary index experimental vs. control group contrast (E-C) in column 6, the adjusted p-value was .03 for females. Thus, if one searched for significance over the entire

[47] The delta-method p-values in column 2 are slightly smaller than the asymptotic p-values in column 1 when the treatment effects are beneficial (as they are for females) and slightly larger when the treatment effects are adverse (as they are for males). In our application where most of the components of our summary measures are binary outcomes with means above .5 (e.g. absence of risky behaviors), the covariance between the treatment effect and the control group standard deviation is positive. This covariance dampens the sampling variability of the mean effect size measure when the treatment is positive in sign, and amplifies it when the treatment effect for a summary measure is negative in sign.

[48] Boostrap p-values are slightly larger than delta method p-values for all contrasts except S-C female mental health.

[49] For example, the asymptotic p-values for male risky behavior appear to be too small when simulating the null hypothesis using this data. This type of inaccuracy causes the Westfall-Young estimates to be too large, in some

set of 30 tests, an effect this large would be observed for at least one among this family of tests about 3 percent of the time if the null hypothesis of no effect were true. Similarly, the magnitude of the overall mean effect size for males was adverse enough that an effect this large would be observed over this family of tests about 8 percent of the time if the null hypothesis of no effect were true. For the Section 8 vs. control group (S-C) contrast, there is a greater likelihood of at least one result at least as large in absolute magnitude as those observed for the family even if the null hypothesis were true: about 13 percent for females and about 42 percent for males.

The adjusted p-values for mental health for females were .03 and .02 for the E-C and S-C contrasts. Thus, even when we apply the very stringent criteria that 95 percent of the time no true null hypotheses in the entire family of 30 tests will be falsely rejected, the beneficial effects on female mental health were sufficiently large to reject the null hypothesis of no effect. Adjusted p-values for the risky behavior and physical health summary measures were greater than .05, and were greater than .15 for the nineteen members of the family not shown in Table 8.

Although the female-male differences do not fit formally within the testing framework outlined above since they were not prespecified, we can calculate the size that the family of tests would have to have been in order to have an adjusted p-value of .05 given the estimated magnitude and standard error of the difference. For the overall female-male difference for the experimental-control contrast, the family could have included over 6000 tests, and we still would have rejected the null hypothesis of no true effect at a 5 percent level of familywise significance given the large magnitude of the overall mean effect size difference by gender.

We have focused our analysis of familywise significance on a relatively small set of summary measures, since the probability of at least one false significance increases as the size of the family increases. It is possible to apply the same multiple testing framework discussed above to a larger family, such as the 90 estimates for specific outcomes (pooled/male/female, two contrasts, and fifteen outcomes in Tables 3-6). However, the estimated treatment effects need to be extremely large to be 95 percent confident that there are no false significances in the entire family. Only one adjusted p-value for a specific outcome is below this threshold, .02 for female generalized anxiety in the S-C contrast (using the Westfall-Young method). While 16 of the 90 estimates in this family have per-comparison p-values of less than .05 (and we would expect only

---

cases exceeding their Bonferonni bound. Our modified Westfall-Young procedure uses the empirical distribution of this simulation to set appropriate p-values.

4.5 by chance under the null hypothesis), it is reasonably likely that at least one of those 16 is a true null hypothesis that was falsely rejected. Thus, we are more cautious in interpreting results about specific outcomes.

In sum, the multiple testing analysis described in this section indicates that the strongest conclusions about primary outcomes -- overall beneficial effects for females, with particularly large benefits for female mental health -- are unlikely to be the result of sampling error.

IX. Mechanisms

This section presents evidence that can help distinguish among alternative mechanisms that may have produced the observed results. We begin by discussing experimental impacts on intermediate outcomes, which we refer to as mediators. Then we turn to evidence on mobility rates, time spent in new neighborhoods, and the impacts of different types of neighborhoods.

Our analysis of mechanisms is complicated by the need to explain the strikingly different patterns of effects for males and females. We had initially hypothesized that we might observe differential impacts for male and female youth. This hypothesis was motivated by evidence in earlier MTO work which showed a reduction in behavior problems for boys and no such impact on girls (Katz, Kling, and Liebman 2001) as well as evidence from studies of welfare reform that indicated that gains in maternal self-sufficiency led to improvements in mental health for male children but not for female children (Bos et al 1999).[50] Thus our hypothesis was that we would observe larger benefits from the treatment for male youth than for female youth – the opposite of what we found.

We further speculated that two mechanisms might produce the larger benefits for boys. First, because males are more likely than females to be involved in gangs, violence, and drug use in high poverty neighborhoods, the gains from escaping these neighborhoods might accrue disproportionately to boys. Second, boys might have an easier time adjusting to the new neighborhoods and making new friends, possibly because they are more likely to engage in

[50] All of the nonexperimental papers that we are aware of showing gender differences in neighborhood effects find larger beneficial effects for boys from living in advantaged neighborhoods than for girls. See Entwisle, Alexander, Olson (1994), Ensminger, Lamkin and Jacobson (1996), Ramirez-Valles, Zimmerman, and Juarez (2002), Crane (1991), Halpern-Felsher et al (1997). The predominant mechanism proposed in these papers is that boys spend more time hanging out in the neighborhood (rather than in the home) and therefore are influenced more heavily by the neighborhood.

sports.[51] While we have no shortage of alternative theories that can explain why gains were concentrated among females, we want to emphasize that these theories were developed after seeing the results. This section of the paper is thus an exploratory analysis helping to provide a new set of hypotheses that are consistent with the MTO data, but which need to be verified in subsequent work.

Mediators. In the MTO interim evaluation, significant resources were allocated to collecting data on intermediate outcomes that could help explain the results and distinguish among alternative theories for why the intervention had any effects that it had. The mediators include measures of neighborhood quality and safety, home environment, school environment, peers and adult role models, experiences in school, and health-related behaviors. Experimental impacts for 74 mediators are presented in Tables A3-A7. These results are summarized in Table 9 which presents mean effect sizes for 14 types of mediators. [52]

The results indicate that the intervention had large, statistically significant, impacts on the general characteristics of the neighborhoods in which MTO youth lived (including poverty rates, neighborhood satisfaction, and perceptions of safety). These results are present for males and females in both treatment groups.[53] The point estimates are larger for females than for males and larger for the experimental group than for the Section 8 group. However, none of these between-group differences are statistically significant. For experimental group males and females, there are also statistically significant impacts on the school environment (including reduction in the share of classmates who were non-white, receiving free lunch, and limited English proficient), but the male-female differences are not statistically significant.

The two mediator types with a significant male-female differences were adult role models (the difference is significant for the experimental group only) and adult mental health (the

[51] In Katz, Kling, and Liebman (2001) we found that girls in the treatment groups had fewer friends than girls in the control group, whereas the treatment had no impact on number of friends for boys.

[52] In our discussion of mediators we are analyzing a large family of estimates, again raising issues of multiple testing. Because the goal of this exercise is to describe sets of theories that are and are not consistent with our data (rather than to formally test specific mechanisms), we do not conduct a formal multiple testing analysis in this section of the paper. However, we do present some simple Bonferroni bounds in order to provide some perspective on how results would be interpreted in a confirmatory analysis.

[53] Note that the four summary general neighborhood measures (male and female for each treatment group) have t-statistics ranging from 3.3 to 6.8 and p-values .001 or smaller. Thus, even in a family of 84 tests (14 rows in Table 9 x 6 estimates each), the Bonferroni bound on the adjusted p-value would have been below 0.10 for all four estimates, assuming this family of tests had been pre-specified. More generally, any estimate in this table with a t-statistic greater than 3.3 will have an adjusted p-value below 0.10 based on the Bonferroni bound.

difference significant for the Section 8 group only).[54] The probability of having two or more significant male-female contrast in a family of 28 tests is high, even under the null hypothesis that none of the mediators have any effect, and we do not place great weight on these results. There were also statistically significant effects for experimental group females on absence of housing problems, future expectations (chances of completing college or finding a good job) and healthy environment (including sports participation and frequency of exercise), but the male-female differences on these mediator types were not statistically significant.

The MTO intervention had no statistically significant impacts on many mediator types including parenting practices, peer characteristics, school engagement, and access to health care. These findings could mean that these channels are not the ones through which MTO's effects occurred, or they could indicate that the standard survey methods for measuring these concepts are not adequate. Overall, the extensive set of mediators we have reviewed does not present a clear answer to the question of why the MTO intervention had more benefits for girls than for boys. Indeed, for some specific mediators that show evidence of "correctly signed" gender differences – such as "friends who use drugs" or "works hard in school" – it seems at least as likely that these intermediate outcomes are the result of the same mechanisms that produced the gender differences in the main outcomes as it is that they are an earlier link in a causal chain.

The results on mediators do provide clear evidence that the MTO intervention changed major features of the neighborhoods in which the youth were living such as the socioeconomic status of neighbors and their sense of safety. The fact that families with female children and families with male children appear to have moved to similar neighborhoods suggests that the differential outcomes by gender are not the result of exposure to different types of environments. Instead, male and female youth appear to respond to their environments in different ways. [55]

[54] The pattern of results for youth could have come about in part from feedback effects between adult mental health and youth outcomes, but this mechanism appears to be strong only for households with youth ages 15-20 and not for all ages of children. Kling, Liebman, Katz, and Sanbonmatsu (2004) find that the gender composition of the children in the household overall does not interact with mental health treatment effects or with economic self-sufficiency or physical health effects.

[55] Another approach to addressing the question of whether neighborhood characteristics differing between females and males could be driving the outcome differences would be to focus the analysis on opposite-gendered sibling pairs where the neighborhoods themselves are shared by the pair. However, there are only 128 of these pairs among our sample of 1807 youth. In related work, Kling, Ludwig, and Katz (2004) use a larger sample of sibling pairs ages 15-25 from administrative arrest data, and find that the overall sample and the sibling pair sample generate similar results.

Residential mobility. The use of a program voucher necessarily involved a residential move. One possibility is that the moves themselves could have generated the program effects, independent of the attributes of the new location. An important aspect of the residential location patterns in this study is that roughly two-thirds of control group youth had moved since the time of random assignment. In addition, it turns out that mobility was higher among treatment group females than among treatment group males.[56] We also examined the changes in location among compliers after their initial program move, and the proportion moving to higher poverty neighborhoods was higher among females than males.[57] We had hypothesized that greater mobility (holding neighborhood quality constant) would be disruptive and therefore have a deleterious effect on outcomes. So the greater mobility and more beneficial treatment effects experienced by females are at variance. In light of these results, we believe it is likely that it was the characteristics of the neighborhoods to which people moved rather than the moves themselves that produced the results reported in this study.

Exposure time. Because families in the MTO experiment were randomly assigned over a three-year period, it would, in principle, be possible to learn how treatment effects vary with the duration of exposure to new neighborhoods by comparing effects for youth randomly assigned in the first half of the enrollment process to those from the second half of the enrollment process. In practice, our sample sizes are simply too small for within-gender subgroup results to be informative. In addition, the groups assigned earlier and later may not be comparable and there may be calendar time as well as exposure time effects. An alternative is to compare our results to those in several earlier, smaller scale studies.[58] In the domains of education, risky behavior,

[56] The number of moves was about 1.5 for females in both treatment groups versus 1 in the control group, while for males the number of moves was about 1.3 in the experimental group and 1.4 in the Section 8 group versus 1.2 in the control group. The difference in these treatment effects, with larger effects for females, was statistically significant for both treatments. This pattern of greater effects on mobility for females is evident in analysis of the probability of making a program move, making at least one move, and making two or more moves, although none of the gender differences in these indicators is individually statistically significant for the experimental group. These data are from administrative data on addresses collected from credit bureaus, National Change of Address, housing authorities, program files, and in-person tracking for this study. Estimates based on self-reports by adults of mobility also show treatment effects on mobility were larger for females than males, but these differences are smaller in magnitude than those based on address tracking file data and not statistically significant. Results on mobility from both administrative and self-reported data are shown in Appendix Table A8.

[57] The share of compliers moving to higher poverty neighborhoods between their initial program move and the survey in 2002 was .50 for females versus .38 for males in the experimental group, and .38 for females versus .30 for males in the Section 8 group.

[58] Early MTO work based on about 350 Baltimore children ages 5-12 found large changes in neighborhood circumstances for the experimental group relative to the control group and positive effects on reading and math test scores over the first four years after random assignment (Ludwig, Ladd, and Duncan, 2001). A study of 168

mental health, and physical health, none of the youth results reported as statistically significant at a single site in the early results is also statistically significant in this paper. For example, the reductions in injuries and asthma attacks in 1997 data from Boston are not confirmed for the entire 5-site sample using 2002 data, nor are they confirmed using data only from the Boston site for the same age range or same birth cohort. Given the small sample sizes of the early studies, we cannot reject the hypothesis that there was no difference between the results in this paper and any of the earlier results in these domains, even though the results do not have the same patterns.

Using panel data on a large sample of arrest records, Kling, Ludwig, and Katz (2004) find that the treatment effects for males are more pronounced 3-4 years after random assignment than after 1-2 years for increased nonviolent crime arrests in the experimental group relative to the control group. These results, in combination with the imprecise comparisons to the earlier survey studies, suggest that it is likely that the effects reported in this paper did not occur shortly after the initial moves, but manifested themselves years later. Since we had hypothesized that if moving per se was an important factor for outcomes then the effect on the outcome would be seen in the period shortly after random assignment, we interpret this as further evidence against the importance of moving per se as a mechanism leading to the treatment effects on outcomes found in this paper.

Neighborhood Types. One of the main policy questions underlying the Moving to Opportunity demonstration is whether most of the gains from leaving high-poverty public housing accrue simply from moving out of the most distressed neighborhoods into slightly-lower poverty neighborhoods or whether a move to a very-low poverty neighborhood is required. A simple way to explore this issue is to compare the patterns of experimental group and section 8 group TOT estimates. For females, most of the mental health TOT estimates are similar for the two treatment groups, suggesting that the mental health benefits may come primarily from getting out of the crime-ridden housing projects (and possibly that any further improvements in

---

children ages 6-10 at the MTO site in New York did not find effects on test scores for the experimental versus control group overall after three years -- although it did find positive effects on test scores for a sample of male youth (Leventhal and Brooks-Gunn 2003b). In samples of about 175 boys and 175 girls ages 8-18 at the New York site of MTO three years after random assignment, there were no effects of the MTO intervention on cigarette smoking, but higher alcohol use among female (but not male) youth in the experimental group relative to the control group; boys in the experimental group were less likely to report that they were anxious or that they were depressed relative to the control group, and there were no significant effects for girls or for either gender in the Section 8 group (Leventhal and Brooks-Gunn 2003a). Early research with data from only the Boston site of MTO 1-3 years after random assignment found that the experimental group (but not the Section 8 group) had lower prevalence of asthma attacks and injuries than the control group (Katz, Kling, and Liebman 2001).

mental health from moving to even lower poverty neighborhoods are offset by an increase in feelings of isolation from being further from the city). For female risky behaviors, the experimental TOT estimates are somewhat bigger than the Section 8 ones; thus while the majority of the gains seems to come from getting out of the highest poverty neighborhoods, additional gains accrue from moving to even lower poverty neighborhoods. Finally, for female education there are negligible effects for the Section 8 group but much larger gains for the experimental group. A possible explanation for this finding is that Section 8 group moves often resulted in the youth remaining in the same school system whereas experimental group moves often resulted in a change of municipality and therefore of school district.

There are at least two reasons to be cautious about reading too much into these results. First, the differences between the experimental and Section 8 TOT effects for the same gender are not statistically significant on any of our fifteen primary outcomes. Second, the compliance patterns of the two treatment groups appear to be different. Thus the TOT effects are estimated on different populations of compliers.[59]

Competing hypotheses. Because we are not aware of any other intervention of this sort that shows more beneficial impacts on girls than on boys, we speculate on some possible reasons for these different results. In particular, we are interested both in theories that would explain why girls are the main beneficiaries of moves out of high-poverty neighborhoods and in theories that would suggest negative influences of more affluent neighborhoods on boys and which could therefore explain negative results as well as zero results that might have come from offsetting negative and positive influences. We offer the following eight hypotheses with the hope that further investigation in future studies can distinguish among them:

1. *Relative performance and self-esteem*. Moving to a neighborhood with higher mean educational performance could be demoralizing for low-achieving students, particularly if the students are more visible because they are racial or ethnic minorities. It is possible that these effects differ by gender.

[59] A common assumption is that, among families with similar observed characteristics, the "more motivated" families would be less likely to reside in high poverty neighborhoods and would have had better outcomes -- such as well-adjusted children -- regardless of their neighborhood. In preliminary studies of the MTO demonstration's Baltimore and Boston sites, however, there was evidence that households whose children would have had high rates of delinquency and behavior problems if they had stayed were more likely to leave the high poverty neighborhoods; this alternative sorting behavior would lead non-experimental studies to give understated estimates of the causal effect of neighborhoods on negative behaviors (Ludwig, Duncan, and Hirschfield, 2001; Katz, Kling and Liebman 2001). In related work on the selection process using MTO data from 2002, we find evidence for both types of selection (Liebman, Katz, and Kling 2003). The process appears to be complex, to differ by outcome, and to vary across demographic subgroups.

2. *Peer matching*. If youth from families who moved through MTO find themselves shifting from the middle of the achievement distribution at the old schools to the bottom of the distribution in the new schools they may end up associating with people who are similarly placed in the distribution. This lower tail of the achievement distribution in the new neighborhoods, especially for boys, could disproportionately be made up of drug users and people who otherwise provide negative peer influences.
3. *Cultural conflict.* Use of language, style of dress, and taste in music could make adjustment to new circumstances more difficult for male youth.[60]
4. *Domestic violence and sexual abuse.* Girls may suffer disproportionately from domestic violence and may be victims of sexual abuse, and the MTO intervention may have reduced their exposure to such events.
5. *Stereotyped expectations*. Male youths may feel greater pressure to fulfill peer expectations about how someone who grew up in housing projects should act.[61]
6. *Institutional Capacity*. It is possible that the destination neighborhoods lacked some of the institutional support for at-risk boys that might have been present in the origin neighborhoods.
7. *Connections to old neighborhoods*. Among MTO movers, perhaps boys were less successful in breaking ties to undesirable peers and activities in the prior neighborhoods.
8. *Role models*. Given that nearly all of the MTO families are headed by single mothers, a female MTO youth will generally have a same-gendered role model available in her home. Male youth may have fewer male role models, particularly ones of their same race or ethnicity, in the new neighborhoods.[62]

There has been a broader trend over the past two decades of gains in education and employment for minority women that has not been shared by minority men (Altonji and Blank, 1999). MTO moves may remove the barriers to benefiting from these gains for female sample members whereas the males, even in lower-poverty neighborhoods, have poor prospects. Yet even if this historical parallel contains a kernel of truth, identifying the nature of these barriers is itself a challenge.

Our current data offer little support for the relative performance or peer matching hypotheses, in the sense that the educational environments of youth do not appear to have been affected a great deal or to have differed by gender. Moreover, the minority representation in most schools and residential areas of youth remained very high regardless of MTO program assignment -- making MTO youth less obviously noticeable than they would be in, say, all-white schools. Of course, cultural conflict could be quite important even in a racially homogeneous

[60] Akerlof and Kranton (2000) discuss models in which group identity enters the utility function.
[61] Bernheim (1994) analyzes a model of social interactions in which status concerns produce conformity. Ferguson (2001) presents evidence that there are gender differences in what confers status among teenagers.
[62] Borjas (1995) studies the importance of within-neighborhood ethnicity externalities. See also Lazear (1999).

environment, and our survey did not collect information on whether there were differences in speech, taste, or style. Similarly, we currently have no data about expectations of youth behavior or the time path over which they were formed in order to assess the importance of stereotypes. We do have some data about the prevalence of visits to baseline neighborhoods, and the sign is consistent with boys making more visits but the male-female difference is not statistically significant. Regarding role models, there is some significant evidence that female youth are more likely to have three or more adults to whom they are comfortable talking about problems, and the mean effect size on adult role models did rise for experimental group girls relative to controls. In ongoing work, we are collecting open-ended interview data from both parents and youth from which we hope to garner more information about the relative plausibility of these hypotheses and how they might be tested and distinguished more systematically.

X. Conclusion

This paper has examined the effects of moving out of high-poverty neighborhoods on the outcomes of teenage youth. The randomized design of the Moving To Opportunity demonstration was used to compare groups of otherwise similar youth all initially living in high-poverty public housing, some of whom were in families offered housing vouchers to help them move to neighborhoods that had lower poverty rates, improved safety, and more resources. The general direction of the effects, with females benefiting from program participation and males being adversely affected, is summarized in Table 7. For females, seven of eight domain-specific mean effect size estimates were positive. The overall mean effect size is positive and statistically significant at the 0.1 percent level. For males, six of the eight domain-specific coefficients are negative, and overall effects are negative and statistically significant. All of the male-female differences are in the direction of greater improvement for females.

Among females, the mean effect sizes were largest for mental health (major depression, generalized anxiety, and non-specific distress). For experimental group females there was also less risky behavior (such as use of marijuana, cigarettes, and alcohol) and better educational outcomes (although specific outcomes such as dropout status and test scores were not significant).

Male youth in both the experimental and Section 8 groups had significantly more risky behavior than controls. The average effects for education and mental health were insignificant

for males in both treatment groups. Males in both treatment groups had more physical health problems than in the control group (with especially large effects on injury rates).

To account for the large number of statistical tests performed, we based our statistical inference on a multiple testing framework. Our family of tests contains three subgroups (pooled genders, females, and males), two contrasts (experimental vs. control and Section vs. control), and five summary measures -- for a total of 30 estimates. A conventional five percent level of statistical significance for any one outcome would lead one to expect 1.5 false significances in a family of 30 estimates if the null hypotheses of no effect were true. Using a more stringent familywise error rate approach, we show in Table 8 that our strongest results hold at the 95 percent level of familywise confidence. These strongest results are for overall effects on females in the experimental group, and for effects on mental health for females in both treatment groups. At this 95 percent familywise confidence level, there is less than a five percent chance under the null hypothesis of no effect that we would observe even one mean effect size from among the family of 30 as large as those actually estimated.

The conclusion that this intervention had beneficial effects for female teenagers is consistent with comparisons of the control group with national samples, demographically weighted to match our sample population. In contrast, these comparisons suggest that the adverse effects we find for males may have resulted from a control group with atypically positive outcomes, due to some combination of sample attrition and random sampling variation.

Regardless of whether the effects of this intervention on males were adverse or negligible, there remains the puzzle of why the female and male results differ. Families with female children and families with male children moved to similar neighborhoods, suggesting that the differential outcomes by gender are not the result of exposure to different types of environments. Instead, the male and female youth in these families appear to respond to their environments in different ways.

In sum, we reject the hypothesis that neighborhoods had only limited effects on these youth. We have identified some important beneficial effects of moving out of high poverty neighborhoods on the outcomes of female teenage youth, and we have established that similar benefits did not accrue to males.

References

Aaronson, Daniel (1998). "Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes." *Journal of Human Resources* 33: 915-946.

Akerlof, George A. and Kranton, Rachel E. (2000). "Economics and Identity." *Quarterly Journal of Economics* 115: 715-753.

Altonji, Joseph and Blank, Rebecca (1999). "Race and Gender in the Labor Market." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley Ashenfeleter and David Card. Amsterdam: North Holland.

Anand, Rajen, S.; Basiotis, P. Peter and Klein, Bruce. W. (1999). "Profile of Overweight Children." *Nutrition Insights* 13 (May). Washington, DC: U.S. Department of Agriculture.

Angrist, Joshua D.; Imbens, Guido.W. and Rubin, Donald B. (1996). "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91: 444-472.

Angrist, Joshua D. and Lang, Kevin (2002). "How Important are Classroom Peer Effects? Evidence from Boston's Metco Program." Working Paper No. 9263. Cambridge, MA: NBER.

Arnott, Richard and Rowse, John (1987). "Peer Group Effects and Educational Attainment." *Journal of Public Economics* 32: 287-305.

Augustyn, Marilyn; Parker, Steven; McAlister-Groves, Betsy and Zuckerman, Barry (1995). "Silent Victims: Children Who Witness Violence." *Contemporary Pediatrics* 12: 35-57.

Beardslee, William R.; Versage, Eve M. and Galdstone, Tracy R. (1998). "Children of Affectively Ill Parents: a Review of the Past 10 Years." *Journal of the American Academy of Child and Adolescent Psychiatry* 37: 1134-1141.

Benabou, Roland (1993). "Workings of a City: Location, Education, and Production." *Quarterly Journal of Economics* 108 (3): 619-652.

Bernheim, B. Douglas (1994). "A Theory of Conformity." *Journal of Political Economy* 102: 841-877.

Billy, John O.G.; Brewster, Karin L. and Grady, William R. (1994). "Contextual Effects on the Sexual Behavior of Adolescent Women." *Journal of Marriage and the Family* 56: 387-404.

Bitler, Marianne P.; Gelbach, Jonah B. and Hoynes, Hilary W. (2003). "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." Working Paper no. 10121. Cambridge, Mass.: NBER.

Bloom, Howard (1984). "Accounting for No-shows in Experimental Evaluation Designs." *Evaluation Review* 8 (April): 225-46.

Blume, Lawrence E. and Durlauf, Steven N. (2001). "The Interactions-Based Approach to Socioeconomic Behavior." In *Social Dynamics,* edited by Steven N. Durlauf and H. Peyton Young. Cambridge: MIT Press.

Borjas, George J. (1992). "Ethnic Capital and Intergenerational Mobility." *Quarterly Journal of Economics* 107 (1): 123-150.

Borjas, George J. (1995). "Ethnicity, Neighborhoods, and Human-Capital Externalities." *American Economic Review* 85: 365-390.

Bos, Johannes; Brock, Thomas; Duncan, Greg; Granger, Robert; Huston, Aletha and McLloyd, Vonnie. (1999). "New Hope for People with Low Incomes: Two-Year Results of a Program to Reduce Poverty and Reform Welfare." New York: Manpower Demonstration Research Corporation.

Bouchard, Thomas J. Jr. (1997). "IQ Similarity in Twins Reared Apart: Findings and Responses to Critics." In *Intelligence, Heredity, and Environment*, edited by Robert J. Sternberg and Elena Grigorenko. Cambridge: Cambridge University Press, 126-162.

Boyce, W. Thomas; Jensen, Eric W.; James, Sherman, A. and Peacock, James L. (1983). "The Family Routines Inventory: Theoretical Origins." *Social Sciences and Medicine* 17 (4): 193-200.

Brener, Nancy D.; McManus, Tim; Galuska, Deborah A.; Lowry, Richard and Wechsler, Howell (2003). "Reliability and Validity of Self-reported Height and Weight Among High School Students." *Journal of Adolescent Health* 32: 281-287.

Brock, William A. and Durlauf, Steven N. (2001). "Interactions-based Models." In *Handbook of Econometrics, Volume 5*, edited by James J. Heckman and Edward E. Leamer. Amsterdam: North-Holland, 3297-3380.

Bronfenbrenner, Urie (1979). *The Ecology of Human Development*. Cambridge: Harvard University Press.

Brooks-Gunn, Jeanne (1990). "Identifying the Vulnerable Young Child." In *Improving the Life Chances of Children at Risk*, edited by David E. Rogers and Eli Ginzberg. Boulder, CO: Westview Press, 104-124.

Brooks-Gunn, Jeanne, Duncan, Greg J; Klebanov, Pamela and Sealand, Naomi (1993). "Do Neighborhoods Influence Child and Adolescent Development?" *American Journal of Sociology* 99 (2): 353-395.

Case, Anne C. and Katz, Lawrence F. (1991). "The Company You Keep: The Effects of Family and Neighborhood on Disadvantaged Youths." Working Paper no. 3705. Cambridge, Mass.: NBER.

Case, Anne C.; Lubotsky, Darren and Paxson, Christine H. (2002). "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review* 92 (5): 1308-1334.

Collins, Rebecca L. (1996). "For Better or Worse: The Impact of Upward Social Comparison on Self-Evaluations." *Psychological Bulletin* 119: 51-69.

Connell, James P. and Halpern-Felsher, Bonnie (1997). "How Neighborhoods Affect Educational Outcomes in Middle Childhood and Adolescence: Conceptual Issues and an Empirical Example." In *Neighborhood Poverty: Context and Consequences for Children, Volume 1*, edited by Jeanne Brooks-Gunn; Greg J. Duncan; and J. Lawrence Aber. New York: Russell Sage Foundation Press. 174-199.

Crane, Jonathan (1991). "The Epidemic Theory of Ghettos and Neighborhood Effect on Dropping Out and Teenage Childbearing." *American Journal of Sociology* 96 (5): 1226-1259.

Currie, Janet and Reagan, Patricia (2003). "Distance to Hospitals and Children's Access to Care: Is Being Closer Better, and for Whom?" *Economic Inquiry* 41(3), 378-391.

Currie Janet and Stabile, Mark (2003). "Socioeconomic Status and Health: Why is the Relationship Stronger for Older Children?" *American Economic Review* 93(5), 1813-1823.

Currie, Janet and Yelowitz, Aaron (2000). "Are Public Housing Projects Good for Kids?" *Journal of Public Economics* 75: 99-124.

Darling, Nancy and Steinberg, Lawrence (1997) "Assessing Neighborhood Effects Using Individual Data." In *Neighborhood Poverty: Context and Consequences for Children, Volume 1*, edited by Jeanne Brooks-Gunn; Greg J. Duncan; and J. Lawrence Aber. New York: Russell Sage Foundation Press. 120-131.

Dornbusch, Sanford M.; Ritter, Philip L.; and Steinberg, Lawrence (1991). "Community Influences on the Relations of Family Statuses to Adolescent School Performance: Differences

between African Americans and Non-Hispanic Whites." *American Journal of Education* 38: 543-567.
Downey, Geraldine and Coyne, James C. (1990). "Children of Depressed Parents: an Integrative Review." *Psychological Bulletin* 108: 50-76.
Duflo, Esther and Saez, Emmanuel (2003). "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118: 815-842.
Duncan, Greg J.; Brooks-Gunn, Jeanne, and Klebanov, Pamela (1994). "Economic Deprivation and Early-childhood Development." *Child Development* 65 (2): 296-318.
Duncan, Greg J.; Bos, Johannes; Levy, Dan; Kremer, Michael and Eccles, Jacque (2003). "Empathy or Antipathy? The Consequences of Racially and Socially Diverse Peers on Attitudes and Behaviors." Manuscript.
Duncan, Greg J. and Raudenbush, Stephen W. (2001). "Neighborhoods and Adolescent Development: How Can We Determine the Links?" In *Does it Take a Village? Community Effects on Children, Adolescents, and Families*, edited by Alan Booth and Nan Crouter. State College, PA: Pennsylvania State University Press, 105-136.
Ellen, Ingrid and Turner, Margery A. (1997). "Does Neighborhood Matter? Assessing Recent Evidence." *Housing Policy Debate* 8 (4): 833-866.
Ensminger, Margaret E.; Lamkin, Rebecca P. and Jacobson, Nora (1996). "School Leaving: A Longitudinal Perspective Including Neighborhood Effects." *Child Development* 67: 2400-2416.
Entwisle, Doris R.; Alexander, Karl L. and Olson, Linda S. (1994) "The Gender Gap in Math: Its Possible Origins in Neighborhood Effects." *American Sociological Review* 59 (6): 822-838.
Epple, Dennis; and Romano, Richard E. (1998). "Competition between Private and Public Schools, Vouchers, and Peer-Group Effects." *American Economic Review* 88 (1): 33-62.
Esbensen, Finn-age and Huizinga, David (1990). "Community Structure and Drug Use: From a Social Disorganization Perspective." *Justice Quarterly* 7 (4): 691-709.
Evans, William N.; Oates, Wallace E. and Schwab, Robert M. (1992). "Measuring Peer Group Effects: A Study of Teenage Behavior." *Journal of Political Economy* 100: 966-991.
Famularo, Richard; Fenton, Terence; Kinscherff, Robert and Augustyn, Marilyn (1996). "Psychiatric Comorbidity in Childhood Post Traumatic Stress Disorder." *Child Abuse & Neglect* 20: 953-961.
Ferguson, Ronald F. "A Diagnostic Analysis of the Black-White GPA Disparities in Shaker Heights, Ohio.: *Brookings Papers on Education Policy* 2001: 347-414.
Fernandez, Raquel (2002). "Sorting, Education, and Inequality." In *Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress,* edited by Mathias Dewatriport, Lars P. Hansen, and Steven J. Turnovsky. Cambridge: Cambridge University Press.
Follman, Dean (1996). "A Simple Multivariate Test for One-Sided Alternatives." *Journal of the American Statistical Association*, 91:434, 854-861.
Garfinkel, I.; Manski, C. F. and Michalopoulos, C. (1992). "Micro Experiments and Macro Effects." In *Evaluating Welfare and Training Programs*, edited by Charles Manski and Irwin Garfinkel. Cambridge: Harvard University Press.
Gelber, Lawrence E.; Seltzer, Leonard H.; Bouzoukis, James K.; Pollart, Susan M.; Chapman, Martin D. and Platts-Mills, Thomas (1993). "Sensitization and Exposure to Indoor Allergens

as Risk Factors for Asthma among Patients Presenting to Hospital." *American Review of Respiratory Disease* 174: 573-578.
Glass, Gene V.; McGaw, Barry and Smith, Mary Lee (1981). *Meta-analysis in Social Research*. Beverly Hills, Calif. : Sage Publications.
Glaeser, Edward; Sacerdote, Bruce and Scheinkman, Jose A. (2003). "The Social Multiplier." *Journal of the European Economic Association* 1: 345-353.
Goering, John M. and Feins, Judith D., eds. (2003). *Choosing a Better Life: Evaluating the Moving To Opportunity Experiment.* Washington DC: Urban Institute Press.
Goodman, Elizabeth; Hinden, Beth R. and Khandelwal, Seema (2000). "Accuracy of Teen and Parental Reports of Obesity and Body Mass Index." *Pediatrics* 106: 52-58.
Goodman, Elizabeth and Whitaker, Robert C. (2002). "A Prospective Study of the Role of Depression in the Development and Persistence of Adolescent Obesity." *Pediatrics* 109: 497-504.
Gould, Eric D.; Lavy, Victor and Paserman, M. Daniel (2004). "Immigrating to Opportunity: Estimating the Effect of School Quality Using a Natural Experiment on Ethiopians in Israel." *Quarterly Journal of Economics* 119(2): forthcoming.
Groves, Betsy M.; Zuckerman, Barry; Marans, Steven and Cohen, Donald J. (1993). "Silent Victims: Children Who Witness Violence." *Journal of American Medical Association* 269: 262-264.
Gruber, Jonathan (2001). "Introduction." In *Risky Behavior Among Youth: An Economic Analysis*, edited by Jonathan Gruber. Chicago: University of Chicago Press.
Halpern-Felsher, Bonnie L.; Connell, James P.; Spencer, Margaret Beale; Aber, J. Lawrence; Duncan, Greg J.; Clifford, Elizabeth; Crichlow, Warren E.; Usinger, Peter A.; Cole, Steven P.; Allen, LaRue and Seidman, Edward. (1997). "Neighborhood and family factors predicting educational risk and attainment in African-American and European-American children and adolescents." In *Neighborhood Poverty: Context and Consequences for Children, Volume 1*, edited by Jeanne Brooks-Gunn, Greg J. Duncan and J. Lawrence Aber. New York: Russell Sage Foundation Press. 146-173.
Hammen, Constance and Brennan, Patricia A. (2003). "Severity, Chronicity, and Timing of Maternal Depression and Risk for Adolescent Offspring Diagnoses in a Community Sample." *Archives of General Psychiatry* 60: 253-258.
Hauser, Robert M.; Brown, Brett V. and Prosser, William R, eds. (1997). *Indicators of Children's Well-being*. New York: Russell Sage Foundation Press.
Heckman, James J. (2001). "Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programs." *Economic Journal* 111 (475): 654-699.
Heckman, James J.; LaLonde, Robert J. and Smith, Jeffrey A. (1999). "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley Ashenfeleter and David Card. Amsterdam: North Holland, 1865-2097.
Hedges, Larry V. and Olkin, Ingram (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.
Henderson, Vernon; Mieszkowski, Peter and Sauvageau, Yvon (1978). "Peer Group Effects and Educational Production Functions." *Journal of Public Economics* 10: 97-106.
Hogan, Dennis; Astone, Nan-Marie and Kitagawa, Evelyn (1985). "Social and Environmental Factors Influencing Contraceptive Use Among Black Adolescents." *Family Planning Perspectives* 17: 165-169.

Hogan, Dennis and Kitagawa, Evelyn (1985). "The Impact of Social Status, Family Structure, and Neighborhood on the Fertility of Black Adolescents." *American Journal of Sociology* 90 (4): 825-55.

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6: 65-70.

Horowitz, Joel L. (2001). "The Bootstrap." In *Handbook of Econometrics*, *Vol. 5*, edited by James J. Heckman and Edward E. Leamer. Elsevier Science: B.V. 3159-3228.

Hoxby, Caroline M. (2001). "Peer Effects in the Classroom: Learning from Gender and Race Variation." Working Paper no. 7867. Cambridge, Mass.: NBER.

Idler, Ellen L. and Kasl, Stanislav V. (1995). "Self-Ratings of Health: Do They Also Predict Changes in Functionality." *Journal of Gerontology: Social Sciences* 50 (6): S344-353.

Jacob, Brian A. (2004) "Public Housing, Housing Voucher, and Student Achievment: Evidence from Public Housing Demolitions in Chicago." *American Economic Review* 94 (1): 233-258.

Jargowsky, Paul A. *Poverty and place : ghettos, barrios, and the American city.* New York: Russell Sage Foundation, 1997.

Jencks, Christopher and Mayer, Susan (1990). The Social Consequences of Growing up in a Poor Neighborhood." In *Inner-city Poverty in the United States*, edited by Lawrence E. Lynn and Michael G. H. McGeary. Washington, DC: National Academy Press. 111-186.

Kane, Thomas J. (2004). "The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations." Working paper. William T. Grant Foundation.

Katz, Lawrence F.; Kling, Jeffrey R. and Liebman, Jeffrey B. (2001). "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment." *Quarterly Journal of Economics* 116 (2): 607-654.

Kessler, Ronald C.; Andrews, G.; Colpe, Lisa J.; Hiripi, Eva; Mroczek, Daniel K.; Normand, Sharon-Lise T.; Walters, Ellen E. and Zaslavsky, Alan M. (2002). "Short Screening Scales to Monitor Population Prevalances and Trends in Nonspecific Psychological Distress." *Psychological Medicine* 32 (6): 959-976.

Kling, Jeffrey R.; Liebman, Jeffrey B.; Katz, Lawrence F. and Sanbonmatsu, Lisa. (2004). "Moving to Opportunity and Tranquility: Neighborhood Effects on Adult Economic Self-sufficiency and Health in a Randomized Housing Voucher Experiment." Princeton University IRS Working Paper 481.

Kling, Jeffrey R.; Liebman, Jeffrey B. and Katz, Lawrence F. (2004). "Bullets Don't Got No Name: Consequences of Fear in the Ghetto." In *Discovering Successful Pathways in Children's Development: New Methods in the Study of Childhood and Family Life*, edited by Thomas S. Eisner. Chicago: The University of Chicago Press.

Kling, Jeffrey R.; Ludwig, Jens and Katz, Lawrence F. (2004). "Youth Criminal Behavior in the Moving to Opportunity Experiment." Princeton University IRS Working Paper 482.

Kremer, Michael and Levy, Dan (2003). "Peer Effects and Alcohol Use Among College Students." Cambridge, MA: NBER Working Paper No. 9876, July.

Krosnick, Jon A. and Judd, Charles M. (1982). "Changes in Social Influence at Adolescence: Who Induces Cigarette Smoking?" *Developmental Psychology* 18: 359-368.

Krueger, Alan B. (1999). "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114:497-532.

Lazear, Edward P. (1999). "Culture and Language." *Journal of Political Economy* 107: S95-S126.

Lee, Rachel E. and Cubbin, Catherine (2002). "Neighborhood Context and Youth Cardiovascular Health Behaviors." *American Journal of Public Health* 92: 428-436.

Leventhal, Tama and Brooks-Gunn, Jeanne (2000). "The Neighborhoods They Live In: The Effects of Neighborhood Residence on Child and Adolescent Outcomes." *Psychological Bulletin* 126 (2): 309-337.

Leventhal, Tama and Brooks-Gunn, Jeanne (2003a). "The Early Impacts of Moving to Opportunity on Children and Youth." In *Choosing a Better Life: Evaluating the Moving To Opportunity Experiment*, edited by John Goering and Judith D. Feins. Washington DC: Urban Institute Press.

Leventhal, Tama and Brooks-Gunn, Jeanne (2003b). "A Randomized Study of Neighborhood Effects on Low-Income Children's Educational Outcomes." *Developmental Psychology,* forthcoming.

Liebman, Jeffrey B.; Katz, Lawrence F. and Kling, Jeffrey R. (2003). "Are Neighborhood Effects Nonlinear? Estimates from the MTO Experiment." Manuscript, Harvard University.

Logan, Brent R. and Tamhane, Ajit C. (2003). "On O'Brien's OLS and GLS Tests for Multiple Endpoints." Manuscript, Northwestern University.

Ludwig, Jens; Duncan, Greg J. and Hirschfield, Paul (2001). "Urban Poverty and Juvenile Crime: Evidence from a Randomized Housing-mobility Experiment." *Quarterly Journal of Economics* 116 (2): 655-679.

Ludwig, Jens; Ladd, Helen F. and Duncan, Greg J. (2001). "Urban Poverty and Educational Outcomes." *Brooking Papers on Urban Affairs 2001*. 147-197.

Manski, Charles F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60: 531-542.

Manski, Charles F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

Manski, Charles F. (2000). "Economic Analysis of Social Interactions." *Journal of Economic Perspectives* 14: 115-136.

Manski, Charles F. and Pepper, John V. (2000). "Monotone Instrumental Variables: with an Application to the Returns to Schooling." *Econometrica* 68 (4): 997-1010.

Marsh, Herbert W. and Parker, John W. (1984). "Determinants of Student Self-concept: is it Better to be a Relatively Large Fish in a Small Pond Even if You Don't Learn to Swim as Well? *Journal of Personality and Social Psychology* 47: 213-231.

Mayer, Susan E. and Jencks, C. (1989). "Growing Up in Poor Neighborhoods: How Much Does it Matter?" *Science* 243: 1441-1445.

McCormick, Marie C. and Brooks-Gunn, Jeanne (1989). "The Health of Children and Adolescents." In *Handbook of Medical Sociology*, edited by Howard E. Freeman and Sol Levine. Englewood Cliffs, NJ: Prentice Hall. 347-380.

McCleod Jane D. and Shanahan, Michael J. (1993). "Poverty, Parenting, and Children's Mental Health." *American Sociological Review* 58: 351-366.

Meisels, Samuel J. and Shonkoff, Jack P., eds. (1990). *Handbook of Early Childhood Intervention*. New York: Cambridge University Press.

Meyers, Alan; Frank, Deborah A.; Roose, Nicole; Peterson, Karen E.; Casey, Virginia A.; Cupples, Adrienne and Levenson, Suzette H. (1995). "Housing Subsidies and Pediatric Undernutrition." *Archives of Pediatric and Adolescent Medicine* 149: 1079-1084.

Miguel, Edward and Kremer, Michael. (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72 (1): 159-217.

Moffitt, Robert A. (2001). "Policy Interventions, Low-Level Equilibria, and Social Interactions." In *Social Dynamics*, edited by Steven N. Durlauf and H. Peyton Young. Cambridge: MIT Press, 45-82.

Norton, Edward C.; Lindrooth, Richard C. and Ennett, Susan T. (1998). "Controlling for the Endogeneity of Peer Substance Use on Adolescent Alcohol and Tobacco Use." *Health Economics* 7: 439-53.

O'Brien, Peter C. (1984). "Procedures for Comparing Samples with Multiple Endpoints." *Biometrics* 40: 1079-1087.

O'Regan, Katherine M. and Quigley, John M. (1996). "Teenage Employment and the Spatial Isolation of Minority and Poverty Households," *Journal of Human Resources* 31 (3): 692-702.

Oreopoulos, Philip (2003). "The Long-run Consequences of Living in a Poor Neighborhood." *Quarterly Journal of Economics* 118 (4): 1533-1575.

Orr, Larry.; Feins, Judith D.; Jacob, Robin; Beecroft, Eric; Sanbonmatsu, Lisa; Katz, Lawrence F.; Liebman, Jeffrey B. and Kling, Jeffrey R. (2003). *Moving To Opportunity Interim Impacts Evaluation*. Washington DC: U. S. Department of Housing and Urban Development.

Pearce, Neal; Beasley, Richard; Burgess, Carl and Crane, Julian (1998). *Asthma Epidemiology: Principles and Methods*. Oxford: Oxford University Press.

Quinlan, Kyran P. (1996). "Injury Control in Practice." *Archives of Pediatrics Adolescent Medicine* 150: 954-957.

Ramirez-Valles, Jesus; Zimmerman, Marc A. and Juarez, Lucia (2002). "Gender Differences of Neighborhood and Social Control Process: A Study of the Timing of First Intercourse among Low-achieving, Urban, African American youth." *Youth and Society* 33 (3): 418-442.

Richter, Kimber P.; Harris Karl-Jo; Paine-Andrews, Adrienne, Fawcett, Stephen B.; Schmid, Thomas L.; Lankenau, Becky, H. and Johnston, Judy (2000). "Measuring the Health Environment for Physical Activity and Nutrition among Youth: A Review of the Literature and Applications for Community Initiatives." *Preventive Medicine* 31: S98-S111.

Rockoff, Jonah E. (2003). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." Manuscript, Harvard University.

Rosenbaum, James E. (1992). "Black Pioneers: Do Their Moves to the Suburbs Increase Economic Opportunity for the Mothers and Children?" *Housing Policy Debate* 2 (4): 1179-1213.

Rosenthal, Robert (2000). "Effect Sizes in Behavioral and Biomedical Research." In *Validity and Social Experimentation*, edited by Leonard Bickman. Thousand Oaks, CA: Sage Publications. 121-139.

Rowe, David C. (1994). *The Limits of Family Influence: Genes, Expression, and Behavior*. New York: Guilford Press.

Sacerdote, Bruce (2001). "Peer Effect With Random Assignment: Results for Dartmouth Roomates." *Quarterly Journal of Economics* 116 (2): 681-704.

Sampson, Robert J. and Groves, W. Byron (1989). "Community Structure and Crime: Testing Social-disorganization Theory." *American Journal of Sociology* 94(4): 774-780.

Sampson, Robert J.; Morenoff, Jeffrey D. and Gannon-Rowley, Thomas (2002). "Assessing Neighborhood Effects: Social Processes and New Directions in Research." *Annual Review of Sociology* 28: 443-78.

Sampson, Robert J.; Raudenbush, Stephen W. and Earls, Felton (1997). "Neighborhoods and Violent Crime: A Multi-level Study of Collective Efficacy." *Science* 277 (August 15): 918-924.

Savin, Nathan E. (1984). "Multiple Hypothesis Testing." In *Handbook of Econometrics, Volume II*, edited by Zvi Griliches and Michael D. Intriligator. Amsterdam: Elsevier Science Publishers BV, 827-879.

Scarr, Sandra (1992). "Developmental Theories for the 1990s: Development and Individual Differences." *Developmental Psychology* 63: 1-19.

Scharfstein, Joshua and Sandel, Megan, eds. (1998). "Not Safe at Home: How America's Housing Crisis Threatens the Health of Its Children." The Doc4Kids Project. Boston Medical Center: Department of Pediatrics.

Solon Gary; Page, Marianne E. and Duncan, Greg J. (2000). Correlations Between Neighboring Children in Their Subsequent Educational Attainment. *Review of Economics and Statistics* 82: 383-392.

Substance Abuse and Mental Health Services Administration. "Results from the 2002 National Survey on Drug Use and Health: Detailed Tables." Washington, DC: Department of Health and Human Services.

Tamhane, Ajit C. and Logan, Brent R. (2003). "Multiple Endpoints: An Overview and New Developments." Technical Report 43, Division of Biostatistics, Medical College of Wisconsin.

Taylor Laura; Zuckerman, Barry; Harik, Vaira and Groves, Betsy, M. (1994). "Witnessing Violence by Young Children and Their Mothers." *Developmental and Behavioral Pediatrics* 15: 120-123.

Thurston, George D. and Bates, David V. (2003). "Asthma as an Underappreciated Cause of Asthma Symptoms." *Journal of the American Medical Association.* 290(14): 1915-1916.

Weiss, Kevin B.; Gergen, Peter and Crain, Ellen (1992). "Inner-city Asthma: The Epidemiology of an Emerging US Public Health Concern." *Chest* 101: 362-367.

Weissman, Myrna M.; Warner, Virginia; Wickramaratne, Priya; Moreau, Donna and Olfso, Mark (1997). "Offspring of Depressed Parents: 10 Years Later." *Archives of General Psychiatry* 54: 932-940.

Weitzman, Michael, Gortmaker, Steven and Sobol, Arthur (1990). "Racial, Social, and Environmental Risks for Childhood Asthma." *American Journal of Diseases of Children* 144: 1189-1194.

Westfall, Peter H., and Young, S. Stanley (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment.* New York: Wiley.

Wilson, William J. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: University of Chicago Press.

Wood, Joanne V. (1989). "Theory and Research Concerning Social Comparisons and Personal Attributes." *Psychological Bulletin* 106: 231-248.

Wright, Rosalind J.; Rodriguez, Mario and Cohen, Sheldon (1998). "Review of Psychosocial Stress and Asthma: An Integrated Biopsychosocial Approach." *Thorax* 53: 1066-1074.

Zimmerman, David J. (2003). "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *Review of Economics and Statistics* 85, 9-23.

TABLE 1. BASELINE CHARACTERISTICS

	Female			Male		
	Exp (1)	Sec8 (2)	Con (3)	Exp (4)	Sec8 (5)	Con (6)
African-American	.68	.64	.67	.64	.65	.59
Special class for gifted students or did advanced work	.15	.17	.17	.17*	.15*	.27
Special school, class, or help for learning problem in past two years	.13	.13	.12	.29	.25	.30
Special school, class, or help for behavioral or emotional problems in past two years	.07	.08	.05	.18	.17	.11
Problems that made it difficult to get to school and/or to play active games	.03	.06	.06	.11*	.08	.05
Problems that required special medicine and/or equipment	.05	.07	.05	.13	.14	.09
School asked to talk about problems child having with schoolwork or behavior in past two years	.19	.23	.19	.41	.37	.33
Suspended or expelled from school in past two years	.09	.10	.07	.23	.20	.15

Notes. Exp: Experimental. Sec8: Section 8. Con: Control. * indicates p-value <.05 on difference between experimental or Section 8 and control group. Baseline data was collected at random assignment, during 1994-1997. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001 for a total sample size of 1807. Covariates used in equations (4) and (5) include those in Table 1 and four indicators for site, six Legendre polynomials for date of birth, and five indicators for missing data on: special class for gifted students or did advanced work; special school, class or help for learning problem in past two years; special school, class or help for behavioral or emotional problems in past two years; problems that made it difficult for him/her to get to school and/or to play active games or sports; suspended or expelled from school in past two years. All statistical estimates use survey weights, which have three components, described in detail in Orr et al. (2003), Appendix B. First, subsample members receive greater weight since, in addition to themselves, they represent individuals whom we did not attempt to contact during the subsampling phase. Second, youth from large families receive greater weight since we randomly sampled two children per household implying that youth from large families are representative of a larger fraction of the study population. Third, all individuals are weighted by the inverse of their probability of assignment to their experimental group to account for changes in the random assignment ratios over time. The ratio of individuals randomly assigned to treatment groups was changed during the course of the demonstration to minimize the minimum detectable effects after take-up of the vouchers turned out to be much higher than had been projected. This last component of the weights is, therefore, necessary to prevent time or cohort effects from confounding the results. Our weights imply that each random assignment period is weighted in proportion to the number of people randomly assigned in that period.

TABLE 2. SHARE OF CURRENT LOCATION IN 2000 CENSUS TRACT POVERTY RATE CATEGORIES

	Poverty Rate (percent)				
	0-12 (1)	12-24 (2)	24-36 (3)	36-48 (4)	48+ (5)
Experimental -- overall	.17	.25	.19	.21	.19
Section 8 -- overall	.08	.26	.23	.27	.16
Control -- overall and no subsidized move through MTO	.04	.17	.20	.27	.32
Experimental -- 57% of group with no subsidized move through MTO	.06	.14	.22	.28	.30
Section 8 -- 45% of group with no subsidized move through MTO	.07	.18	.17	.31	.26
Section 8 -- 55% of group with subsidized move through MTO	.08	.33	.28	.23	.08
Experimental -- 43% of group with subsidized move through MTO	.31	.39	.15	.12	.04

Notes. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001.

TABLE 3. EDUCATION OUTCOMES

Outcome	Females				Males				M-F TOT
	CM (1)	ITT (2)	TOT (3)	CCM (4)	CM (5)	ITT (6)	TOT (7)	CCM (8)	(9)
A. Experimental - Control									
Graduated HS or still in school	.772	.060	.117	.710	.759	-.036	-.091	.820	-.208
		(.037)	(.072)			(.038)	(.098)		(.118)
In school or working	.771	.058	.120	.757	.758	.022	.053	.722	-.067
		(.036)	(.074)			(.037)	(.089)		(.115)
WJ-R reading test z-score	.059	.091	.190	-.131	-.110	.001	.002	-.173	-.188
		(.086)	(.181)			(.107)	(.263)		(.313)
WJ-R math test z-score	.005	.117	.243	-.219	-.042	-.063	-.156	-.064	-.399
		(.092)	(.193)			(.105)	(.260)		(.316)
B. Section 8 - Control									
Graduated HS or still in school	.772	.042	.074	.752	.759	-.016	-.031	.762	-.105
		(.037)	(.066)			(.040)	(.079)		(.102)
In school or working	.771	-.022	-.038	.775	.758	-.010	-.020	.705	.018
		(.038)	(.065)			(.041)	(.080)		(.102)
WJ-R reading test z-score	.059	.059	.105	.006	-.110	.136	.264	-.346	.159
		(.090)	(.160)			(.112)	(.221)		(.270)
WJ-R math test z-score	.005	.087	.157	-.017	-.042	.059	.114	-.202	-.043
		(.095)	(.170)			(.116)	(.226)		(.281)

Notes. CM: Control Mean. ITT: Intent-to-treat. TOT: Treatment-on-treated. CCM: Control complier mean. M-F TOT: TOT for males - TOT for females. Robust standard errors adjusted for household clustering are in parentheses. P-value denoted by * is <.05. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001. ITT estimates in columns 2 and 6 for each row are from separate models estimated using equation (4), using OLS to control for variables discussed in Table 1. TOT estimates in columns 3 and 7 are from separate models estimated using equation (5), using 2SLS to control for the same baseline characteristics as with OLS. CCM is from equation (6). Graduated HS or still in school is based on parental report; others from youth interviews.

TABLE 4. RISKY BEHAVIOR OUTCOMES

Outcome	Females				Males				M-F TOT
	CM (1)	ITT (2)	TOT (3)	CCM (4)	CM (5)	ITT (6)	TOT (7)	CCM (8)	(9)
A. Experimental - Control									
Used marijuana in past 30 days	.131	-.059* (.028)	-.121* (.058)	.193	.118	.053 (.030)	.131 (.074)	.045	.252* (.094)
Smoked in past 30 days	.191	-.062 (.033)	-.128 (.069)	.252	.125	.109* (.034)	.267* (.086)	0	.395* (.110)
Had alcohol in past 30 days	.206	-.064 (.035)	-.134 (.075)	.305	.140	.067* (.033)	.162* (.081)	.069	.295* (.110)
Ever pregnant or gotten someone pregnant	.267	-.025 (.039)	-.052 (.081)	.228	.119	.030 (.031)	.071 (.076)	.069	.123 (.110)
B. Section 8 - Control									
Used marijuana in past 30 days	.131	-.052 (.030)	-.091 (.053)	.176	.118	.075* (.035)	.142* (.068)	.055	.233* (.085)
Smoked in past 30 days	.191	-.054 (.034)	-.094 (.060)	.243	.125	.146* (.040)	.279* (.079)	.027	.373* (.100)
Had alcohol in past 30 days	.206	-.073 (.037)	-.125 (.065)	.276	.140	.079* (.036)	.152* (.071)	.070	.277* (.097)
Ever pregnant or gotten someone pregnant	.267	.060 (.042)	.104 (.075)	.234	.119	.032 (.036)	.061 (.068)	.143	-.043 (.099)

Notes. CM: Control Mean. ITT: Intent-to-treat. TOT: Treatment-on-treated. CCM: Control complier mean. M-F TOT: TOT for males - TOT for females. Robust standard errors adjusted for household clustering are in parentheses. P-value denoted by * is <.05. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001. ITT estimates in columns 2 and 6 for each row are from separate models estimated using equation (4), using OLS to control for variables discussed in Table 1. TOT estimates in columns 3 and 7 are from separate models estimated using equation (5), using 2SLS to control for the same baseline characteristics as with OLS. CCM is from equation (6). Outcomes from youth self-reports.

TABLE 5. MENTAL HEALTH OUTCOMES

Outcome	Females				Males				
	CM (1)	ITT (2)	TOT (3)	CCM (4)	CM (5)	ITT (6)	TOT (7)	CCM (8)	M-F TOT (9)
A. Experimental - Control									
Psychological distress -- K6 scale z-score	.268	-.246*	-.508*	.556	-.162	.069	.167	-.251	.675*
		(.091)	(.195)			(.091)	(.223)		(.293)
Ever had serious depression symptoms	.137	-.049	-.099	.187	.031	.015	.036	0	.135
		(.029)	(.060)			(.022)	(.054)		(.082)
Ever had generalized anxiety symptoms	.121	-.064*	-.131*	.199	.055	-.006	-.015	.076	.116
		(.025)	(.053)			(.026)	(.065)		(.084)
B. Section 8 – Control									
Psychological distress -- K6 scale, z-score	.268	-.133	-.232	.398	-.162	-.027	-.053	-.065	.179
		(.104)	(.183)			(.096)	(.187)		(.258)
Ever had serious depression symptoms	.137	-.067*	-.114*	.195	.031	.001	.002	.045	.116
		(.029)	(.050)			(.022)	(.042)		(.066)
Ever had generalized anxiety symptoms	.121	-.071*	-.125*	.162	.055	-.032	-.063	.091	.062
		(.026)	(.046)			(.025)	(.049)		(.068)

Notes. CM: Control Mean. ITT: Intent-to-treat. TOT: Treatment-on-treated. CCM: Control complier mean. M-F TOT: TOT for males - TOT for females. Robust standard errors adjusted for household clustering are in parentheses. P-value denoted by * is <.05. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001. ITT estimates in columns 2 and 6 for each row are from separate models estimated using equation (4), using OLS to control for variables discussed in Table 1. TOT estimates in columns 3 and 7 are from separate models estimated using equation (5), using 2SLS to control for the same baseline characteristics as with OLS. CCM is from equation (6). Outcomes from youth self-reports.

TABLE 6. PHYSICAL HEALTH OUTCOMES

Outcome	Females				Males				M-F TOT
	CM (1)	ITT (2)	TOT (3)	CCM (4)	CM (5)	ITT (6)	TOT (7)	CCM (8)	(9)
A. Experimental - Control									
Has fair or poor health	.101	.021	.044	.090	.045	.033	.079	.023	.035
		(.027)	(.056)			(.019)	(.046)		(.072)
Asthma attack or wheezing in past year	.201	.011	.022	.196	.122	.021	.050	.122	.028
		(.037)	(.076)			(.030)	(.074)		(.107)
Serious non-sports accident or injury in past year	.115	-.019	-.039	.169	.062	.085*	.206*	0	.246*
		(.025)	(.051)			(.026)	(.062)		(.082)
Body Mass Index > 95th percentile	.173	-.012	-.027	.160	.161	.023	.057	.132	.083
		(.034)	(.074)			(.037)	(.089)		(.114)
B. Section 8 - Control									
Has fair or poor health	.101	-.003	-.005	.108	.045	.033	.065	.028	.070
		(.027)	(.046)			(.023)	(.045)		(.064)
Asthma attack or wheezing in past year	.201	-.026	-.045	.244	.122	.045	.086	.085	.132
		(.037)	(.064)			(.036)	(.069)		(.096)
Serious non-sports accident or injury in past year	.115	-.038	-.065	.139	.062	.081*	.156*	0	.222*
		(.026)	(.045)			(.028)	(.057)		(.073)
Body Mass Index > 95th percentile	.173	-.012	-.020	.177	.161	-.012	-.022	.181	-.002
		(.039)	(.067)			(.037)	(.072)		(.098)

Notes. CM: Control Mean. ITT: Intent-to-treat. TOT: Treatment-on-treated. CCM: Control complier mean. M-F TOT: TOT for males - TOT for females. Robust standard errors adjusted for household clustering are in parentheses. P-value denoted by * is <.05. Surveys were completed in experimental, Section 8 and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001. ITT estimates in columns 2 and 6 for each row are from separate models estimated using equation (4), using OLS to control for variables discussed in Table 1. TOT estimates in columns 3 and 7 are from separate models estimated using equation (5), using 2SLS to control for the same baseline characteristics as with OLS. CCM is from equation (6). Outcomes from youth self-reports.

TABLE 7. MEAN EFFECT SIZES FOR OUTCOMES

	Females		Males		Males-Females	
	E-C	S-C	E-C	S-C	E-C	S-C
Education (4 measures)	.129 (.066)	.053 (.065)	-.021 (.051)	.026 (.059)	-.150 (.082)	-.027 (.087)
Absence of Risky Behavior (4 measures)	.137* (.051)	.084 (.054)	-.194* (.063)	-.250* (.072)	-.331* (.080)	-.334* (.090)
Absence of Mental Health Problems (3 measures)	.188* (.062)	.179* (.062)	-.046 (.085)	.056 (.087)	-.234* (.106)	-.123 (.107)
Absence of Physical Health Problems (4 measures)	-.001 (.050)	.056 (.051)	-.159* (.053)	-.150* (.059)	-.158* (.073)	-.206* (.079)
Overall (15 measures)	.109* (.034)	.089* (.033)	-.109* (.035)	-.088* (.039)	-.218* (.049)	-.177* (.052)

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. * denotes p-value < .05. Absence of risky behavior, mental health problems, and physical health problems are defined so that positive signs represent beneficial effects. As described in section II, mean effect sizes are the mean of the intent-to-treat estimates in Tables 3-6, where each estimate was normalized by the gender-specific standard deviation of the outcome in the control group from equation (7). Standard errors account for correlation between measures, based on equation (10).

TABLE 8. P-VALUES FOR SELECTED DOMAINS

Domain	Gender	Contrast	Per-Comparison p-values			Familywise adjusted p-values		
			Asymptotic (1)	Asymptotic w/δ method (2)	Bootstrap (3)	Bonferroni & Asymptotic (4)	Bonferroni & Bootstrap (5)	Westfall-Young (6)
Overall	Females	E-C	.0013	.0007	.0012	.039	.035	.029
	Males	E-C	.0019	.0031	.0039	.052	.100	.076
	Females	S-C	.0074	.0049	.0082	.171	.181	.132
	Males	S-C	.0239	.0289	.0413	.455	.743	.424
Risky Behavior	Females	E-C	.0078	.0036	.0067	.173	.160	.118
	Males	E-C	.0020	.0043	.0063	.054	.156	.116
	Males	S-C	.0006	.0015	.0029	.018	.079	.062
Mental Health	Females	E-C	.0023	.0008	.0011	.060	.033	.027
	Females	S-C	.0038	.0011	.0007	.090	.022	.019
Physical Health	Males	E-C	.0028	.0074	.0109	.070	.230	.164
	Males	S-C	.0118	.0210	.0273	.248	.519	.322

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. Results shown only for contrasts from Table 7 with per-comparison p-values less than .05. Column 1 shows p-values based on Table 7, using the asymptotic standard error calculation in equation (10). Column 2 uses point estimates from Table 7 and standard errors based on equation (11). Column 3 uses bootstrap standard errors calculated as described in Appendix A. Columns 4-6 are adjusted p-values, with the family defined as five domains (overall, education, risky behavior, mental health, physical health), three subgroups (pooled gender, females, males), and two contrasts (E-C, S-C), for a total of 30 estimates. Column 4 is based on column 1, using the Bonferroni-Holm adjustment described in section II. Column 5 is a Bonferroni-Holm adjustment to column 3. Column 6 is based on equation (12), using the Westfall-Young bootstrap adjustment described in Appendix A with 10,000 bootstrap replications.

TABLE 9. MEAN EFFECT SIZES FOR MEDIATORS

	Females		Males		Males-Females	
	E-C	S-C	E-C	S-C	E-C	S-C
General Neighborhood (11 measures)	.308* (.051)	.223* (.053)	.266* (.055)	.183* (.053)	-.043 (.072)	-.040 (.072)
Absence of Victimization (3 measures)	.125 (.067)	.127 (.066)	.040 (.064)	.081 (.066)	-.085 (.091)	-.046 (.094)
Absence of Housing Problems (4 measures)	.256* (.067)	.097 (.079)	.141 (.078)	.132 (.083)	-.115 (.103)	.034 (.114)
Parenting Practices (5 measures)	.046 (.048)	.019 (.054)	-.066 (.050)	-.057 (.054)	-.112 (.070)	-.077 (.077)
School Environment (6 measures)	.116* (.056)	.019 (.059)	.156* (.061)	.117 (.062)	.040 (.083)	.098 (.085)
Peers (8 measures)	.057 (.036)	.015 (.039)	-.019 (.038)	.027 (.040)	-.076 (.052)	.012 (.055)
Adult Role Models (8 measures)	.184* (.045)	.090* (.045)	.046 (.044)	.019 (.053)	-.137* (.062)	-.072 (.069)
School Engagement (6 measures)	.094 (.063)	.015 (.067)	-.018 (.061)	.006 (.068)	-.112 (.085)	-.009 (.094)
Educational Track (4 measures)	.030 (.057)	.018 (.059)	-.081 (.054)	-.003 (.056)	-.111 (.079)	-.021 (.082)
Absence of Educational Problems (5 measures)	.012 (.052)	-.007 (.061)	.019 (.057)	.059 (.062)	.007 (.077)	.066 (.086)
Future Expectations (2 measures)	.170* (.077)	.058 (.090)	-.026 (.082)	-.021 (.085)	-.196 (.111)	-.079 (.124)
Healthy Environment (5 measures)	.125* (.051)	-.018 (.052)	.078 (.049)	.089 (.058)	-.047 (.071)	.107 (.078)
Access To Health Care (2 measures)	.053 (.067)	.024 (.076)	.033 (.071)	-.085 (.083)	-.020 (.096)	-.109 (.108)
Adult mental health (5 measures)	-.181* (.072)	-.200* (.074)	-.028 (.092)	.092 (.081)	.153 (.102)	.292* (.110)

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. * denotes p-value < .05. Estimates are the mean of intent-to-treat estimates in Tables A3-A7, normalized by the gender-specific standard deviation of the outcome in the control group from equation (7). Standard errors account for correlation among measures, based on equation (10). Positive signs represent effects hypothesized to be beneficial to outcomes, so the following signs from tables A3 to A7 have been reversed: general neighborhood (all except pro/mgmt occupations), victimization (all), housing (all), parenting (no adult present), school environment (free lunch, limited English, pupil teacher ratio), peers (friends use drugs/weapons, friends in gangs, old friends visit, visits old neighborhood), educational track (special education), educational problems (all), healthy environment (asthma triggers), adult mental health (distress, depression, worrying).

Appendix A. Westfall-Young familywise adjusted p-value bootstrap algorithm

This appendix describes our algorithm for calculating adjusted p-values. It is based on the Westfall-Young (1993, algorithm 2.8) free step-down resampling method, modified to utilize per-comparison p-values based on bootstrap estimates instead of asymptotic approximations.

For each parameter of interest, τ_j, define $\hat{\tau}_j$ as the estimated value from the actual data and $p^c{}_j$ as the asymptotic per-comparison p-value on the test of the null hypothesis that τ_j equals zero. Define N as the number of bootstrap replications. The per-comparison bootstrap p-value for τ_j is $p^b{}_j$, and the Westfall-Young familywise adjusted p-value for τ_j is $p^a{}_j$.

/* Calculate bootstrap p-values ($p^b{}_j$) */
For j = 1 to J {
    $p^a{}_j = p^b{}_j = 0$
}
For i = 1 to N {
    Draw a sample of households with replacement.
    For j = 1 to J {
        Calculate τ^*_{ij}, the estimated value of τ^*_j for this bootstrap replication.
        Calculate the p-value r_{ij} for the test that $\tau^*_{ij} = \hat{\tau}_j$.
        If $r_{ij} < p^c{}_j$, then $p^b{}_j = p^b{}_j + 1/N$
    }
}

/* Calculate p-values for each replication under null hypothesis (s_{ij}), ordering by r_{ij}
    and imposing uniform p-value distribution across replications
    for each of J parameters */
Define r_j as a vector of length N with elements r_{ij}
For j = 1 to J {
    Sort elements of r_j so r_{kj} is smallest value of r_j when k is 1
    For k = 1 to N {
        $s_{kj} = (k-.5)/N$
    }
}

/* Calculate adjusted p-value (p^a_j) */

For the J parameters in the family of tests, sort p^b_j such that j indexes family members in descending order of significance, so p^b_1 is the smallest bootstrap p-value.

For k = 1 to N {

$q_{J+1} = 1$

For j = J to 1 {

$q_j = \min(s_{kj} , q_{j+1})$

If $q_j < p^b_j$, then $p^a_j = p^a_j + 1/N$

}

}

/* Enforce monotonicity so that the order of outcomes according to bootstrap per-comparison p-values is weakly preserved according to adjusted p-values */

$p^a_0 = 0$

For j = 1 to J {

$p^a_j = \max(p^a_{j-1} , p^a_j)$

}

Appendix B. Achievement Test Scoring

The test scores used in this paper have been adjusted for potential interviewer effects. The Woodcock Johnson tests used in the MTO study indicate the level of achievement within a very wide range, as opposed to many tests in school which test proficiency at a particular threshold appropriate for specific grade levels. These same tests have also been used in other large social science studies, such as the Panel Study of Income Dynamics Child Development Supplement, the Los Angeles Family and Neighborhoods Study, and the Welfare, Children and Families Three City Study. In order to adapt to the achievement level of each individual and to avoid confounding reading skills with other skills, the tests involve considerable interaction between the sample youth and the interviewer conducting the test.

There are two subtests on which the Broad Reading (BR) score is based, Letter-Word (LW) and Passage Comprehension (PC), and two on which the Broad Math (BM) score is based, Applied Problems (AP) and Calculation (CA). Since the tests contain items simple enough for five year olds, the youth studied in this paper started in the middle of most tests (LW: item 18; PC: item 13; AP item 20; CA item 1). If the youth did not begin with six in a row correct, easier items were then asked to establish a "basal" level of performance. The test score is based on the number of correct answers, imputing correct answers for all items below the basal. The items increase in difficulty until the youth gives six incorrect responses in a row, which establishes a "ceiling" of performance, at which point the test is stopped. Thus, while administering the test, the interviewer must score the items.

There are two types of interviewer effects that we suspect are most likely. First, interviewers read items out loud during the test (LW: 1-5 but not 6-57; PC: items 1-30 but not 31-43; AP: all items 1-60; CA: none of 1-58), and the reading and pronunciation skills of the interviewers varied. Second, there is some interviewer judgment required in scoring. For example, many LW items ask for pronunciation of words, such as "sufficient," but correct pronunciation is subject to interpretation. One PC item is: "A good composition has an interesting introduction and a strong conclusion. The body is ____ the beginning and the end." Correct answers are "between" or "in between" and examples of incorrect answers are "interesting, supporting, both." While the interviewers were instructed in training to only score the item as correct if the youth said "between" or "in between", it may have been the case that some interviewers were more inclusive and marked items as correct if the response seemed correct to them. The CA test consists of math problems in a workbook and involves little interviewer interaction or judgment.

We have several pieces of statistical evidence suggesting that some interviewers may systematically score respondents higher and some lower on the reading and math tests. There is wide dispersion in the mean scores by interviewer, though interpretation of this statistic is confounded with the systematic assignment of interviewers to neighborhoods that they were most familiar with and where they could best locate sample members. Our main analysis of interviewer effects relies on differences between the test scores of different interviewers who tested sample members in the same census tract. In order to examine a sufficient number of tests per interviewer, we pool data for all 5239 children ages 5-20 who were tested in the MTO study. Specifically, we analyze the regressions of test scores on interviewer indicator variables conditional on census tract fixed effects and on individual child characteristics. Interviewer coefficients are estimated relative to the interviewer conducting the largest number of tests in each of the five main MTO sites. Essentially, we assume that interviewers are as good as randomly assigned to children within census tracts, even though they are systematically assigned to children across tracts.

Simple summary measures of potential interviewer effects are the F-statistics on the 125 interviewer coefficients, which are: BR=3.0; BM=3.3; LW=2.4; PC=4.9; AP=4.3; CA=2.3. All p-values are less than .001. This pattern is consistent with a hypothesis that the CA test (requiring the least interviewer reading or judgment) would have the smaller interview effects -- although its p-value is also less than .001. Looking at the individual t-statistics on the test that particular interviewer coefficients differ from the mean for their site, the number of interviewers with |t| greater than 2.6 on each test was: BR=1; BM=1; LW=1; PC=5; AP=2; CA=0. While we might expect one or two to be this large in a sample of 125, the binomial probability of 5 independent events with t greater than 2.6 is less than .01. Bonferroni-adjusted p-values for the most extreme interviewer, who reported systematically high scores, are .03 on BM, and less than .01 on PC and AP.

We conducted a similar set of analyses on the PC-LW and AP-CA differences, testing the hypothesis that interviewers should not have a systematic relationship to differences on two tests of reading or two tests of math for the same individual. The F-statistics on the 125 interviewer coefficients are: PC-LW=3.8, AP-CA=3.1. An extreme interviewer (who reported the largest differences) had Bonferroni-adjusted p-values of .00006 on PC-LW and .01 on AP-CA.

Based on this evidence, we conclude there is a reasonable possibility that scores may have differed systematically by interviewer. Although all interviewers conducted interviews with sample members of all three treatment groups, the proportions differed. The geographical

mix also differed, with some interviewers mainly interviewing experimental and Section 8 group non-compliers still living in inner-city areas in addition to controls.

In order to investigate the sensitivity of the treatment results to potential test score effects, we computed an adjusted test score. Using the logic described above, we estimated the interviewer coefficient conditional on census tract fixed effects and individual characteristics, and calculated the estimated interviewer effect as the deviation of the interviewer from the site mean. The adjusted score is simply the unadjusted W-score minus the estimated interviewer effect, then rescaled by its standard deviation.

The adjusted BR and BM results are given in Table 3. The ITT estimates [and p-values] for the experimental group females using adjusted data are: BR=.091 [.29]; BM=.117 [.21]. The corresponding estimates using unadjusted data are: BR=.123 [.15]; BM=.170 [.07]. The differences between the adjusted and unadjusted treatment effect estimates for females in the Section 8 group and for males are also within the range of -.04 and +.06 standard deviations, and using adjusted versus unadjusted estimates does not change our inference about whether the treatment effects are significantly different from zero. Nevertheless, based on the evidence described above, we believe that the adjusted scores are our best estimate of the treatment effect. Substantively, this leads to a similar but slightly more conservative interpretation of the results. We have replicated other analyses of the test score effects for MTO using these adjusted test scores. While there is some effect on specific point estimates, all of the inferences of no significant effects on test scores also hold when using the adjusted test score (for example, in the specifications reported in Orr et al 2003).

Appendix C. Missing Outcome Data.

Although our effective response rate of 88 percent is high, it remains possible that the individuals for whom we did not collect outcome data could have had systematically different behavior than those for whom we did collect data, with this attrition resulting in bias of our estimated treatment effects. In this section we explore the implications of different assumptions about these missing data.[63] We also investigate the importance of the survey subsampling in reducing the sensitivity of our results to imputation of missing data, the implications of assumptions that noncomplying attriters in the treatment and control groups have the same average outcomes, and the differential importance of baseline covariates in predicting attrition in the treatment and control groups.

Smoking example. To be concrete, we focus our initial discussion on the example of the experimental - control difference in the prevalence of smoking among males. Analysis is discussed without regression adjustment to avoid tangential issues. The experimental group mean was .243 and the control group mean was .126, producing a treatment effect estimate of .117.

A worst case of positive attrition bias would be to assume that all missing data for the experimental group (experimental attriters) was zero, and all missing data for the control group (control attriters) was one. This would result in a treatment effect estimate of -.07. If this worst case assumption were true, then the observed estimate of .117 would be positively biased. A worst case of negative attrition bias (experimental attriters = 1; control attriters=0) would result in a treatment effect estimate of .24. This implies the worst case range is .24-(-.07)=.31.

An example of an alternative assumption about positive attrition bias could be based on experimental attriters being below the observed experimental group mean by half a standard deviation (.243-.227=.016) and the control attriters being above the observed experimental group mean by half a standard deviation(.126+.167=.293).[64] This would result in a treatment effect of .06, with a 95 percent confidence interval of [-.022,.139]. Negative attrition bias based on the

[63] Our emphasis in this section differs from recent work by Lee (2003), in that we explicitly make assumptions about all missing data. Lee focuses on differences in response rates between treatment groups and provides bounds for what would have been observed if the response rates had been the same across groups. See also Manski (1995).

[64] These calculations are based on imputations of zeros and ones to actual individuals in the data, with estimates calculated using sample weights. This method has the advantage of being as realistic as possible, but it does not lead to a unique solution for non-worst case imputation. For example, the mean of the Control attriters is close to but not exactly one half a standard deviation above the observed Control mean (.141 imputed and .165 calculated). Because the Control attriters are a finite sample of 17 individuals each with their own survey weight, imputation cannot match the exact calculation. Results reported here are based on a single random ordering of individuals. We have also investigated a more complex procedure involving multiple imputation, which does not affect the substantive character of the results reported in this section.

one half standard deviation assumption for experimental attriters (.243+.210=.453) and control attriters (.126-.126=0) would result in a treatment effect estimate of .17. The half standard deviation range is .17 - .06 = .11.

Importance of subsampling. In order to gauge the sensitivity of our analysis to weighting for survey subsampling, we performed the same calculations described above but ignoring information about who was in the subsample -- which lowers the effective response rate from 88% to 78%. This additional missing data in the smoking example increases the worst case range substantially, to .37 - (-.23) = .60, which is 1.9 times as large as the worst case range of .31 under subsampling. Similarly, the half standard deviation range without subsampling is .22 - .00 = .22, also about twice as large as with subsampling.

In a simple model in which the fraction of experimental and control attriters are exactly equal, the ratio of the range under imputation assumptions without subsampling to the range with subsampling will equal the ratio of the effective non-response rates for any outcome. In our data, this is approximately .22/.12, or 1.8. The actual attrition pattern is not this simple, differing between groups and also affected by survey weights for each individual, but the ratios of non-subsample to subsample ranges do fall in the neighborhood of 1.8 for all outcomes examined in this paper.

Note that the use of the subsample does not have a substantial impact on the point estimates (simple mean difference of .117 with subsample weighting and .120 without subsample weighting), and the subsample is too small to statistically reject any meaningful difference between the subsample and the initial main sample. However, the use of the subsample information does substantially reduce the sensitivity of the estimation to assumptions about data missing from survey attrition.

Sensitivity of effect sizes estimates. Using the subsample weighting and applying the type of missing data assumptions used in the example above about smoking, we present the sensitivity of selected effect size estimates in Table A9.[65] Columns 1 and 7 represent worst case bounds. Columns 2 and 6 represent half standard deviation differences between the observed mean and the imputed value for survey attriters, with column 2 increasing effect sizes and column 6 reducing effect sizes. Columns 3 and 5 represent half standard deviation differences between the

[65] Since there are no missing data after imputation, the effect size estimates in this section use treatment effect estimates from equation (2) rather than equation (10), where the outcome is the average of the normalized measures within the domain. Normalization for each outcome by the standard deviation of the Control group from equation (7). This is done for computational convenience, since the point estimates from the two approaches are numerically

observed mean and the imputed value for survey attriters. Column 4 imputes the treatment group mean to treatment attriters, and the control group mean to control attriters. Note that the finite and small number of attriters leads to small differences between the calculated means for attriters and non-attriters, and that this analysis is not regression-adjusted -- leading to some differences between Column 4 and the effect size estimates reported in Table 7.

The results show that, when comparing the opposing worst cases in columns 1 and 7, the estimates of the summary measures can change a great deal. For less extreme assumptions about the outcomes of attriters, the sign of the summary measure estimates in columns 3 - 5 does not change. Note, however, that while the sign does not change, as the results move closer to zero in absolute value (for females from column 5 to 4 and for males from column 3 to 4) the p-values are greater than .06 for eight of the eleven measures.

An alternative assumption to impute attriter data could be to impute the control mean to treatment attriters and the treatment mean to control attriters, effectively imputing the negative of the treatment effect to the attriters. The effect sizes in Table 7 range roughly from 0 to +.25 for females and 0 to -.25 for males. The assumption in column 5 of Table A9, for example, implies an effect size of -.50 among attriters. Thus, this approach results in imputations that are less than halfway between columns 4 and 5 for females, and 4 and 3 for males.

One way to obtain a sense of how different attriters' data might be from the observed data is to compare the initial main sample and subsample. As the subsample is very small, however, these results are highly variable. Subsample members in the treatment groups have more positive outcomes on all eleven measures in Table A9 than treatment group members in the initial main sample. The E/S - C difference between subsample members and initial main sample members ranges from .09sd to .50sd for females, and -.04sd to .16sd for males. Assuming the same pattern for attriters as for subsample members would imply that complete data estimates largely would lie somewhere between column 3 (E/S - C difference between observed sample and attriters of .50sd) and column 4 (no bias) for both females and males -- making results for females more positive and results for males less negative.

Assuming noncomplier attriter outcomes are equal. Use of additional assumptions can reduce the sensitivity of the treatment effect estimates to alternative imputations. Two assumptions that may be reasonable are to assume that the fraction of noncomplying attriters is the same in the treatment and control groups and to assume that the average outcome would have

---

identical. A common weight is used for all outcomes, assigning subsample status to individuals subsampled for either the survey, parental report, or WJR test.

been the same in the treatment and control groups for these noncomplying attriters. Since interviewers were not provided with information about treatment group assignment status and address histories of noncompliers were similar in treatment and control groups, search intensity for noncomplying attriters in both groups should have been approximately equal. We can observe the fraction of noncomplying attriters in the treatment group. Using these two assumptions, we can then impute noncompliance to the same fraction in the control group. Then instead of imputing outcomes to these noncomplying attriters in both the treatment and control groups, we can assume they are perfectly offsetting and remove them all from our imputation procedure.[66]

Table A10 has the same format as Table A9, but uses imputation assuming that noncomplying attriter proportions and outcomes are equal on average in the treatment and control groups. This has no effect on the E-C results for males, among whom there are few noncomplying treatment attriters. The half standard deviation range is somewhat reduced for the S-C male (from .29 in Table A9 to .23 in Table A10 for the overall summary measure) and S-C females (from .18 to .15 for the overall summary measure) contrasts. There is a very large reduction in the sensitivity of the E-C results for females using this approach. For the overall summary measure, for example, the half standard deviation range decreased from .24 to .13. It turns out that that the fraction of female noncomplying experimental group attriters is approximately equal to the fraction of all female control attriters. So, under the assumptions of this method, all female control attriters are assumed to be offset by noncomplying experimental group attriters -- and the results are therefore insensitive to assumptions about missing data for the control group.

Attrition and baseline covariates. One approach to ascertaining whether outcome data for attriters is likely to differ from outcome data for the observed sample is to analyze how treatment effects on attrition rates differ for subgroups defined by baseline covariates. For example, among females in LA and NY the experimental group attrition rate was .12 in the experimental group and .13 in the control group, whereas in Baltimore, Boston, and Chicago the attrition rate was .20 in the experimental group and .05 in the control group. Here the difference-in-difference attrition estimate is (.20-.05) - .(.12-.13) = .16.

[66] For this procedure, noncomplying attriters are only removed from the sample if they have missing data on all 15 outcomes in this study. If there was item nonresponse on some items and not others, then they are left in the sample and values for missing data are imputed.

If treatment interactions with covariates were jointly insignificant predictors of attrition rates, this would be some evidence that the differences in attrition process between treatment and control groups was independent of covariates and perhaps less systematic in nature. To assess this, we estimated linear probability models of attrition on our standard baseline covariates (Xs from Table 1), a main effect for treatment assignment, and full interactions of covariates with treatment assignment. For females, the p-value on the joint hypothesis that all interactions were zero was .16 for the experimental group and .57 for the Section 8 group. For males, the p-value was .01 for the experimental group and .10 for the Section 8 group.

TABLE A1. RISKY BEHAVIOR OUTCOME MEANS

	MTO			NLSY97	
	Exp (1)	Sec8 (2)	Control (3)	adjusted (4)	unadjusted (5)
A. Females					
Used marijuana in past 30 days	.07	.08	.13	.13	.16
Smoked in past 30 days	.14	.14	.19	.25	.33
Had alcohol in past 30 days	.14	.14	.21	.28	.44
Been or gotten someone pregnant	.25	.33	.27	.21	.14
B. Females -- gifted dropped					
Used marijuana in past 30 days	.07	.08	.14	.12	.15
Smoked in past 30 days	.14	.11	.20	.26	.33
Had alcohol in past 30 days	.13	.12	.23	.28	.43
Been or gotten someone pregnant	.26	.32	.27	.22	.14
C. Males					
Used marijuana in past 30 days	.19	.21	.12	.21	.18
Smoked in past 30 days	.24	.29	.13	.29	.33
Had alcohol in past 30 days	.21	.24	.14	.30	.46
Been or gotten someone pregnant	.16	.19	.12	.16	.07
D. Males -- gifted dropped					
Used marijuana in past 30 days	.18	.23	.13	.21	.18
Smoked in past 30 days	.22	.30	.12	.29	.33
Had alcohol in past 30 days	.19	.22	.13	.30	.45
Been or gotten someone pregnant	.17	.17	.12	.16	.07

Notes. Exp: Experimental. Sec8: Section 8. Con: Control. Columns 1-3 are unadjusted means using MTO survey weights. Column 5 is the unadjusted sample mean of NLSY97 Round 3 outcomes for ages 15-20 using NLSY97 survey weights. Using the same NLSY97 data, column 4 contains the predicted values from regressions of outcomes on covariates, based on MTO covariate means. Covariates were sixth order polynomial in age, race white, race other non-black, adult head age 19-29, adult head age 30-39, adult head age 40-49, household size 2, household size 3, household size 4, adult head has car, adult head employed, adult head GED or high school graduate, adult head receiving welfare, missing parental interview, youth gifted classes, youth remedial classes, youth disabled, youth special medical needs. MTO covariates are from the MTO baseline survey. NLSY97 age is as of Round 3 interview; other NLSY covariates are from Round 1, recoded to match MTO baseline covariates. Regressions were estimated separately for females and males and evaluated at the gender-specific means of the MTO baseline covariates (except missing parental interview indicator evaluated at NLSY97 mean). Panels B and D drop observations where youth had earlier been in gifted classes.

TABLE A2. P-VALUES FOR ALTERNATE MEAN EFFECT SIZE MEASURES

		Females			Males		
Domain	Contrast	OLS (1)	Probit (2)	Logit (3)	OLS (4)	Probit (5)	Logit (6)
Risky Behavior	E-C	.008	.004	.002	.002	.002	.002
	S-C	.122	.135	.117	.001	.0005	.0003
Physical Health	E-C	.982	.983	.992	.003	.002	.002
	S-C	.269	.225	.208	.012	.020	.012

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. OLS mean effect size estimates calculated using equation (9), based on treatment effect estimates from OLS estimation of equation (10). Probit estimates based on the mean of treatment effect z-scores from probit estimation of equation (10). Logit estimates based on the mean of treatment effect log odds from logit estimation of equation (10). All estimates rely on standard errors clustered by household.

TABLE A3. INTENT-TO-TREAT EFFECTS FOR NEIGHBORHOOD AND VICTIMIZATION MEDIATORS

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
A. General Neighborhood								
Youth lives in baseline neighborhood [SR]	.455	-.136* (.044)	-.147* (.046)	.485	-.101* (.049)	-.117* (.047)	.036 (.063)	.030 (.064)
Poverty rate in current neighborhood [ADDRESS]	.402	-.093* (.018)	-.075* (.017)	.396	-.089* (.019)	-.064* (.018)	.005 (.025)	.011 (.025)
Pct minority in neighborhood [ADDRESS]	.876	-.039 (.021)	.013 (.020)	.889	-.057* (.020)	-.051 (.026)	-.017 (.028)	-.064* (.032)
Pct youth in neighbrhood not in school or work [ADDRESS]	.121	-.020* (.009)	-.015 (.009)	.132	-.024 (.012)	-.019 (.012)	-.004 (.014)	-.004 (.014)
Pct adults in pro/mgmt occupations [ADDRESS]	.205	.046* (.011)	.016 (.010)	.220	.025* (.011)	.006 (.011)	-.020 (.015)	-.010 (.015)
Not satisfied with neighborhood [PR]	.555	-.204* (.052)	-.209* (.054)	.511	-.123* (.053)	-.062 (.058)	.081 (.075)	.147 (.080)
Feels unsafe in neighborhood at night [PR]	.437	-.160* (.048)	-.130* (.055)	.509	-.209* (.052)	-.129* (.056)	-.049 (.071)	.001 (.080)
Fraction of 4 types of discrimination in 'hood [SR]	.107	-.015 (.017)	-.005 (.019)	.134	-.021 (.020)	.004 (.025)	-.006 (.026)	.009 (.031)
Fraction of 6 problems with neighborhood [PR]	.565	-.175* (.037)	-.127* (.042)	.509	-.127* (.039)	-.054 (.038)	.048 (.055)	.073 (.058)
Saw drugs in neighborhood 1+/week in past 30 days [SR]	.437	-.103* (.048)	-.130* (.051)	.441	-.062 (.048)	-.019 (.054)	.041 (.067)	.111 (.073)
Heard gunshots in 'hood 1+/week in past 30 days [SR]	.118	-.030 (.029)	-.059* (.028)	.155	-.055 (.036)	-.070* (.034)	-.025 (.046)	-.011 (.044)
B. Victimization								
Household member was crime victim past 6 mths [PR]	.273	-.110* (.042)	-.109* (.044)	.247	-.042 (.047)	-.048 (.049)	.068 (.063)	.060 (.065)
Saw someone shot or stabbed in past 12 mths [SR]	.150	-.035 (.034)	-.040 (.035)	.209	-.016 (.039)	-.023 (.042)	.019 (.051)	.017 (.055)
Was "jumped" in past 12 months [SR]	.085	.005 (.028)	.009 (.027)	.181	.002 (.038)	-.013 (.039)	-.003 (.046)	-.022 (.047)

Notes. CM: Control mean. E-C: experimental - control difference. S-C: Section 8 - control difference. ADDRESS: Address history from tracking file, linked to Census. SR: Self-report. PRY: Parental report about youth. PR: Parental report about household. Differences regression-adjusted, using equation (4) with standard errors clustered by household. * indicates p-value <.05. For PR measures, analysis was conducted at household level using household average right-hand side variables, as discussed in section II. Surveys were completed in experimental, Section 8, and control groups with 749, 510, and 548 respondents respectively ages 15-20 on 12/31/2001. Types of discrimination were: at school or work, neighborhood recreation program, shopping or restaurant, with police. Problems with neighborhood were: litter, graffiti, public drinking, abandoned buildings, people hanging out, police not coming. Types of criminal victimization were: purse or wallet snatched, threatened with weapon, beaten or assaulted, break-in to home, stabbed or shot.

TABLE A4. INTENT-TO-TREAT EFFECTS FOR HOUSING, PARENTING, AND SCHOOL MEDIATORS

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
A. Housing								
Overall housing condition is fair/poor [PR]	.477	-.107* (.051)	-.022 (.056)	.507	-.090 (.053)	-.076 (.055)	.017 (.073)	-.055 (.079)
Fraction of 7 problems with home [PR]	.334	-.065* (.028)	-.042 (.032)	.333	-.076* (.029)	-.038 (.031)	-.011 (.040)	.004 (.045)
Fraction of 7 problems with home interior [OBS]	.215	-.064* (.022)	-.013 (.026)	.217	-.026 (.028)	-.020 (.032)	.039 (.035)	-.007 (.039)
Fraction of 7 problems with home exterior [OBS]	.217	-.082* (.027)	-.036 (.028)	.229	-.048 (.028)	-.023 (.030)	.034 (.037)	.014 (.039)
B. Parenting Practices								
Mother /primary caregiver is very supportive [SR]	.670	.024 (.044)	.016 (.049)	.842	-.039 (.037)	-.038 (.038)	-.064 (.057)	-.053 (.062)
Parent knows all about friends & whereabouts [SR]	.258	-.015 (.039)	-.047 (.044)	.173	-.044 (.035)	-.031 (.041)	-.029 (.053)	.016 (.060)
No adult present after school [SR]	.242	.046 (.049)	.018 (.054)	.301	.034 (.050)	.087 (.060)	-.012 (.069)	.069 (.080)
Fraction days/week family eats together [PR]	.571	.046 (.039)	.026 (.042)	.596	-.011 (.043)	.012 (.044)	-.057 (.058)	-.013 (.060)
Fraction of 4 types of parental contact w/schl [PR]	.370	.024 (.031)	.008 (.034)	.418	-.027 (.035)	-.005 (.036)	-.051 (.047)	-.013 (.050)
C. School Environment								
% free lunch [ADMIN]	.516	-.047* (.022)	.011 (.023)	.524	-.069* (.024)	-.038 (.027)	-.022 (.033)	-.049 (.035)
% limited English proficient [ADMIN]	.155	-.029* (.013)	.001 (.015)	.163	-.030* (.015)	-.030 (.015)	-.001 (.019)	-.030 (.021)
% white [ADMIN]	.114	.058* (.021)	.012 (.021)	.112	.059* (.023)	.076* (.030)	.001 (.031)	.063 (.037)
Pupil-teacher ratio [ADMIN]	18.6	.66 (.35)	-.28 (.43)	17.4	1.28* (.39)	.63 (.41)	.63 (.52)	.92 (.59)
Percentile rank on state exam [ADMIN]	.240	.033 (.024)	-.028 (.023)	.188	.059* (.027)	.046 (.026)	.026 (.036)	.074* (.034)
Fraction of 5 positive school climate items [SR]	.621	.009 (.031)	.045 (.033)	.599	.050 (.030)	-.013 (.035)	.042 (.043)	-.058 (.047)

For notes, see Table A8.

TABLE A5. INTENT-TO-TREAT EFFECTS FOR PEER AND ADULT ROLE MODEL MEDIATORS

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
A. Peers								
Has at least one close friend [PRY]	.890	.050 (.027)	.002 (.035)	.917	.018 (.026)	.043 (.027)	-.032 (.037)	.041 (.044)
Has 5 or more friends [SR]	.382	.057 (.044)	.037 (.048)	.530	.025 (.048)	.077 (.050)	-.032 (.065)	.040 (.069)
Friends involved in school activities [SR]	.615	.091* (.045)	.055 (.049)	.710	-.019 (.042)	-.022 (.048)	-.110 (.063)	-.077 (.069)
Has friends who use drugs [SR]	.295	.028 (.042)	.021 (.044)	.327	.121* (.049)	.136* (.052)	.093 (.064)	.114 (.067)
Has friends who carry weapons [SR]	.098	.010 (.025)	.024 (.030)	.157	.036 (.040)	-.027 (.037)	.027 (.046)	-.051 (.047)
Has relatives or friends who belong to a gang [SR]	.154	.008 (.035)	-.043 (.033)	.187	-.072* (.034)	-.056 (.036)	-.081 (.048)	-.013 (.047)
Friends from baseline visit new neighborhood [SR]	.178	-.016 (.036)	.022 (.042)	.164	-.018 (.035)	-.045 (.040)	-.002 (.050)	-.067 (.058)
Visits baseline 'hood but doesn't live there [SR]	.234	-.028 (.040)	.003 (.045)	.205	.043 (.040)	.038 (.047)	.071 (.057)	.035 (.065)
B. Adult Role Models								
Likely neighbors intervene vs. graffiti [PR]	.497	.188* (.052)	.103 (.059)	.575	.054 (.052)	-.026 (.058)	-.134 (.074)	-.129 (.083)
Likely neighbors intervene if kids skipping school [PR]	.343	.158* (.055)	.039 (.057)	.370	.097 (.052)	.061 (.059)	-.061 (.076)	.021 (.083)
Structured activity after school [SR]	.275	.075 (.044)	-.011 (.044)	.248	.069 (.042)	.070 (.047)	-.006 (.061)	.081 (.064)
Attended 1+ church youth activities per month [SR]	.380	.011 (.048)	-.022 (.049)	.313	-.033 (.044)	.016 (.048)	-.045 (.064)	.038 (.067)
Saw father at least once a week in past 12 months [SR]	.253	.036 (.042)	.078 (.046)	.365	-.046 (.045)	-.018 (.048)	-.081 (.060)	-.096 (.065)
Father has been very supportive [SR]	.235	.002 (.041)	.013 (.042)	.271	.023 (.042)	-.022 (.044)	.022 (.057)	-.034 (.059)
Comfortable talking about problems w/3+ adults [SR]	.305	.141* (.044)	.094* (.047)	.397	.002 (.047)	.025 (.053)	-.139* (.064)	-.069 (.070)
Has 4+ adults who care and will help if trouble [SR]	.448	.069 (.047)	.034 (.048)	.498	.030 (.048)	-.023 (.053)	-.039 (.067)	-.057 (.072)

For notes, see Table A8.

TABLE A6. INTENT-TO-TREAT EFFECTS FOR EDUCATIONAL MEDIATORS

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
A. School Engagement								
Always pays attention in class [SR]	.490	.124* (.056)	.041 (.062)	.484	.014 (.056)	.062 (.061)	-.110 (.078)	.020 (.087)
Works hard in school [SR]	.508	.051 (.058)	.030 (.059)	.449	-.096 (.057)	-.017 (.064)	-.147 (.080)	-.047 (.087)
B grades or higher last year [SR]	.415	.009 (.047)	-.004 (.049)	.293	-.061 (.043)	-.103* (.044)	-.070 (.063)	-.098 (.066)
Always finishes homework [SR]	.505	.041 (.058)	-.006 (.066)	.406	-.015 (.055)	-.044 (.063)	-.056 (.080)	-.038 (.092)
At least 5 hours/week of homework [SR]	.488	.068 (.057)	.002 (.061)	.354	.083 (.055)	.091 (.064)	.015 (.078)	.089 (.087)
At least 5 hours/week of reading [SR]	.377	-.011 (.046)	-.017 (.049)	.250	.019 (.047)	.031 (.052)	.029 (.065)	.048 (.070)
B. Educational Track								
Ever took SAT, ACT, or AP exams [SR]	.426	-.019 (.049)	.025 (.052)	.358	-.039 (.048)	.055 (.055)	-.020 (.069)	.030 (.075)
Ever took algebra or higher math [SR]	.833	.015 (.034)	.005 (.035)	.827	-.075* (.037)	-.039 (.037)	-.090 (.051)	-.044 (.051)
Gifted class in past 2 years [PRY]	.068	.068* (.032)	.000 (.030)	.147	-.024 (.043)	-.049 (.037)	-.092 (.052)	-.049 (.048)
Special education in past 2 years [PRY]	.154	.056 (.039)	-.003 (.042)	.324	-.009 (.049)	-.054 (.050)	-.065 (.061)	-.051 (.066)
C. Educational Problems								
Ever repeated a grade [PRY]	.200	.093* (.038)	.007 (.037)	.326	-.038 (.043)	-.061 (.043)	-.131* (.057)	-.067 (.057)
Late for school once a month or more [SR]	.679	-.044 (.042)	.005 (.048)	.616	.045 (.048)	.027 (.054)	.089 (.064)	.022 (.072)
Absent from school 5% or more of the school year [SR]	.426	-.097* (.046)	-.048 (.052)	.389	-.003 (.047)	-.012 (.053)	.094 (.066)	.036 (.073)
School requested meet about prob past 2 yrs [PRY]	.184	-.004 (.040)	-.017 (.044)	.337	.034 (.050)	.011 (.057)	.038 (.062)	.028 (.071)
Was suspended/expelled from school past 2 yrs [PRY]	.117	.023 (.033)	.037 (.042)	.301	-.046 (.049)	-.105* (.053)	-.069 (.056)	-.143* (.067)
D. Future Expectations								
Believes chances high will complete college [SR]	.543	.103* (.046)	.045 (.050)	.449	-.033 (.046)	-.059 (.049)	-.136* (.065)	-.104 (.071)
Believes chances high will find good job as adult [SR]	.742	.059 (.037)	.011 (.044)	.652	.007 (.045)	.036 (.048)	-.051 (.058)	.025 (.066)

For notes, see Table A8.

TABLE A7. INTENT-TO-TREAT EFFECTS FOR HEALTH MEDIATORS

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
A. Healthy environment								
Fraction of past 7 days did aerobic exercise [SR]	.353	.059 (.032)	.014 (.036)	.555	.012 (.033)	.019 (.036)	-.047 (.045)	.005 (.051)
Fraction of past week moderate activity [SR]	.412	.034 (.033)	-.022 (.038)	.476	.062 (.036)	.050 (.038)	.028 (.049)	.072 (.054)
Participates in sport after school [SR]	.032	.050* (.022)	.005 (.018)	.138	.003 (.032)	.040 (.037)	-.047 (.039)	.035 (.042)
Fraction of past 7 days some fruits/vegetables [SR]	.574	.013 (.032)	-.032 (.036)	.568	-.001 (.030)	.027 (.032)	-.014 (.044)	.059 (.048)
Fraction of 6 asthma triggers [PR]	.263	-.005 (.019)	-.008 (.021)	.265	-.033 (.020)	-.005 (.022)	-.028 (.028)	.003 (.032)
B. Access to care								
Youth has health insurance [PRY]	.876	-.006 (.029)	.006 (.030)	.819	.066* (.030)	.015 (.036)	.073 (.040)	.009 (.044)
Talked to a doctor about health in past 6 mths [PRY]	.736	.072 (.047)	.000 (.055)	.728	-.079 (.055)	-.106 (.062)	-.152* (.073)	-.106 (.083)
C. Adult mental health								
Adult distress K6 z-score [PR]	.170	-.247* (.111)	-.210 (.127)	-.058	.112 (.106)	.165 (.118)	.360* (.158)	.376* (.177)
Adult probability of depression [PR]	.186	-.025 (.035)	-.071* (.035)	.137	-.002 (.031)	.044 (.039)	.023 (.048)	.115* (.053)
Adult fraction worried, tense or anxious [PR]	.424	-.055 (.051)	-.068 (.055)	.453	-.060 (.054)	-.033 (.058)	-.005 (.076)	.035 (.082)
Adult fraction calm and peaceful [PR]	.375	.139* (.049)	.138* (.054)	.508	-.040 (.051)	-.128* (.058)	-.179* (.072)	-.267* (.081)
Adult fraction sleeping 7-8 hours/night [PR]	.362	.123* (.052)	.097 (.054)	.409	.126* (.052)	.037 (.057)	.003 (.075)	-.061 (.080)

For notes, see Table A8.

TABLE A8. INTENT-TO-TREAT EFFECTS FOR RESIDENTIAL MOBILITY

Outcome	Female			Male			Male - Female	
	CM (1)	E-C (2)	S-C (3)	CM (4)	E-C (5)	S-C (6)	E-C (7)	S-C (8)
Program move [ADDRESS]	0	.483*	.577*	0	.411*	.515*	-.073	-.062
		(.033)	(.037)		(.032)	(.039)	(.045)	(.052)
One or more moves [ADDRESS]	.649	.188*	.205*	.673	.086*	.076	-.101	-.129*
		(.041)	(.041)		(.041)	(.041)	(.055)	(.055)
One or more moves [PR]	.623	.080	.081	.654	.024	-.025	-.056	-.106
		(.047)	(.048)		(.048)	(.046)	(.067)	(.065)
Two or more moves [ADDRESS]	.295	.138*	.176*	.346	.028	.053	-.110	-.122
		(.045)	(.046)		(.047)	(.051)	(.062)	(.066)
Two or more moves [PR]	.266	.017	.099	.301	-.006	.032	-.023	-.067
		(.046)	(.051)		(.043)	(.050)	(.064)	(.070)
Number of moves [ADDRESS]	1.01	.484*	.524*	1.20	.133	.139	-.351*	-.386*
		(.090)	(.098)		(.105)	(.107)	(.134)	(.140)
Number of moves [PR]	1.11	-.001	.192	1.25	-.088	-.070	-.087	-.262
		(.125)	(.129)		(.126)	(.132)	(.158)	(.165)

Notes to Tables A3-A8. CM: Control mean. E-C: experimental - control difference. S-C: Section 8 - control difference. SR: Self-report. PRY: Parental report about youth. PR: parental report about household. ADDRESS: Address history from tracking file, linked to Census. ADMIN: Administrative data from school reported to attend or last attended. OBS: interviewer observation of housing unit. Differences regression-adjusted, using equation (4) with standard errors clustered by household. * indicates p-value <.05. For PR measures, analysis was conducted at household level using household average right-hand side variables, as discussed in section II. Sample is ages 15-20 as of 12/31/01. Some items not asked for youth ages 19-20, resulting in smaller sample sizes: no adult present after school, fraction of school climate items, pays attention, works hard, finishes homework, 5+ hours homework, gifted class, special education, school requested meeting, suspended/expelled. Problems with home were: peeling paint, plumbing, rats or mice, cockroaches, broken locks, broken windows, heat. Interviewer observations of problems with home interior were: cracks in walls, peeling paint, mold, cigarette smoke, noisy inside, noisy outside, cluttered. Interviewer observations of problems with home exterior were: condition of unit, condition of other units on block, metal bars on unit, metal bars on other units, condition of block, broken windows, junk on block. Items parental knows everything about were: who friends are, who with when not home. No adult present was: no adult at either 3:45, 5:30, or 7:30 on selected day of week. Parental contact with school (for any child in household) was: went to general school meeting, went to a school event, volunteered at school, volunteered for team or club. School climate was: teachers interested in students, students disruptive, cheating on tests, discipline fair, felt safe. Structured activity was: at school, church or community center -- participating in a sport, club, tutoring, or other organized activity. Asthma triggers were: rats or mice, cockroaches, wall-to-wall carpet, pets with fur, cigarette smoke, mold.

TABLE A9. EFFECT SIZES UNDER VARIOUS ASSUMPTIONS ABOUT MISSING DATA

Gender	Domain	Contrast	E/S: =max C: =min (1)	E/S: +.5sd C: -.5sd (2)	E/S: +.25sd C: -.25sd (3)	E/S: +0 C: -0 (4)	E/S: -.25sd C: +.25sd (5)	E/S: -.5se C: +.5sd (6)	E/S: =min C: =max (7)
A. Females	Overall	E-C	.41	.22	.17	.11	.04	-.02	-.46
	Overall	S-C	.33	.16	.12	.08	.03	-.02	-.26
	Risky Behavior	E-C	.36	.23	.19	.12	.06	.01	-.33
	Mental Health	E-C	.48	.32	.29	.22	.15	.10	-.43
	Mental Health	S-C	.42	.25	.22	.18	.13	.09	-.20
B. Males	Overall	E-C	.34	.03	-.04	-.12	-.21	-.28	-.68
	Overall	S-C	.34	.03	-.03	-.11	-.20	-.26	-.61
	Risky Behavior	E-C	.25	-.05	-.12	-.22	-.31	-.40	-.67
	Risky Behavior	S-C	.19	-.13	-.22	-.31	-.41	-.50	-.74
	Physical Health	E-C	.33	-.05	-.01	-.10	-.18	-.24	-.57
	Physical Health	S-C	.32	-.02	-.02	-.10	-.19	-.21	-.51

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. As described in Appendix C, estimates are not regression-adjusted. Column 1 imputes the maximum value to survey attriters in the experimental or Section 8 (E/S) groups and the minimum value to attriters in the control (C) group. Column 2 imputes .5sd above the E/S mean to E/S attriters and .5sd below the C mean for C attriters. Column 3 imputes .25sd above the E/S mean for E/S attriters and .25sd below the C mean for C attriters. Column 4 imputes the E/S mean for E/S attriters and the C mean for C attriters. Column 5 imputes .25sd below the E/S mean for E/S attriters and .25sd above the C mean for C attriters. Column 6 imputes imputes .5sd below the E/S mean for E/S attriters and .5sd above the C mean for C attriters. Column 7 imputes the minimum value to E/S attriters and the maximum value to C attriters. For binary variables, the minimum and maximum are 0 and 1. For continuous variables (test scores and the K6 mental health scale), the minimum and maximum are the most extreme scores observed in the data.

TABLE A10. EFFECT SIZES UNDER ASSUMPTIONS ABOUT MISSING DATA, ASSUMING EQUAL OUTCOMES FOR NONCOMPLYING ATTRITERS

Gender	Domain	Contrast	E/S: =max C: =min (1)	E/S: +.5sd C: -.5sd (2)	E/S: +.25sd C: -.25sd (3)	E/S: +0 C: -0 (4)	E/S: -.25sd C: +.25sd (5)	E/S: -.5se C: +.5sd (6)	E/S: =min C: =max (7)
A. Females	Overall	E-C	.27	.17	.14	.11	.07	.04	-.22
	Overall	S-C	.30	.15	.12	.08	.03	-.00	-.22
	Risky Behavior	E-C	.23	.18	.16	.12	.09	.06	-.13
	Mental Health	E-C	.36	.28	.26	.23	.19	.16	-.16
	Mental Health	S-C	.40	.24	.22	.18	.14	.10	-.17
B. Males	Overall	E-C	.34	.03	-.04	-.13	-.21	-.27	-.66
	Overall	S-C	.30	.01	-.04	-.11	-.18	-.24	-.53
	Risky Behavior	E-C	.25	-.05	-.12	-.23	-.32	-.40	-.66
	Risky Behavior	S-C	.15	-.15	-.21	-.32	-.40	-.46	-.67
	Physical Health	E-C	.33	-.05	-.01	-.10	-.17	-.23	-.55
	Physical Health	S-C	.29	.01	-.02	-.10	-.19	-.22	-.43

Notes. E-C: experimental - control difference. S-C: Section 8 - control difference. As described in Appendix C, estimates are not regression-adjusted. Column 1 imputes the maximum value to survey attriters in the experimental or Section 8 (E/S) groups and the minimum value to attriters in the control (C) group. Column 2 imputes .5sd above the E/S mean to E/S attriters and .5sd below the C mean for C attriters. Column 3 imputes .25sd above the E/S mean for E/S attriters and .25sd below the C mean for C attriters. Column 4 imputes the E/S mean for E/S attriters and the C mean for C attriters. Column 5 imputes .25sd below the E/S mean for E/S attriters and .25sd above the C mean for C attriters. Column 6 imputes imputes .5sd below the E/S mean for E/S attriters and .5sd above the C mean for C attriters. Column 7 imputes the minimum value to E/S attriters and the maximum value to C attriters. For binary variables, the minimum and maximum are 0 and 1. For continuous variables (test scores and the K6 mental health scale), the minimum and maximum are the most extreme scores observed in the data.