STRUCTURAL IVE FOR DYNAMIC TREATMENT EFFECTS: SPANKING EFFECT ON BEHAVIOR

(June 27, 2006)

Myoung-jae Lee*
Department of Economics
Korea University
Anam-dong, Sungbuk-gu
Seoul 136-701, South Korea
myoungjae@korea.ac.kr

Fali Huang
School of Economics and Social Sciences
Singapore Management University
90 Stamford Road
Singapore 178903.
flhuang@smu.edu.sg

ABSTRACT

Finding the effects of multiple sequential treatments on a response variable measured at the end of a trial is difficult, if some treatments are affected by interim responses; e.g., assessing the effects of spanking on behavior when parents adjust their spanking level depending on interim behaviors. A headway, '*G estimation*', has been made in 1980's generalizing the usual static treatment effect analysis under 'selection on observables'. But G estimation is hard to implement. In this paper, firstly, we propose a much simpler alternative to G estimation—a single or multiple *IVE's for a linear structural model*—and show that our proposal and G estimation identify the same effect under some assumptions. Secondly, we explore the relation between our proposal and *Granger causality* to show that our approach is more general, although the two become equivalent for testing non-causality under a stationarity assumption. Thirdly, our approach and G estimation are applied to find the effects of spanking on behavior. We find that mild spanking at early years reduces a child's behavior problems later, which seems to differ from most findings in the psychology literature.

Key words: dynamic model, treatment effect, panel data, causality, spanking

JEL Classification Numbers: C33, I20, J13, E60

*Myoung-jae Lee gratefully acknowledges the financial support of Wharton-SMU Research Center, Singapore Management University.

1 Introduction

Non-cognitive skills including personal traits such as discipline, conscientiousness, or motivation seem to be important determinants of earnings; see, e.g., Heckman (1999), Bowles et al. (2001), and Persico et al. (2004). The important role of these non-cognitive skills in worker performance has been recognized for a long time (Kandel and Lazear (1992) and Kreps (1997)). There is also some evidence that the non-cognitive skill formation in early childhood is crucial since "success or failure at this stage feeds into success or failure in school which in turn leads to success or failure in post-school learning" (Heckman (1999)). Keane and Wolpin (1997) show that the skill heterogeneity at age 16 may account for as much as 90% of the total variance of lifetime earnings.

Childhood non-cognitive skills are closely related to their behavioral problems. It is thus interesting to know whether spanking corrects or worsens the behavioral problems. Provided that a causal effect of childhood good behavior on adulthood earnings exists, if a causal link from spanking to childhood behavior is found, spanking could have lingering economic—as well as psychological—consequences. The *empirical goal of this paper is to find the effect of spanking on child behavior.*

Whether mild to moderate spanking works has been hotly debated in psychology and education as well as among the public. In a meta analysis comparing many studies over 60 years, Gershoff (2002) concludes that there are strong negative associations between corporal punishment and a range of child behaviors. This is a non-causal statement only to suggest that corporal punishment may worsen behavioral problems. The difficulty in establishing the causal link is the endogeneity of spanking arising from various sources. First, children and the parents may share predisposition (e.g., genes for violence) to misbehave and spank, respectively; here, the source of endogeneity is time-constant. Second, inappropriate home inputs (e.g., economically depressed environment) may foster poor behavior of children and violent behavior of the parents; here, the source of endogeneity is time-variant. Third, spanking can affect behavior which can in turn affect spanking. For example, effective spanking may stop a child's bad behaviors and hence prevent bad habits from forming in the beginning. So spanking at early ages may reduce the need to spank later (Larzelere 1996).

The first two sources of endogeneity—unobserved common factors—arise in static treatment-response framework as well, and the third forms the core of the dynamic feature that will

be the focal point of this paper. Our *analytic goal is to set up a dynamic treatment effect framework with linear structural equations and estimate the effects with instrumental variable estimator (IVE) and 'G estimation (or G computation algorithm)'* that has been developed in the epidemiological/medical literature. Our dynamic treatment effect analysis extends the usual static treatment effect analysis as in Angrist and Krueger (1999), Heckman et al. (1999), Rosenbaum (2002), and Lee (2005).

The rest of this paper is organized as follows. Section 2 shows that the usual dynamic panel data approach fails to identify the desired dynamic treatment effects by missing '*indirect effects' through interim responses*. Section 3 presents simple modifications to the usual approach to find the desired effects; those modifications are our main proposals. This section also compares our approaches to Granger causality and motivates 'G estimation' as another alternative. Section 4 reviews G estimation and show how it is related to our approaches. Section 5 explains our data and presents the empirical findings; a coherent finding emerging from various specifications is that moderate spanking at early ages reduces child behavior problems several years later. Finally, Section 6 concludes.

Throughout the main text of this paper, we examine only two period/treatment cases to simplify our presentation. In the appendix, three period/treatment generalizations are illustrated for our main results, which shows that a further generalization to four or more periods/treatments is straightforward.

2 Failure of Dynamic Panel Data Model

Suppose we have

$$(x_0, y_0),\ (d_1, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}),\ (d_2, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix})$$

where x_0 and y_0 are the baseline covariate and response, and treatment d_t at period t temporally precedes x_t and y_t (no temporal ordering yet between x_t and y_t), $t = 1, 2$. We are interested in the total effect of the treatment 'profile' $d \equiv (d_1, d_2)'$ on the last response y_2, while *(i) allowing for d_1 to affect both y_1 and y_2, and (ii) allowing for d_2 to depend on the interim response y_1*. If d_2 depends on y_1, then it is also natural for d_1 to depend on y_0.

In the spanking-behavior question, (i) means that spanking in period 1 may affect behavior in period 1 and 2, and (ii) means that spanking in period 2 may depend on behavior

in period 1. It will be ill-advised to rule out either. Particularly, ruling out (ii) implies continued spanking despite improved behavior, which is nonsensical unless 'preventive spanking' is practiced. But allowing for both creates a conflict: the lagged response y_1 should be controlled in view of (ii), but then the indirect effect of d_1 on y_2 through y_1 in (i) is missed. We elaborate on this key point in the following.

The usual approach in econometrics would be setting up a dynamic panel data model

$$y_{i2} = \beta_1 + \beta_y y_{i1} + \beta_{d1} d_{i1} + \beta_{d2} d_{i2} + \beta'_{x0} x_{i0} + \beta'_{x1} x_{i1} + \beta'_{x2} x_{i2} + v_{i2}, \text{ iid across } i = 1, ..., N \quad (2.1)$$

where the β parameters are to be estimated, and some regressors are possibly correlated with the error term v_{i2} (an endogeneity problem). Besides the above motivation of controlling y_{i1} in the preceding paragraph, another often-invoked motivation to control for y_{i1} in (2.1) is that y_{i1} captures the unobserved variables relevant for y_{i2} to lessen the endogeneity problem of the other regressors. In view of the iid assumption, often we will drop the subscript i in the remainder of this paper.

In (2.1), the *indirect effect of* d_1 *on* y_2 *through* y_1 *is missed* because y_1 is controlled. Specifically, if the effect of d_1 on y_1 is γ_y, then the indirect effect of d_1 on y_2 through y_1 is $\beta_y \gamma_y$, whereas (2.1) identifies only the direct effects β_{d1} and β_{d2} of d_1 and d_2 on y_2. The desired *total effect of the treatment profile is then* $\beta_{d1} + \beta_y \gamma_y$ *(from* d_1*) plus* β_{d2} *(from* d_2*).* As shown in detail in the next section, if y_1 is substituted out of (2.1) to leave the last-lagged y_0 on the right-hand side, then the sum of the coefficients of d_1 and d_2 is the total effect. This solution, however, not just gives an odd-looking model because y_0 instead of y_1 appears for the y_2 equation, but also makes it more difficult to find instruments for the endogeneity problem. Of course, if there are extra sources for instruments as the list of ingenious instruments in Angrist and Krueger (2001) illustrates, then this would not be much of a worry. But typically such extra variables are hard to find, and one then has to find instruments from within the model, namely, the past (or future) variables. In the rest of this section, we briefly discuss sources of instruments, because this issue is unavoidable for our empirical model later lacking any extra instrument source.

Instruments typically come from exclusion restrictions such as $\beta_{x0} = \beta_{x1} = 0$—i.e., only the contemporaneous covariates appear—combined with assumptions on the error term; e.g.,

$$v_{it} = \delta_i + u_{it}, \quad COR(\delta_i, x_{it}) \neq 0 \ \forall t, \quad (2.2)$$

$$COR(x_{is}, u_{it}) = 0 \text{ for } \forall s < t. \quad (2.3)$$

Condition (2.3) with $\forall s \leq t$ would be called the 'predeterminedness' of x_t; the equality $s = t$ is removed in (2.3) because x_t may be simultaneously related to y_t (i.e., to v_t). Conditions (2.2) and (2.3) imply, e.g., $COR(v_2 - v_1, x_0) = 0$: IVE with instruments x_0 can be applied to the $\Delta y_2 \equiv y_2 - y_1$ equation. Alternatively, if we assume

$$x_{it} = \zeta_i + e_{it} \text{ and } COR(\zeta_i, \delta_i) \neq 0, \; COR(e_{is}, v_{it}) = 0 \text{ for } \forall s < t, \tag{2.4}$$

then we can use the condition $COR(v_2, x_1 - x_0) = 0$ for the y_2 equation; here, x_t is first-differenced, not the y_2 equation.

The conditions (2.2) to (2.4) reflect three main concerns about the endogeneity of x_t in the y_t equation:

$$\begin{aligned} &\text{(i) } x_t \text{ related to the 'unit-specific effect' } \delta; \\ &\text{(ii) } x_t \text{ related to } v_t \text{ due to a simultaneous relation with } y_t; \\ &\text{(iii) future } x_t \text{ affected by the past } v_t \text{ (or } y_t). \end{aligned} \tag{2.5}$$

As to be seen later, these concerns are germane to our data. But these endogeneities may not occur in all components of x_t. We can then classify the components of x_t so that each component can be used to its fullest extent as an instrument when 'purified' (i.e., the endogeneity removed) properly. See Lee (2002) for more on assumptions on the error terms and regressors, and the ensuing moment conditions for panel data IVE.

3 IVE for Linear Structural Models

Define the 'potential responses' for the observed responses y_1 and y_2:

y_1^j : potential response when $d_1 = j$ is exogenously set,

y_2^{jk} : potential response when $d_1 = j$ and $d_2 = k$ are exogenously set, $\quad j, k \in [0, \infty)$.

Suppose the goal is to find the mean treatment effect $E(y_2^{jk} - y_2^{00})$ for treatment levels j and k versus no treatment at all. Our interest is in the 'intervention effect' of setting d_1 and d_2 exogenously, not in the 'self-selection' effect of allowing the subjects to choose d_1 and d_2.

The observed responses y_1 and y_2 are y_1^j and y_2^{jk} when $d_1 = j$ and $d_2 = k$; i.e., only the potential responses corresponding to the realized treatment levels are observed, and all the other potential responses—'counter-factuals'—are not. With d_1 and d_2 observed, we have

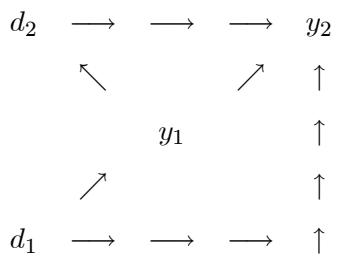
thus $y_1 = y_1^{d_1}$ and $y_2 = y_2^{d_1 d_2}$. Since we will be modeling y_1^j and y_2^{jk}, not y_1 and y_2 directly, we need to express y_1 and y_2 in terms of y_1^j and y_2^{jk}. For this, rewrite the observed y_1 and y_2 as

$$y_1 = \int y_1^j \cdot \partial 1[d_1 \leq j] \quad \text{and} \quad y_2 = \int y_2^{jk} \cdot \partial 1[d_1 \leq j, d_2 \leq k],$$

where ∂ is used instead of d for integration to prevent confusion, the first integral is with respect to (wrt) to the distribution $1[d_1 \leq j]$ for j that is degenerate at d_1, and the second integral is wrt to the distribution $1[d_1 \leq j, d_2 \leq k]$ for (j,k) that is degenerate at (d_1, d_2).

Observe the following figure that omits y_0, x_0, x_1, x_2:

Two Period Effects



d_2 has only a direct effect on y_2, but d_1 has both direct and indirect (through y_1) effects on y_2. If y_1 is controlled as in the dynamic panel model, then the indirect effect of d_1 on y_2 is not identified. If y_1 is not controlled, however, then the effect of d_2 on y_2 can be distorted because y_1 becomes a 'common factor' for d_2 and y_2. That is, even if there is no effect of d_2 on y_2, we may find a spurious effect of d_2 due to not controlling y_1. In the following, we will find the total effect of d using IVE for a linear structural model, and then compare our approach to Granger causality in Granger (1969,1980).

3.1 First- and Last-Lag Response IVE

Consider a '*contemporaneous covariate' model*' ($\beta_{x0} = \beta_{x1} = 0$ in (2.1)):

$$\begin{aligned} y_{i1}^j &= \gamma_1 + \gamma_y y_{i0} + \gamma_d j + \gamma_x' x_{i1} + v_{i1}, \\ y_{i2}^{jk} &= \beta_1 + \beta_y y_{i1}^j + \beta_{d1} j + \beta_{d2} k + \beta_{x2}' x_{i2} + v_{i2} \end{aligned}$$

where γ's and β's are parameters. The coefficients of the y_1^j and y_2^{jk} equations differ to allow for nonstationarity. Observe

$$\begin{aligned} y_{i1} &= \int y_{i1}^j \partial 1[d_1 \le j] = \int (\gamma_1 + \gamma_y y_{i0} + \gamma_d j + \gamma_x' x_{i1} + v_{i1}) \partial 1[d_1 \le j] \\ &= \gamma_1 + \gamma_y y_{i0} + \gamma_d d_1 + \gamma_x' x_{i1} + v_{i1}; \\ y_{i2} &= \int (\beta_1 + \beta_y y_{i1}^j + \beta_{d1} j + \beta_{d2} k + \beta_{x2}' x_{i2} + v_{i2}) \partial 1[d_1 \le j, d_2 \le k] \\ &= \beta_1 + \beta_y y_{i1}^{d_1} + \beta_{d1} d_1 + \beta_{d2} d_2 + \beta_{x2}' x_{i2} + v_{i2} \\ &= \beta_1 + \beta_y y_{i1} + \beta_{d1} d_1 + \beta_{d2} d_2 + \beta_{x2}' x_{i2} + v_{i2}. \end{aligned} \tag{3.1}$$

We can estimate the two equations separately with IVE to find

$$\begin{aligned} \textit{direct and indirect effects of } d_1 \text{ on } y_2 &: \quad \beta_{d1} + \beta_y \gamma_d, \\ \textit{direct effect of } d_2 \text{ on } y_2 &: \quad \beta_{d2}; \end{aligned}$$

the total effect of d is then the sum of these two lines. The source for the instruments in the y_1 equation is x_0, and the source for the instruments in the y_2 equation is x_0 and x_1. This is a two-step IVE method, for IVE is applied twice.

As often done in the panel data literature, first-differencing the y_t equation to get rid of the unit-specific effect δ_i under (2.2) yields

$$y_{i2} - y_{i1} = \beta_1 - \gamma_1 + \beta_y y_{i1} - \gamma_y y_{i0} + (\beta_{d1} - \gamma_d) d_1 + \beta_{d2} d_2 + \beta_{x2}' x_{i2} - \gamma_x' x_{i1} + u_{i2} - u_{i1}$$

where β_{d1} and γ_d are not separated. One may estimate this and the y_1 equation with IVE to find the desired parameters. But this procedure does not seem coherent, because the $y_1 - y_0$, not y_1, equation ought to be estimated along with the $y_2 - y_1$ equation. A more coherent procedure would be keeping the y_2 and y_1 equations intact and using IVE with first-differenced (or transformed) x_t. In this case, if we are concerned about all three endogeneity sources in (2.5), then $COR(v_2, x_1 - x_0) = 0$ may be used for the y_2 equation. But, this does not work for the y_1 equation, for there is no $x_0 - x_{-1}$.

One way to overcome the lack of instruments for the y_1 equation is, instead of applying condition (2.5) to all components of x_t, classifying the regressors to get enough moment conditions. To see this point, let w_t denote a component of x_t (if w_t is time-constant, only (a) is applicable in the following classifications that are not necessarily exhaustive) and consider

- (a) w_t is uncorrelated to v_s at all leads and lags: $COR(v_s, w_t) = 0 \ \forall s, t$.

- (b) w_t may be correlated to v_s only through its time-constant component: $COR(v_s, w_t - w_{t-1}) = 0 \ \forall s, t$; alternatively, $COR(v_s, w_t - \bar{w}) = 0$ where $\bar{w}_i \equiv T^{-1} \sum_i w_{it}$.
- (c) w_t may be only simultaneously related to v_s: $COR(v_s, w_t) = 0 \ \forall s \neq t$.
- (d) w_t may be related only to the past v_s: $COR(v_t, w_s) = 0 \ \forall s \leq t$.

Once this kind of classification is done, we get two sets of instruments for v_1 and v_2, respectively, and IVE can be applied to each equation separately.

Another way to overcome the problem of insufficient instruments for the y_1 equation is assuming

$$\textit{equal contemporaneous effects}: \ \gamma_d = \beta_{d2}$$

that the effect of d_1 on y_1 is the same as the effect of d_2 on y_2. This is a stationarity assumption, under which we get

$$d_1 \text{ effect } \beta_{d1} + \beta_y \beta_{d2} \quad \text{and} \quad d_2 \text{ effect } \beta_{d2}.$$

These are estimable with the y_2 equation only.

Instead of doing IVE twice or only once under $\gamma_d = \beta_{d2}$, substitute out y_1^j in the y_2^{jk} equation. Then replace y_2^{jk}, j, k with y_2, d_1, d_2, respectively, to get

$$y_{i2} = (\beta_1 + \beta_y \gamma_1) + \beta_y \gamma_y y_{i0} + (\beta_{d1} + \beta_y \gamma_d) d_{i1} + \beta_{d2} d_{i2} + \beta_y \gamma_x' x_{i1} + \beta_{x2}' x_{i2} + (\beta_y v_{i1} + v_{i2}). \quad (3.2)$$

IVE can be applied only once to this equation to find the total effect of d_1 and d_2 as the sum of the coefficients of d_1 and d_2. This equation looks unusual in that the last-lag response y_0 is included, but not the first lag response y_1. This *last-lag response IVE* for (3.2) is simpler, being one-step IVE than the above *first-lag response IVE* for (3.1), but there are two disadvantages: decomposition of the total effect of d_1 cannot be done with (3.2) alone, and there is in general less instrument source because x_1 and x_2 are included in the right-hand side and the error term consists of v_1 and v_2. The main source for the instruments for (3.2) is x_0, but since $x_0 - x_{-1}$ is not available, conditions such as $COR(\beta_y v_1 + v_2, \ x_0 - x_{-1})$ cannot be used. Classification of covariates as in (a) to (d) above is thus called for.

It should be noted that the our IVE so far are based on the contemporaneous covariate model. Despite criticism as in Todd and Wolpin (2004),[1] given our data, it is impossible

[1] When using a current summary indicator for home environment, they find earlier indicators significantly

to include all current and past covariates in the y_2 equation, which essentially eliminates all sources for instruments. It is possible, however, to relax the assumption of contemporaneous covariate model selectively only for some covariates, which again requires classifying covariates as in (a) to (d) above. Even when we use the contemporaneous model, there still occur issues such as 'whether to control or not the regressors affected by the treatments' and 'what happens to the various effects when nonlinear functions of d_1 and d_2 are used'. These issues are addressed in the appendix.

Going one step further from (3.2), we may assume $y_{0i} = \gamma_{10} + \gamma'_{x0} x_{i0} + v_{i0}$ and substitute this into the y_2 equation (3.2) to get

$$\begin{aligned} y_{i2} &= (\beta_1 + \beta_y\gamma_1 + \beta_y\gamma_y\gamma_{10}) + (\beta_y\gamma_d + \beta_{d1})d_{i1} + \beta_{d2}d_{i2} \\ &\quad + \beta_y\gamma_y\gamma'_{x0}x_{i0} + \beta_y\gamma'_x x_{i1} + \beta'_{x2}x_{i2} + (\beta_y\gamma_y v_{i0} + \beta_y v_{i1} + v_{i2}) \end{aligned}$$

which has only d_1, d_2, x_0, x_1, x_2 as the regressors, and the coefficients for d_1 and d_2 are still the same as in (3.2). But the problem is that it is hard to think of any instrument source, because all of x_0, x_1, x_2 appear on the right-hand side. If we apply the Least Squared Estimator (LSE) of y_2 on d_1, d_2, x_0, x_1, x_2 for this equation, is it possible to assume that the regressors are orthogonal to the error term? Unfortunately, even if x_t is unrelated to v_t at all leads and lags, this cannot hold due to the dependence of d_2 on y_1 and d_1 on y_0, because the error term $\beta_y\gamma_y v_0 + \beta_y v_1 + v_2$ includes the errors v_1 and v_0 for y_1 and y_0, respectively. Hence, both IVE and LSE fail. Surprisingly, however, *G estimation* of Robins (see Robins (1998,1999) and the references therein) is applicable in this case, as explained in the next section.

3.2 Comparison to Granger Causality

Granger non-causality of d_t on y_t is often tested by doing LSE and testing for $H_0 : \beta_{d1} = \beta_{d2} = 0$ in

$$y_{i2} = \beta_1 + \beta_{y1}y_{i1} + \beta_{y0}y_{i0} + \beta_{d1}d_1 + \beta_{d2}d_2 + \beta'_{x2}x_{i2} + \beta'_{x1}x_{i1} + \beta'_{x0}x_{i0} + v_{i2} \tag{3.3}$$

where no components of x_2 that are related to y_2 simultaneously should be included.

different from zero in child cognitive achievement regressions. Since some detailed current home inputs are missing there, it is not clear whether the earlier inputs merely capture the influence of the missing current inputs. Our regressions, using 25 current home inputs in addition to family and maternal variables, should mitigate the problems, if any, caused by excluding earlier inputs.

Granger causality is a probabilistic causality and it does not need the potential response concept; there are fundamental differences between probabilistic causality and potential-response-based 'counter-factual causality' as noted, e.g., in Holland (1986) among many others. In the former, the main interest is on testing for whether or not d_t causes y_t, and in the latter, the main interest is on finding the magnitude of the causal effect for a known cause (i.e., treatment). But if one tries to test for whether the effect magnitude is zero or not, then the two approaches become similar, as noted in Robins et al. (1999). Granger causality has been applied mostly to macro-economic data, but its application to micro-panel data can be seen in Holtz-Eakin et al. (1988,1989).

The similarity notwithstanding, a number of remarks are in order, comparing (3.1), (3.2), and (3.3). First, some of x_0, x_1, x_2 are excluded from (3.1) and (3.2), which is simply owing to our assumption of the contemporaneous covariate model, not due to anything intrinsic to the dynamic treatment effect framework. Second, the endogeneity issue of y_1, y_0, d_1, d_2 are ignored in (3.3), whereas this is tackled in (3.1) and (3.2). Third, even if the LSE is valid for (3.3), *one can test only for the direct effect of* d *in Granger causality*, because y_1 and y_0 are included in the model. This is the most important distinction between our approaches and the Granger causality as implemented by the LSE for (3.3).

The important distinction disappears, however, under the stationarity assumption $\gamma_d = \beta_{d2}$, with which the indirect effect $\beta_y \gamma_d$ is zero as well when the two direct effects β_{d1} and β_{d2} are zero. It is important to be aware that, *under the stationarity of equal contemporaneous effects, the equivalence between the counter-factual and Granger causalities holds only for the test* of non-causality. For the effect magnitude, (3.3) still misses the indirect effect. Inclusion of both y_1 and y_0 is not a distinguishing character of the Granger causality, for both may be included in (3.1) as well. It is the lack of awareness that the confounding by y_1 affecting both d_2 and y_2 is avoided by controlling for y_1, which then unfortunately misses the indirect effect of d_1 on y_2 through y_1. The appendix shows that, for three periods, the equivalence holds under an analogous—yet stronger for more periods are involved—stationarity condition.

4 G Algorithm

4.1 No Unmeasured Confounder Assumption

Define

$$X_2 \equiv (x_0', x_1', x_2')'.$$

Before we introduce G estimation and its requisite assumptions, we present structural form (SF) models for the treatments:

$$d_{i1} = \alpha_{11} + \alpha_{1x}' x_{i0} + \alpha_{1y} y_{i0} + \varepsilon_{i1} \quad \text{and} \quad d_{i2}^j = \alpha_{21} + \alpha_{2x}' x_{i1} + \alpha_{2y} y_{i1}^j + \varepsilon_{i2}; \tag{4.1}$$

d_2^j is the potential version of d_2 because d_2 depends on y_1^j. Also observe the y_2^{jk} reduced form (RF) obtained by removing y_1^j in the y_2^{jk} SF in the display preceding (3.1):

$$y_{i2}^{jk} = (\beta_1 + \beta_y \gamma_1) + (\beta_{d1} + \beta_y \gamma_d) j + \beta_{d2} k + \beta_y \gamma_y y_{i0} + \beta_y \gamma_x' x_{i1} + \beta_{x_2}' x_{i2} + (\beta_y v_{i1} + v_{i2}). \tag{4.2}$$

Define '$a \amalg b|c$' as the conditional independence of a and b given c. G estimation assumes '*no unobserved confounder' (NUC)*:

NUC (a) : $y_2^{jk} \amalg d_1 | (y_0, X_2)$ ($\iff$ $(\beta_y v_1 + v_2) \amalg \varepsilon_1 | (y_0, X_2)$ in view of y_2^{jk} RF and d_1 SF),

NUC (b) : $y_2^{jk} \amalg d_2^j \, | (d_1 = j, y_1^j, y_0, X_2)$ ($\impliedby$ $v_2 \amalg \varepsilon_2 | (\varepsilon_1, v_1, y_0, X_2)$ from y_2^{jk}, y_1^j SF and d_2^j SF).

In NUC (b),

$$(d_1 = j, y_1^j, y_0, X_2) \iff (\varepsilon_1 = j - \alpha_{11} - \alpha_{1x}' x_0 - \alpha_{1y} y_0, \; v_1, y_0, X_2).$$

Thus, conditioning on $(\varepsilon_1, v_1, y_0, X_2)$ is stronger than conditioning on this display because ε_1 is arbitrary in conditioning on $(\varepsilon_1, v_1, y_0, X_2)$, which explains '$\impliedby$' in NUC (b).

NUC (a) holds if d_1 is determined by (y_0, X_2) and some error term independent of y_2^{jk} given (y_0, X_2). NUC (b) holds if d_2^j is determined by $(d_1 = j, y_1^j, y_0, X_2)$ and some error term independent of y_2^{jk} given $(d_1 = j, y_1^j, y_0, X_2)$. NUC allows for dependence between $(\varepsilon_1, \varepsilon_2)$ and (v_1, v_2) through the conditioned variables; e.g., ε_2 may be related to v_2 through v_1. NUC (a) and (b) are nothing but 'selection-on-observables' where the observables are (y_0, X_2) and $(d_1 = j, y_1^j, y_0, X_2)$, respectively.

4.2 Main Integral for G Estimation

G estimation under NUC is

$$E(y_2^{jk}|y_0, X_2) = \int E(y_2|d_1 = j, d_2 = k, y_1, y_0, X_2) f(y_1|d_1 = j, y_0, X_2) \partial y_1 \qquad (4.3)$$

where $f(y_1|d_1 = j, y_0, X_2)$ denotes the conditional density. The important point is that the right-hand side is identified, and so is the conditional effect $E(y_2^{jk}|y_0, X_2)$. The equality holds because the right-hand side is

$$\begin{aligned}
& \int E(y_2^{jk}|d_1 = j, d_2^j = k,\ y_1^j, y_0, X_2) f(y_1^j|d_1 = j, y_0, X_2) \partial y_1^j \\
= & \int E(y_2^{jk}|d_1 = j,\ y_1^j, y_0, X_2) f(y_1^j|d_1 = j, y_0, X_2) \partial y_1^j \qquad \text{(due to NUC (b))} \\
= & E(y_2^{jk}|d_1 = j, y_0, X_2) \qquad \text{(for } y_1^j \text{ is integrated out)} \\
= & E(y_2^{jk}|y_0, X_2) \qquad \text{(due to NUC (a)).}
\end{aligned}$$

As to be discussed in the next section, for the IVE approaches, there occurs the issue of which covariates to include in the model. For instance, consider a covariate m_1 analogous to y_1 in its role that m_1 is affected by d_1 and affects d_2 and y_2. Then including m_1 in the y_2 equation leads to exactly the same problem as including y_1 does for (2.1): the indirect effect of d_1 through m_1 is missed. For G estimation, m_1 does not pose any problem in principle, because we can redefine of y_1 as (y_1, m_1) to apply (4.3). In practice, however, a multi-dimensional integration is needed when m_1 is included, which hence does pose a problem.

If d_1 were non-existent, then we would get

$$\begin{array}{ccc} d_2 & \longrightarrow & y_2 \\ \uparrow & \nearrow & \\ y_1 & & \end{array}$$

This is nothing but the static 'common factor' model with y_1 as an observed confounder. Also, NUC becomes $y_2^k \amalg d_2 | (y_1, X_2)$, which is the usual selection-on-observable condition for the one-shot treatment d_2. The G estimation gets reduced to

$$\int E(y_2^k|d_2 = k, y_1, X_2) f(y_1|X_2) \partial y_1 = \int E(y_2^k|y_1, X_2) f(y_1|X_2) \partial y_1 = E(y_2^k|X_2),$$

which is the usual static way of identifying $E(y_2^k|X_2)$ under the selection on observables. As X_2 gets integrated out for the total (marginal) effect eventually, instead of the G estimation, we can just use

$$\int E(y_2^k|d_2 = k, y_1, X_2)dF(y_1, X_2) = \int E(y_2^k|y_1, X_2)dF(y_1, X_2) = E(y_2^k)$$

where $F(y_1, X_2)$ is the distribution for $(y_1, X_2')'$. This shows that the G estimation is a dynamic generalization of the usual static selection-on-observable approach. Pearl (2000) reviews the graphical approach literature to causality and calls this—controlling the observables first and then integrating them out—'back door adjustment'.

When a linear structural model holds, one can estimate the dynamic treatment effect using LSE or IVE, and the same effect can be estimated with G estimation. Lee (2005) shows this in a simpler setting without covariates. With covariates, this is proven in the appendix for our two period model, further assuming 'NUC (c): $y_1^j \amalg d_1|(y_0, X_2)$'. For more on dynamic treatment effects in general, refer to Gill and Robins (2001) and Van der Laan and Robins (2003).

4.3 Simplification with Discrete Responses

Although the G estimation can be implemented nonparametrically in principle, estimating the conditional mean $E(y_2|d_1 = j, d_2 = k, y_1, y_0, X_2)$ and the conditional density $f(y_1|d_1 = j, y_0, X_2)$ nonparametrically and then integrating out y_1 is difficult in practice if the dimension of X_2 is large as in our data. Also in our data, the response variable is an ordinal behavior index, although it takes almost continuously many values. The linear models require cardinality of the variable. Hence, it may be sensible to turn the response variable into a binary or ordered response. Suppose we turn the original response into a binary response. In this case, the G estimation becomes

$$\begin{aligned} E(y_2^{jk}|y_0, X_2) &= P(y_2 = 1|d_1 = j, d_2 = k, y_1 = 0, y_0, X_2) \cdot P(y_1 = 0|d_1 = j, y_0, X_2) \\ +P(y_2 &= 1|d_1 = j, d_2 = k, y_1 = 1, y_0, X_2) \cdot P(y_1 = 1|d_1 = j, y_0, X_2). \end{aligned} \tag{4.4}$$

For our empirical analysis later, we will use this instead of (4.3).

Implementation of this G estimation is much easier. For instance, apply probit (or logit) to y_2 on d_1, d_2, y_1, y_0, X_2 to obtain the two probit probabilities in (4.4) for $y_2 = 1$:

$$\Phi(\psi_1 + \psi_{d1}d_1 + \psi_{d2}d_2 + \psi_{y1}y_1 + \psi_{y0}y_0 + \psi_x'X_2)$$

where the ψ-parameters are to be estimated. Also apply probit (or logit) to y_1 on d_1, y_0, X_2 to get the probit probabilities for $y_1 = 1$ (and $y_1 = 0$):

$$\Phi(\eta_1 + \eta_{d1} d_1 + \eta_{y0} y_0 + \eta_x' X_2)$$

where the η-parameters are to be estimated. Substituting these into (4.4) will do. A caution is warranted, however, as explained in the following.

When it holds that

$$E(y_1^j | y_0, X_2) = \Phi(\eta_1 + \eta_{d1} j + \eta_{y0} y_0 + \eta_x' X_2),$$

can we have

$$E(y_1 | d_1, y_0, X_2) = \Phi(\eta_1 + \eta_{d1} d_1 + \eta_{y0} y_0 + \eta_x' X_2) \ ?$$

In a similar context, Lee and Kobayashi (2001) show that such a replacement is equivalent to a selection-on-observable assumption. In the current context, what holds is, as proven in the appendix,

$$E(y_1 | d_1, y_0, X_2) = \Phi(\eta_1 + \eta_{d1} d_1 + \eta_{y0} y_0 + \eta_x' X_2) \iff E(y_1^j | d_1 = j, y_0, X_2) = E(y_1^j | y_0, X_2)$$

which is a selection-on-observables of d_1 for y_1^j. If d_1 is determined by (y_0, X_2) and some error term that is independent of y_1^j given (y_0, X_2), then this selection-on-observables holds.

An analogous question is, when it holds that

$$E(y_2^{jk} | y_1^j, y_0, X_2) = \Phi(\psi_1 + \psi_{d1} j + \psi_{d2} k + \psi_{y1} y_1^j + \psi_{y0} y_0 + \psi_x' X_2),$$

can we have

$$E(y_2 | d_1, d_2, y_1, y_0, X_2) = \Phi(\psi_1 + \psi_{d1} d_1 + \psi_{d2} d_2 + \psi_{y1} y_1 + \psi_{y0} y_0 + \psi_x' X_2) \ ?$$

This display can be shown to be equivalent to the selection-on-observables

$$E(y_2^{jk} | d_1 = j, d_2^j = k, y_1^j, y_0, X_2) = E(y_2^{jk} | y_1^j, y_0, X_2).$$

4.4 G Estimation with a Structural Nested Model

Instead of G estimation, there are other estimation methods available for dynamic causal inference as can be seen in Robins (1998,1999). But they are weighting-based estimators that deal with dynamic selection-on-observables by weighting; see Imbens (2004) or Lee (2005)

for exposition on the weighting idea. As shown in Frölich (2003) and Lee (2005), weighting estimators tend to be unstable, because some weights can be close to zero. A simple version in Robins (1992) of 'Structural Nested Model' that does not require weighting is available, and this will be applied to our data. An epidemiological application can be seen in Witteman et al. (1998) among others.

Suppose, for given covariates and some unknown parameter ψ_o, we have

$$y_2^{00} = y_2^{jk}\frac{\exp(\psi_o j)+\exp(\psi_o k)}{2} \iff y_2^{jk} = y_2^{00}\frac{2}{\exp(\psi_o j)+\exp(\psi_o k)}. \tag{4.5}$$

Here the treatments multiplicatively alter the no-treatment response y_2^{00}. For the spanking-behavior case with y being Behavior Problem Index (BPI; the lower the better), $\psi_o > 0$ means a good effect of spanking.

Recall NUC (b) that, conditional on the past spanking and input history, d_2 is independent of y_2^{jk}. Due to (4.5), d_2 should be independent of y_2^{00} as well. Defining

$$S_i(\psi) \equiv y_{i2}\frac{\exp(\psi d_{i1})+\exp(\psi d_{i2})}{2},$$

we get $S_i(\psi_o) = y_{i2}^{00}$. Thus, transforming the treatments into binary, the true value of θ in the following logit model should be zero if $\psi = \psi_o$:

$$P(d_2 = 1|y_1, y_0, d_1, X_2) = \frac{\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta S_2(\psi)\}}{1+\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta S_2(\psi)\}}. \tag{4.6}$$

Depending on ψ, we get different t-ratio $t_N(\psi)$ for θ. Following the well known duality between a test and the confidence interval (CI), a 95% CI for ψ is $\{\psi : |t_N(\psi)| < 1.96\}$. The middle point of the CI may be used as a point estimator $\hat{\psi}$ of ψ; alternatively, the ψ for zero θ estimate may be taken as a point estimate for ψ.

The main disadvantage of this simple structural nested model approach is the same effect restriction for d_1 and d_2 in (4.5) and the arbitrary functional form assumption linking all counter-factuals y_2^{jk} to y_2^{00}, but the main advantage—computational ease—is simply incomparable with the other dynamic causal effect estimators.

If desired, the same effect assumption in (4.5) can be relaxed: adopt, instead of $S_2(\psi)$ and (4.6)

$$S_2(\psi_0, \psi_1) \equiv y_2\frac{\exp(\psi_0 d_1)+\exp(\psi_1 d_2)}{2} \quad \text{and} \tag{4.7}$$

$$P(d_2 = 1|y_1, y_0, d_1, X_2) = \frac{\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta_1 S_2(\psi_0, \psi_1) + \theta_2 S_2(\psi_0, \psi_1)^2\}}{1+\exp\{\beta_2'(y_1, y_0, d_1, X_2') + \theta_1 S_2(\psi_0, \psi_1) + \theta_2 S_2(\psi_0, \psi_1)^2\}}$$

A 95% confidence region is $\{(\psi_0, \psi_1) : T_N < 5.99\}$ where T_N is an asymptotic χ^2_2 test. A point estimator for ψ_0 and ψ_1 may be obtained from the "center" of the region, but the concept of the center is ambiguous differently from the preceding single parameter case. To avoid this problem, we will set $\psi_1 = c\psi_0$ in our empirical analysis and then estimate ψ_0 from each fixed level of c. As c changes around one, the estimate for ψ_0 will change, showing how robust the result for (4.6) is as the assumption $\psi_1 = \psi_0$ gets relaxed.

5 Empirical Findings

5.1 Data Description

The NLSY79 child sample contains rich information on children born to the women respondents of the NLSY79. Starting from 1986, a separate set of questionnaires was developed to collect information about the cognitive, social, and behavioral development of the children of the NLSY79 respondents. The sets of child development results and inputs from birth up to age 10 were grouped in three: 0-2 years, 3-5 years, and 6-9 years. The variables include detailed home inputs as well as family backgrounds and some child care information.

Based on children surveyed from 1986 to 1998, we constructed a longitudinal sample of about 4700 children. In this full sample, there are 1329 children who have no missing values in the main variables of interest; this is our basic sample. We track these children for three survey rounds and get detailed information when they were at 2-3, 4-5, and 6-7 years old. Since severe spanking is likely to harm children and since most children are spanked modestly in frequency, our study will focus on the effects of mild to moderate spanking.[2] This motivates us to further restrict the sample to 961 children spanked up to three times a week before age three (73% of the whole sample) and up to five times a week before age five (94% of the whole sample); this is our main working sample on which most of our empirical analyses are based.

For children four years old and above, social and behavioral development is measured by the Behavior Problems Index Total Scores (BPI). BPI is one of the most frequently used variable in the NLSY79 child assessments for a wide range of child attitude and behavior. It is based on 28 questions in the Mother Supplement about specific behaviors that children of

[2] Indeed, the most hotly debated issue was and still is whether modest spanking works or not (Baumrind, Larzelere, and Cowan 2002).

age four and above may have exhibited in the previous three months. Mothers' responses to the individual items are then dichotomized and summed to produce an index for each child. In this recording process, each item answered "often true" or "sometimes true" is given a score of one, and "not true" zero. Thus, *a higher BPI represents more behavior problems.* In a fully representative sample of children, the mean standard score is expected to be 100. The BPI in our sample has mean 105.3 and standard deviation (SD) 14.7 around age 6-7, and mean 104.8 with SD 14.8 around age 4-5. Two binary variables are also constructed for BPI (1 if a child's BPI is higher than the sample mean and 0 otherwise).

Since there is no BPI for age below four, we use Motors and Social Development Scale (MSD) which measures developmental milestones in the areas of motor, cognitive, communication, and social development. The items were derived from standard measures of child development that are known to have high reliability and validity. Differently from BPI, however, *a higher MSD means better development.* MSD for children in our sample has mean 102.7 with SD 14.1, and MSD by age 2 will be used as a 'negative' proxy for BPI.

The frequency of spanking has been recorded when a child was around 2-3, 4-5, and 6-7 years old respectively. The survey question asks the mother "About how many times, if any, have you had to spank your child in the past week?" Spanking is quite common for young kids. In our data, over 90% were spanked by their mothers at least once before they reached age five. As children grew, the probability of being spanked dropped: 87% mothers spanked their toddlers at least once in the past week, but only 68% spanked their five year olds.

We also use a binary variable for spanking (1 if ever spanked and 0 otherwise). Since all children in the main working sample were not spanked more than several times a week, the most important difference among them may be not the exact spanking frequency, but whether or not ever spanked. Also, the construction of spanking variables is based on the reported spanking number in the past week when the mother was surveyed. So the reported values may not be the regular spanking frequency, and the binary variable indicating whether parents ever-spanked could be more reliable in reflecting a mother's disciplinary behavior. In most cases, the estimation results using both discrete and continuous versions of spanking are presented. The summary statistics of all variables in our basic sample are listed in Table A in the appendix, while some are listed in Table 1.

The link between spanking and behavior problems seems to be a complicated one, as it is still hotly debated after many years of investigation (Gershoff 2002, Deater-Deckard

and Dodge 1997). Children are heterogeneous in the first place, and there could be many unobserved heterogenous variables. Table 1 shows that white children are less likely to be spanked, and they have higher earlier development results and fewer behavioral problems. Boys are spanked slightly more often, having lower MSD and more behavior problems than girls. Firstborns are more likely to be spanked than others at age 2-3 but spanked slightly less at age 4-5; they have much higher MSD, more behavior problems at age 4-5, but almost the same BPI at age 6-7.

Detailed home inputs may matter much. Mothers who often read to their children at age 2-3 were less likely to spank them than those that did not; their children had better early development results and fewer behavior problems later. Similar patterns hold for children who have more books and less TV hours at home, who were breast-fed, and who have better home inputs in general. Mother's education does not seem to make much difference. Mothers with more than 12 years schooling in 1988 spanked their children only slightly more than those with less schooling.

Harmful effects of spanking may be over-estimated if detailed home inputs are not properly controlled, given that (already suggested by our data) a child spanked more may also lack other home inputs. The strength of our data is that a rich set of home inputs from birth up to age seven as well as key family background variables are available. This would reduce potential omitted variable biases. The age-specific Home Observation Measurement of the Environment variables (HOME), which is a simple summation of the dichotomized individual input item scores, is often used in child development research as an aggregate quality indicator of home environment. The completion rates of HOME, however, are in general very low for children under age four, which causes many missing values. Whenever possible, HOME is included as a control in addition to the detailed home inputs.

Most home input variables are categorical with multiple levels, which are then converted to dummy variables. The home environment variables are age-specific, where there are 25 home inputs at age 6-7, 18 inputs at age 4-5, and 10 inputs at age 2-3. These inputs include how many books a child has, how often the mother reads to the child, how often the father plays with the child outdoors, whether there are musical instruments and newspapers at home, whether the parents encourage hobbies and bring the child to enriching activities such as visiting museums, how often the child gets together with relatives and friends, how often the child watches TV, and how the mother responds to tantrums. When the sample

size allows (given missing values), child care attendance at age 0-3, mother prenatal care variables and her working hours before child birth are controlled as well.

5.2 Empirical Results

5.2.1 IVE for the Structural Linear Models

We first estimate the structural linear models as in (3.1), where y_2 and y_1 are BPI at age 6-7 and 4-5; d_2 and d_1 are the spanking frequencies at age 4-5 and 2-3, or their binary versions (ever spanking or not). Under the assumption that only the current inputs are related to current behavior problems, we use past inputs as instrumental variables. Note the current home inputs include disciplinary and parenting variables such as the number of times mother grounded child, took away TV or allowance, sent child to room, praised child, showed physical affection, and said positive things are controlled unless otherwise noted. The empirical results are in Table 2.

The first column 'IV' presents results with detailed current inputs at age 6-7 as controls, and detailed inputs at age 2-3 and 4-5 as instrumental variables. The second column 'IV(B)' adds family background variables. The same set of inputs is included in the third column 'IV(B')', where the exact numbers of spanking and their squared terms are used. The effects of spanking at age 2-3 are negative across the three specifications, though none is statistically significant. Similar, though weaker, results apply to spanking at age 4-5. This pattern is robust to changes in the detailed inputs used as controls.

Wald tests show that these IV results are not significantly different from their LSE counterparts. For this reason, the two columns 'LSE (B,D)' and 'LSE(D)' are presented for LSE. Both include two HOME scores at age 2-3 and 4-5, which have many missing values (hence smaller sample sizes). In the model where the exact numbers of spankings are used, the sample size is increased by including kids spanked up to five times a week at age 2-3. The effects of spanking at age 2-3 are negative and significant in the LSE results, and their magnitudes are similar to the IV estimates. In comparison, the effects of spanking at age 4-5 are never significant, again similar to the IV results.

The next four columns show the regression results for BPI at age 4-5. The first two IV results ('IV' and 'IV") use the same sample including HOME at age 2-3 but no family background variables; the third IV model 'IV(B)' replaces HOME with family backgrounds

variables. The coefficients of spanking at age 2-3 are negative and have similar magnitudes across the different samples and specifications. In the final column 'LSE (B,T)', LSE is done using three measures of child temperament instead of MSD, because these three variables may make a better proxy than MSD for BPI. Due to missing values in these measures, the sample size drops so much so that no sensible IV regressions can be done. The coefficient of spanking is negative and significant with a much higher magnitude. This result is robust to including family backgrounds and using the exact spanking frequencies.

Based on the pair of IV regressions with family backgrounds and binary spanking variables (IV(B) at age 6-7 and 4-5), we calculate the total effect of spanking. Modest spanking at age 2-3 reduces BPI at age 6-7 by 4.03 points, while modest spanking at age 4-5 increases it by 1.42 points. These are the direct effects, because BPI score at age 4-5 is controlled. As the regression results of IV(B) for age 4-5 BPI show, modest spanking at age 2-3 reduces BPI at age 4-5 by 4.07 points, while the effect of BPI at age 4-5 is 0.52 on BPI at age 6-7. So the indirect effect of spanking at age 2-3 on the child's age 6-7 BPI is $0.52 \times (-4.07) = -2.12$, which is about half the direct effect (-4.03) in magnitude. Taken together, the effect of d_1 is

$$\textit{direct effect} + \textit{indirect effect through } y_1 : \widehat{\beta}_{d1} + \widehat{\beta}_y \widehat{\gamma}_d = -4.03 + 0.52 \times (-4.07) = -6.15,$$

which is 42% of SD(BPI). The bootstrap bias-corrected 95% CI is $[-59.3, 6.5]$. Since this includes zero, 'H_0 : no d_1 effect' is not rejected, but the interval is nine times longer to the negative side. Although the indirect effect may look small being only one half the direct effect, it can accumulate over time in the long run, leading to a substantial magnitude. The total effect of $d = (d_1, d_2)'$ and its bootstrap bias-corrected 95% CI are

$$-6.15 + 1.42 = -4.73 \quad \text{and} \quad [-37.7, 13.6].$$

Based on results using the exact spanking frequencies, another set of estimates can be calculated. For example, in IV(B'), the quadratic function of d_2 (for y_2) is $-1.6d_2 + .78{d_2}^2$, which is negative for $d_2 \leq 2$ and positive otherwise. The quadratic function of d_1 (for y_2) is $-3.11d_1 + .62{d_1}^2$, which is negative for $d_1 \leq 3$. In IV', the quadratic function of d_1 (for y_1) $-8.79d_1 + 2.31d_1^2$, also negative for $d_1 \leq 3$. Now using the first derivatives, the total effect of spanking at age 2-3 on BPI at age 6-7 is

$$-3.11 + 2 \times 0.62d_1 \ + 0.48(-8.79 + 2 \times 2.31d_1) = -7.33 + 3.46d_1.$$

This is clearly greater in magnitude than the effect of d_2 on y_2 that is $-1.6 + 1.56d_2$. Modest spanking seems to reduce a child's behavior problems as the negative 'intercepts' indicate, but too much spanking is harmful as the positive 'slopes' show.

5.2.2 G Estimation with Discrete Responses

In order to apply the simplified G estimation with discrete responses, we convert the two BPIs to dummy variables (higher than the sample mean or not). The binary spanking variables (ever spanked or not) are used as well to obtain the total effects with ease. The probit results are shown in Table 3, where the entries are the estimated marginal effects calculated at the sample means of the control variables (i.e., the derivatives of $P(y_2 = 1|\cdots)$ evaluated at the variable sample averages). The probit is the discrete analog of the dynamic panel data model (but no unit-specific effect is considered in the probit), and as such, it misses the indirect effects. The estimates for the covariates are omitted as in Table 2. We also tried logit instead of probit, but the logit results are omitted, for they differ little from the probit results.

The first column includes as controls detailed home inputs from birth up to age 6-7 as well as family background variables. Modest spanking at age 2-3 reduces the probability of higher-than-average BPI at age 6-7 by 0.35, which is significant at a 10% level; spanking at age 4-5 increases the same probability by 0.07, but this is not significant. Higher-than-average BPI at age 4-5 increases the probability of higher-than-average BPI at age 6-7 by 0.46. The sample size is small due to missing values especially in early inputs at age 2-3 and family background variables. The second column reports results excluding these variables. The coefficient of spanking at age 2-3 is still negative and significant at a p-value 5.6%, but its level is reduced to -0.23. The explanatory power is also reduced, while the other results are very similar. The same trend continues in column three where the sample size increases further by taking out the disciplinary inputs at age 6-7, which are likely to be affected by BPI. Overall, the general pattern is that modest spanking at age 2-3 reduces BPI at age 6-7, while spanking at age 4-5 tends to increase BPI. The latter effect, however, is not significant. The effects of BPI at age 4-5 are always positive and significant.

The probit results for BPI at age 4-5 are presented in the second part of the table. In these results, child temperament measures as well as MSD are used to control for a child's initial characteristics. The results in the three columns vary with different controls: the

first column includes detailed inputs from birth up to age 4-5, the second column adds family background variables, and the third column uses variables on whether a child attended regular child care in the first, second, and third year after birth. The coefficients of spanking are very similar across these specifications: modest spanking reduces the probability of higher than average BPI at age 4-5 by 0.44, which is negative and significant at a 10% level. The results (not reported) are also similar when disciplinary inputs at age 4-5 are excluded.

The desired total effect using (4.4) can be obtained with estimates in columns 'Probit' and 'Probit (T)' in Table 2: the total effect of spanking at both age 2-3 and 4-5 is

$$E(y_2^{11}) - E(y_2^{00}) = 0.047$$

with the bootstrap bias-corrected 95% CI $[-0.4, 0.48]$. It can be decomposed into two parts: the effect of spanking at age 4-5 (conditional on spanking at age 2-3) $E(y_2^{11}) - E(y_2^{10}) = 0.16$, with the bootstrap CI $[-0.14, 0.30]$; and the effect of spanking at age 2-3 (conditional on no spanking at age 4-5) $E(y_2^{10}) - E(y_2^{00}) = -0.12$, with the bootstrap CI $[-0.64, 0.35]$. Unfortunately, all CI's include zero. A possible reason for this is that the subgroup with no spanking at age 2-3 is very small when relevant inputs are controlled. This suggests that modest spanking at age 2-3 reduces a child's behavior problems at age 6-7, while spanking at age 4-5 tends to slightly increase the problems measured at age 6-7. This opposite pattern was noted also in the IVE results.

5.2.3 Simple Structural Nested Model

Another set of estimates obtained using the structural nested model is in Table 4. The regressors in the logit for (4.6) include detailed home inputs at age 4-5 and 2-3. In the second row, measures of child temperament are also included, and in the third row family backgrounds variables are further added. The point estimate $\widehat{\psi_0}$ increases from 0 to 0.04 across the specifications as more controls are added. The number $\widehat{\psi_0} = 0.04$ corresponds to 4.3 points reduction on average (about 30% reduction of a standard deviation) of BPI at age 6-7. This level is similar to those obtained using the IV methods above.

Since our earlier results suggest that the effects of spanking vary at different ages, we allow $\widehat{\psi_1} = c\widehat{\psi_0}$ and explore the corresponding effects using the third row logit model mentioned just above, where $\widehat{\psi_0}$ still indicates the effect of spanking at age 2-3, $\widehat{\psi_1}$ indicates the effect of spanking at age 4-5, and c is a positive number. The estimated ψ_0 varies from 0.20

to 0.01 as c changes from 1/4 to 4, corresponding to a range of average points reduced from 12.64 to 2.11 on BPI at age 6-7.

5.2.4 Granger Causality

Table 5 presents the results for the Granger causality model (3.3). The binary version of spanking variables is used to ease comparison with earlier results, while the third column also presents results using the exact numbers of spanking. The various specifications differ mainly in the set of control variables used. In the first column, all inputs from birth up to age seven are controlled, whereas the current disciplinary inputs are excluded in the other columns since they may be affected by the current BPI. With lagged BPI controlled for, still the lagged spanking is significant, and thus *Granger non-causality is rejected*. In this case, as noted already, the coefficients of d_1 and d_2 show only their direct effects at best, which should be borne in mind in the following interpretation.

The coefficients of spanking at age 2-3 are always negative and significant in these specifications; the effects of spanking at age 4-5 are also negative, though not often significant. Their magnitudes are similar to the IV estimators in table 2. In the third column where the exact spanking frequencies are used, the effects of spanking are concave with significant estimates. When family background variables are included in the fourth column, the coefficients of spanking become slightly larger than those in column two. The last column has the most comprehensive controls, including child temperament measures at age 2-3 as well as family background variables. The coefficients of the two spanking variables are both negative and significant at p value 0.05, with the highest levels among the specifications listed in the table.

6 Conclusions

In this paper, when a treatment is repeated over time and the final response is measured at the end, we showed how to estimate dynamic treatment effects with IVE applied to linear structural models. In our approach, early treatments are allowed to have an immediate (direct) effect as well as a lingering (indirect) effect through interim responses; also, interim treatments are allowed to be affected by interim responses. These two facts pose a dilemma to the usual dynamic model approach: if the interim responses are not controlled, then they become a confounder, because the treatment and control groups differ systematically in the

interim responses; otherwise, the indirect effects are missed. An extreme form of this can be seen in the usual Granger causality model where all interim responses are controlled and consequently all indirect effects are missed. Nonetheless, we showed that, when the hypothesis of no causality is not rejected, the Granger non-causality inference is valid under a stationary effect assumption. We also showed that our approach of IVE for linear structural models identifies the same total effect of the entire treatment 'profile' as 'G estimation' does; G estimation has been proposed as an innovative way of estimating dynamic treatment effects in epidemiology and biostatistics.

The IVE approach and two practical versions of G estimation were applied to an important issue: the effect of spanking on child behavior problems. The empirical results, though varying across different estimation methods, consistently indicate that moderate spanking works, and spanking at an early age 2-3 has a stronger effect on reducing later behavior problems at age 6-7 than spanking at age 4-5, which is a surprising finding. Our preferred estimate suggests the overall effect (including direct and indirect effects) of spanking at age 2-3 on average reduces 42% of one standard deviation of Behavior Problems Index Total Scores (BPI) at age 6-7. The direct effect of spanking at age 2-3 estimated by Granger causality models ranges from 31% to 46% of one standard deviation in reduction of BPI at age 6-7 as more controls are added. In comparison, the estimated effects of spanking at age 4-5 are small and often ambiguous in sign. These results seem at odds with prevailing findings in the psychology literature where the empirical findings are not backed by a proper causal framework. We hope our approach to be applied to other dynamic causal relations which are widely seen in economics, micro or macro. This will be taking one step further from the simple Granger causality analysis toward the full causal analysis allowing for feedbacks from interim responses.

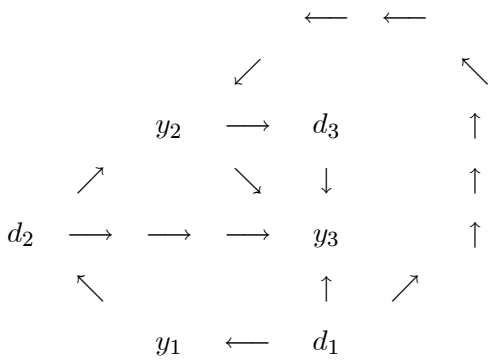
APPENDIX

Extension of Structural IVE to Three Periods/Treatments

Extending two periods to three, the observation sequence is now

$$(x_0, y_0),\ (d_1, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}),\ (d_2, \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}),\ (d_3, \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}),$$

and the treatment profile becomes $d = (d_1, d_2, d_3)'$. The last response y_3 is the response of interest with its potential version y_3^{jkl}. The desired effect is $E(y_3^{jkl} - y_3^{000})$. The following figure shows the three-period direct and indirect effects:



Consider linear contemporaneous-covariate models:

$$\begin{aligned}
y_{i1}^{j} &= \gamma_{11} + \gamma_{1d1} j + \gamma_{1y} y_{i0} + \gamma_{1x}' x_{i1} + v_{i1}, \\
y_{i2}^{jk} &= \gamma_{21} + \gamma_{2d1} j + \gamma_{2d2} k + \gamma_{2y} y_{i1}^{j} + \gamma_{2x}' x_{i2} + v_{i2}, \\
y_{i3}^{jkl} &= \beta_1 + \beta_{d1} j + \beta_{d2} k + \beta_{d3} l + \beta_y y_{i2}^{jk} + \beta_x' x_{i3} + v_{i3}.
\end{aligned}$$

The notations differ somewhat from the two period case, for y_3 is the final response.

The y_2^{jk} RF with y_1^{j} removed is

$$\begin{aligned}
y_{i2}^{jk} &= (\gamma_{21} + \gamma_{2y}\gamma_{11}) + (\gamma_{2d1} + \gamma_{2y}\gamma_{1d1}) j + \gamma_{2d2} k \\
&\quad + \gamma_{2y}\gamma_{1y} y_{i0} + \gamma_{2y}\gamma_{1x}' x_{i1} + \gamma_{2x}' x_{i2} + (\gamma_{2y} v_{i1} + v_{i2}).
\end{aligned}$$

The y_3^{jkl} 'semi-RF' with only y_2^{jk} removed ('semi-RF' because y_{i1}^{jk} appears) is

$$\begin{aligned}
y_{i3}^{jkl} &= (\beta_1 + \beta_y\gamma_{21}) + (\beta_{d1} + \beta_y\gamma_{2d1}) j + (\beta_{d2} + \beta_y\gamma_{2d2}) k + \beta_{d3} l \\
&\quad + \beta_y\gamma_{2y} y_{i1}^{j} + \beta_y\gamma_{2x}' x_{i2} + \beta_x' x_{i3} + (\beta_y v_{i2} + v_{i3}).
\end{aligned}$$

Also, the y_3^{jkl} RF with both y_{i2}^{jk} and y_{i1}^{j} removed is

$$\begin{aligned} y_{i3}^{jkl} &= \{\beta_1 + \beta_y(\gamma_{21} + \gamma_{2y}\gamma_{11})\} + \{\beta_{d1} + \beta_y(\gamma_{2d1} + \gamma_{2y}\gamma_{1d1})\}j + (\beta_{d2} + \beta_y\gamma_{2d2})k + \beta_{d3}l \\ &\quad + \beta_y\gamma_{2y}\gamma_{1y}y_{i0} + \beta_y\gamma_{2y}\gamma_{1x}'x_{i1} + \beta_y\gamma_{2x}'x_{i2} + \beta_x'x_{i3} + (\beta_y\gamma_{2y}v_{i1} + \beta_y v_{i2} + v_{i3}). \end{aligned}$$

This shows five effects to be identified:

$$\begin{aligned} \text{direct and indirect (through } y_1, y_2\text{) effects of } d_1 &: \beta_{d1},\ \ \beta_y(\gamma_{2d1} + \gamma_{2y}\gamma_{1d1}) \\ \text{direct and indirect (through } y_2\text{) effects of } d_2 &: \beta_{d2},\ \ \beta_y\gamma_{2d2} \\ \text{direct effect of } d_3 &: \beta_{d3}. \end{aligned}$$

The first-lag response model IVE for these effects are

- Step 1: estimate γ_{1d1} in the y_1 equation with regressors (d_1, y_0, x_1); x_0 provides the instrument source for d_1 and y_0.
- Step 2: estimate γ_{2d1}, γ_{2d2}, and γ_{2y} in the y_2 equation with regressors (d_1, d_2, y_1, x_2); x_0 and x_1 are the instrument source for d_1, d_2, and y_1.
- Step 3: estimate β_{d1}, β_{d2}, β_{d3}, and β_y in the y_3 equation with regressors $(d_1, d_2, d_3, y_2, x_3)$; x_0, x_1, and x_2 the instrument source for d_1, d_2, d_3, and y_2.

Imposing the equal contemporaneous effect assumption

$$\gamma_{1d1} = \gamma_{2d2}$$

that the effect of d_1 on y_1 is the same as the effect of d_2 on y_2, there is no need to estimate the y_1 equation and two IVE's will do, instead of three in the first-lag response model IVE. There is no more problem of finding instruments for the y_1 equation.

Going further, strengthen $\gamma_{1d1} = \gamma_{2d2}$ to

$$\beta_{d3} = \gamma_{1d1} = \gamma_{2d2}, \quad \gamma_{2y} = \beta_y, \quad \gamma_{2d1} = \beta_{d2}.$$

Under this, we just have to estimate the y_3 equation, and it holds that

$$d_1 \text{ effect } \beta_{d1} + \beta_y(\beta_{d2} + \beta_y\beta_{d3}), \quad d_2 \text{ effect } \beta_{d2} + \beta_y\beta_{d3}, \ \ d_3 \text{ effect } \beta_{d3}.$$

Since only the y_3 equation is estimated, finding instruments becomes even more easier. This display shows that the Granger non-causality test becomes equivalent to our approach under

the strengthened set of stationarity assumptions, because all indirect effects are zero when $\beta_{d1} = \beta_{d2} = \beta_{d3} = 0$.

Turning to the last-lag response IVE, consider the observed version of the above y_3^{jkl} RF with y_{i2}^{jk} and y_{i1}^{j} removed; only y_0 is left as a lagged response on the right-hand side. The observed version has regressors $(d_1, d_2, d_3, y_0, x_1, x_2, x_3)$. The instrument source for d_1, d_2, d_3, y_0 is x_0. This last-lag response IVE is a single step method as in the two period case.

Covariate Choice and Nonlinear Functions of Treatments

Recall $(x_0, y_0),\ (d_1, x_1, y_1),\ (d_2, x_2, y_2)$. In our data, there is no known time order between x_t and y_t; with temporal aggregation, x_t and y_t can be simultaneously related in the data. This raises the issue of which covariates to include in the y_1 and y_2 equations. A component w_1 of x_1 may be affected by y_1 or d_1. In such a case, should w_1 be still included in x_1? We examine this issue here, assuming that w_1 *affects* y_1; if w_1 affects y_2 but not y_1, w_1 should be put into x_2 in our contemporaneous-covariate model; if w_1 does affect neither y_1 nor y_2, then w_1 can be simply ignored. Related to the covariate choice problem is including nonlinear functions of treatments, which is also discussed here. Allowing nonlinearity matters, because excessive spanking can be devastating, even if a moderate spanking is good; at least a quadratic functional form of spanking is called for.

First, suppose that w_1 is affected by d_1, but not by y_1. If w_1 is included in x_1, then the indirect effect $d_1 \to w_1 \to y_1$ is missed because w_1 is controlled; if interested only in the direct effect, however, then including w_1 in x_1 is all right. If we choose not to include w_1 in x_1 to avoid this problem, then we may incur another problem as w_1 may become a confounder, e.g., by affecting d_2 and y_2 as in the two-period effect diagram. To rule out such possibility, we control w_1. For instance, suppose that w_1 is 'reading (books) to children'. A parent may do this because of a guilty feeling after spanking (hence w_1 is affected by d_1), which then influences y_1, and possibly y_2 and d_2 as well. Controlling for reading-to-children entails missing this indirect effect. But not controlling for it may entail confounding. If we are to err, it is safer to err to fall on the conservative side of omitting the indirect effect but still getting the direct effect right, rather than falling on not getting any effect right by not controlling for w_1. In our empirical analysis, we thus include variables such as reading-to-children in the y_1 equation, taking one of the two following positions: either there is no w_1 affected by d_1, or if there is such a w_1, then we are not interested in the indirect effect. When spanking effects

under these assumptions are announced to the public, one can imagine the 'official caveat' that the spanking effect estimates are those without any subsequent spanking-mitigating behaviors such as reading to children (or taking children to a theme-park).

Second, suppose that w_1 is affected by y_1. In this case, w_1 gets simultaneously related to y_1 and becomes an endogenous regressor in the y_1 equation. For instance, various disciplinary measures (e.g., grounding or taking away allowances) other than spanking can be simultaneously related to y_1 (due to the temporal aggregation). Recall that this simultaneity problem does not occur with d_1, as we constructed our data such that d_1 precedes w_1 and y_1. The best way to handle such a w_1 is setting up a bivariate response model where (w_1, y_1) becomes a bivariate response vector. In the y_1 SF, the coefficient of d_1 shows only the direct effects (as if an intervention on d_1 is accompanied by an intervention on w_1). In the y_1 RF with w_1 substituted out, the coefficient of d_1 shows the total effect. For instance, suppose

$$y_1 = \alpha_w w_1 + \alpha_d d_1 + u, \quad w_1 = \beta y_1 + \varepsilon \implies y_1 = \frac{\alpha_d}{1-\alpha_w\beta} d_1 + \frac{u + \alpha_w \varepsilon}{1-\alpha_w\beta} \quad \text{where } |\alpha_w\beta| < 1.$$

In words, an initial change in d_1 causes a change in y_1 of magnitude α_d, but the change in y_1 leads to a change in w_1 of magnitude β, which in turn changes y_1 and so on. The y_1 RF includes this exchange between y_1 and w_1.

In our empirical analysis, we try both including and excluding the disciplinary measures. Including those variables and estimating the y_1 SF with IVE means that the estimated effect of d_1 is only the direct effect without any other disciplinary measures taken to complement or substitute d_1—of course, if desired, the indirect effect can be recovered using the w_1-SF. Excluding those variables means that we are estimating the y_1 RF from the bivariate response model where the total effect of d_1 gets estimated.

The same issue of covariate choice arises for the y_2 equation. In principle, one just have to follow the same model as used for the y_1 equation but augmented by d_2 now, although this could not be done exactly with our data as different sets of variables were available for the y_1 and y_2 equations.

Even if spanking is beneficial, too much spanking is likely to be harmful. That is, nonlinear effects of spanking ought to be taken into account. For this, suppose that the effect

of d_1 and d_2 are quadratic (the subscript q in the following stands for 'quadratic'):

$$\begin{aligned}
y_{i1}^{j} &= \gamma_1 + \gamma_d j + \gamma_{dq} j^2 + \gamma_y y_{i0} + \gamma_x' x_{i1} + v_{i1}, \\
y_{i2}^{jk} &= \beta_1 + \beta_{d1} j + \beta_{d1q} j^2 + \beta_{d2} k + \beta_{d2q} k^2 + \beta_y y_{i1}^{j} + \beta_x' x_{i2} + v_{i2} \\
&= \beta_1 + \beta_{d1} j + \beta_{d1q} j^2 + \beta_{d2} k + \beta_{d2q} k^2 \\
&\quad + \beta_y(\gamma_1 + \gamma_d j + \gamma_{dq} j^2 + \gamma_y y_{i0} + \gamma_x' x_{i1} + v_{i1}) + \beta_x' x_{i2} + v_{i2} \\
&= (\beta_1 + \beta_y \gamma_1) + \{\beta_{d1} j + \beta_{d1q} j^2 + \beta_y(\gamma_d j + \gamma_{dq} j^2)\} + (\beta_{d2} k + \beta_{d2q} k^2) \\
&\quad + \beta_y \gamma_y y_{i0} + \beta_y \gamma_x' x_{i1} + \beta_x' x_{i2} + (\beta_y v_{i1} + v_{i2}).
\end{aligned}$$

With the first derivatives, the three key effects are

$$\begin{aligned}
\text{direct and indirect effects of } d_1 &= j: \ \beta_{d1} + 2\beta_{d1q} j, \ \ \beta_y(\gamma_d + 2\gamma_{dq} j) \\
\text{direct effect of } d_2 &= k: \ \beta_{d2} + 2\beta_{d2q} k.
\end{aligned}$$

These can be identified in two steps with the first-lag response IVE:

- Step 1: estimate γ_d, γ_{dq} in the y_1 equation with regressors d_1, d_1^2, y_0, x_1.
- Step 2: estimate $\beta_{d1}, \beta_{d1q}, \beta_{d2}, \beta_{d2q}$, and β_y in the y_2 equation with regressors $d_1, d_1^2, d_2, d_2^2, y_1, x_2$.

Alternatively, we may estimate the following last-lag response model (from the above y_2^{jk} equation) with a single IVE:

$$\begin{aligned}
y_2 &= (\beta_1 + \beta_y \gamma_1) + (\beta_{d1} + \beta_y \gamma_d) d_1 + (\beta_{d1q} + \beta_y \gamma_{dq}) d_1^2 + \beta_{d2} d_2 + \beta_{d2q} d_2^2 \\
&\quad + \beta_y \gamma_y y_{i0} + \beta_y \gamma_x' x_{i1} + \beta_x' x_{i2} + (\beta_y v_{i1} + v_{i2}).
\end{aligned}$$

Going one step further, we can expand the nonlinearity to cubic terms. For instance, with the subscript c standing for 'cubic'

$$\begin{aligned}
y_{i1}^{j} &= \gamma_1 + \gamma_d j + \gamma_{dq} j^2 + \gamma_{dc} j^3 + \gamma_y y_{i0} + \gamma_x' x_{i1} + v_{i1}, \\
y_{i2}^{jk} &= \beta_1 + \beta_{d1} j + \beta_{d1q} j^2 + \beta_{d1c} j^3 + \beta_{d2} k + \beta_{d2q} k^2 + \beta_{d2c} k^3 + \beta_y y_{i1}^{j} + \beta_x' x_{i2} + v_{i2}.
\end{aligned}$$

This yields

$$\begin{aligned}
\text{direct and indirect effects of } d_1 &= j: \ \beta_{d1} + 2\beta_{d1q} j + 3\beta_{d1c} j^2, \ \ \beta_y(\gamma_d + 2\gamma_{dq} j + 3\gamma_{dc} j^2) \\
\text{direct effect of } d_2 &= k: \ \beta_{d2} + 2\beta_{d2q} k + 3\beta_{d2c} k^2.
\end{aligned}$$

Having seen nonlinear treatment effects, one may consider a nonlinear function of y_1^j in the y_2^{jk} equation, but we will not accommodate this possibility. One reason is that nonlinear lagged response variables are rarely used in economic models. Another reason is that even a quadratic function of y_1^j in the y_2^{jk} equation combined with a quadratic treatment effect results in fourth order polynomials of d_1 and such a function will not be identified easily in practice.

Proof for G estimation Identifying Total Effect in Two Periods

G estimation does not require any functional form specification. But it is instructive to verify that G estimation identifies the same total effect as the SF linear model (the y_2^{jk} equation before (3.1)) identifies. Observe, in the y_2^{jk} equation,

$$\begin{aligned} E(y_2^{jk}|d_1 &= j, d_2^j = k,\ y_1^j, y_0, X_2) = E(y_2^{jk}|d_1 = j,\ y_1^j, y_0, X_2) \\ &= \beta_1 + \beta_{d1}j + \beta_{d2}k + \beta_y y_1^j + \beta_x' x_2 + E(v_2|d_1 = j,\ y_1^j, y_0, X_2) \ \text{ owing to NUC (b).} \end{aligned}$$

Substitute this into the display following (4.3) to get

$$\begin{aligned} \beta_1 + \beta_{d1}j + \beta_{d2}k + \beta_y E(y_1^j|d_1 &= j, y_0, X_2) + \beta_x' x_2 + E(v_2|d_1 = j, y_0, X_2) \\ &= \beta_1 + \beta_{d1}j + \beta_{d2}k + \beta_y E(y_1^j|y_0, X_2) + \beta_x' x_2 + E(v_2|y_0, X_2), \end{aligned}$$

owing to NUC (a) $E(y_2^{jk}|d_1 = j, y_0, X_2) = E(y_2^{jk}|y_0, X_2)$. Substitute $E(y_1^j|y_0, X_2) = \gamma_1 + \gamma_d j + \gamma_y y_0 + \gamma_x' x_1 + E(v_1|y_0, X_2)$ to have (4.3) become

$$\begin{aligned} & \beta_1 + \beta_{d1}j + \beta_{d2}k + \beta_y\{\gamma_1 + \gamma_d j + \gamma_y y_0 + \gamma_x' x_1 + E(v_1|y_0, X_2)\} + \beta_x' x_2 + E(v_2|y_0, X_2) \\ = \ & (\beta_1 + \beta_y\gamma_1) + (\beta_{d1} + \beta_y\gamma_d)j + \beta_{d2}k + \beta_y\gamma_y y_0 + \beta_y\gamma_x' x_1 + \beta_x' x_2 + E(\beta_y v_1 + v_2|y_0, X_2). \end{aligned}$$

From this,

$$E(y_2^{jk}|y_0, X_2) - E(y_2^{00}|y_0, X_2) = (\beta_{d1} + \beta_y\gamma_d)j + \beta_{d2}k.$$

Therefore, the G estimation identifies the same total effect as the SF linear model does.

Proof for Replacing Fixed Treatment with Random Treatment in Two Periods

For y_1^j, under $E(y_1^j|y_0, X_2) = \Phi(\eta_1 + \eta_{d1}j + \eta_{y0}y_0 + \eta_x' X_2)$, we will prove

$$E(y_1|d_1, y_0, X_2) = \Phi(\eta_1 + \eta_{d1}d_1 + \eta_{y0}y_0 + \eta_x' X_2) \iff E(y_1^j|d_1 = j, y_0, X_2) = E(y_1^j|y_0, X_2).$$

First, suppose that the left-hand side holds. Then

$$\begin{aligned} E(y_1^j|d_1 &= j, y_0, X_2) = E(y_1|d_1 = j, y_0, X_2) \quad \text{because } y_1 = y_1^j \text{ given } d_1 = j \\ &= \Phi(\eta_1 + \eta_{d1} j + \eta_{y0} y_0 + \eta_x' X_2) = E(y_1^j|y_0, X_2). \end{aligned}$$

Hence the right-hand side holds. Second, to prove the reverse, suppose $E(y_1^j|d_1 = j, y_0, X_2) = E(y_1^j|y_0, X_2)$. Observe

$$\begin{aligned} E(y_1|d_1, y_0, X_2) &= \int E(y_1|j, y_0, X_2)\partial 1[d_1 \leq j] = \int E(y_1^j|j, y_0, X_2)\partial 1[d_1 \leq j] \\ &= \int E(y_1^j|y_0, X_2)\partial 1[d_1 \leq j] = \int \Phi(\eta_1 + \eta_{d1} j + \eta_{y0} y_0 + \eta_x' X_2)\partial 1[d_1 \leq j] \\ &= \Phi(\eta_1 + \eta_{d1} d_1 + \eta_{y0} y_0 + \eta_x' X_2). \end{aligned}$$

REFERENCES

Angrist, J.D. and A.B. Krueger, 1999, Empirical strategies in labor economics, in Handbook of Labor Economics 3A, edited by O. Ashenfelter and D. Card, North-Holland.

Angrist, J.D. and A.B. Krueger, 2001, Instrumental variables and the search for identification: from supply and demand to natural experiments, Journal of Economic Perspectives 15, 69-85.

Baumrind, D., Larzelere, R. E., and Cowan, P. A., 2002, Ordinary physical punishment: Is it harmful? Comment on Gershoff, Psychological Bulletin 128, 580–589.

Bowles, S., H. Gintis, and M. Osborne, 2001, The determinants of earnings: a behavioral approach, Journal of Economics Literature 39, 1137-1176.

Deater-Deckard, K., and Dodge, K. A., 1997, Externalizing behavior problems and discipline revisited: Nonlinear effects and variation by culture, context, and gender, Psychological Inquiry 8, 161–175.

Frölich, M., 2003, Programme evaluation and treatment choice, Springer-Verlag.

Gershoff, E., 2002, Corporal punishment by parents and associated child behaviors and experiences: a meta-analytic and theoretical review, Psychological Bulletin 128, 539–579.

Gill, R. and J.M. Robins, 2001, Causal inference for complex longitudinal data: the continuous case, Annals of Statistics 29, 1785-1811.

Granger, C.W.J., 1969, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37, 424-438.

Granger, C.W.J., 1980, Testing for causality: a personal viewpoint, Journal of Economic Dynamics and Control 2, 329-352.

Heckman, J.J., 1999, Policies to foster human capital, NBER Working Paper 7288.

Heckman, J.J., R.J. Lalonde, and J.A. Smith, 1999, The economics and econometrics of active labor market programs, in Handbook of Labor Economics 3B, edited by O.C. Ashenfelter and D. Card, North-Holland.

Holland, P.W., 1986, Statistics and causal inference, Journal of the American Statistical Association 81, 945-960.

Holtz-Eakin, D., W. Newey, and H.S. Rosen, 1988, Estimating vector autoregressions with panel data, Econometrica 56, 1371-1395.

Holtz-Eakin, D., W. Newey, and H.S. Rosen, 1989, The Revenue-expenditure nexus: Evidence from local government data, International Economic Review 30, 415-429.

Imbens, G.W., 2004, Nonparametric estimation of average treatment effects under exogeneity: a review, Review of Economic Statistics 86, 4-29.

Kandel, E. and E.P. Lazear, 1992. Peer pressure and partnerships, Journal of Political Economy 100, 801–817.

Keane, M, and K. Wolpin, 1997, Career decisions of young men, Journal of Political Economy 105, 473-522.

Kreps, D., 1997, Intrinsic motivation and extrinsic incentives, American Economic Review 87, 359-364.

Larzelere, R.E., 1996, A review of the outcomes of parental use of nonabusive or customary physical punishment. Pediatrics. 1996; 98 (suppl): 824–828.

Lee, M.J., 2002, Panel data econometrics: methods-of-moments and limited dependent variables, Academic Press

Lee, M.J., 2005, Micro-econometrics for policy, program, and treatment effects, Oxford University Press.

Lee, M.J. and S. Kobayashi, 2001, Proportional treatment effects for count response panel data: effects of binary exercise on health care demand, Health Economics 10, 411-428.

Pearl, J., 2000, Causality, Cambridge University Press.

Persico, N., A. Postlewaite, and D. Silverman, 2004, The Effect of adolescent experience on labor market outcomes: the case of height, Journal of Political Economy 112, 1019-1053.

Robins, J.M., 1992, Estimation of the time-dependent accelerated failure time model in the presence of confounding factors, Biometrika 79, 321-334.

Robins, J.M., 1998, Structural nested failure time models, in Survival Analysis, Vol. 6, Encyclopedia of Biostatistics, edited by P. Armitage and T. Colton, Wiley.

Robins, J.M., 1999, Marginal structural models versus structural nested models as tools for causal inference, in Statistical models in epidemiology: the environment and clinical trials, edited by M.E. Halloran and D. Berry, Springer, 95-134.

Robins, J.M., S. Greenland, and F.C. Hu, 1999, Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome, Journal of the American Statistical Association 94, 687-700.

Rosenbaum, P., 2002, Observational studies, 2nd ed., Springer.

Todd, P.E. and K.I. Wolpin, 2004, The production of cognitive achievement in children: home, school and racial test score gaps, unpublished paper.

Van der Laan, M.J. and J. Robins, 2003, Unified methods for censored longitudinal data and causality, Springer-Verlag.

Witteman, J.C.M., R.B. D'Agostino, T. Stijnen, W.B. Kannel, J.C. Cobb, M.A.J de Ridder, A. Hofman, and J.M. Robins, 1998, G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham heart study, American Journal of Epidemiology 148, 390-401.

Table 1: Spanking and Behavior Scores across Groups
Mean (SD)

	Whether a child was spanked in the past week when survey was conducted		Probability of having higher than average BPI		Motors and Social Development Scale	Group Size
	at age 2-3	at age 4-5	at age 6-7	at age 4-5	at age 0-2	
The Basic Sample	.86 (.35)	.64 (.48)	.48 (.50)	.49 (.50)	102.7 (14.1)	1329
The Main Sample	.81 (.39)	.56 (.50)	.43 (.50)	.44 (.50)	103.2 (14.1)	961
Race						
White	.79 (.41)	.50 (.50)	.38 (.49)	.40 (.49)	104.5 (13.3)	482
Non-White	.83 (.38)	.62 (.49)	.48 (.50)	.49 (.50)	101.9 (14.8)	479
Sex						
Boy	.82 (.39)	.56 (.50)	.45 (.50)	.45 (.50)	100.9 (13.6)	460
Girl	.80 (.40)	.55 (.50)	.41 (.49)	.44 (.50)	105.3 (14.3)	501
Birth order						
First-borns	.84 (.36)	.55 (.50)	.43 (.50)	.46 (.50)	106 (13.5)	385
Others	.78 (.40)	.57 (.50)	.43 (.50)	.43 (.50)	101.3 (14.2)	576
Mother reads to child at age 2-3						
Often	.78 (.41)	.50 (.50)	.38 (.49)	.38 (.49)	105.5 (13.3)	561
Not often	.84 (.36)	.64 (.48)	.51 (.50)	.53 (.50)	100 (14.4)	391
How many children books a child has at home at age 4-5						
> = 10	.79 (.40)	.52 (.50)	.41 (.49)	.42 (.49)	104.2 (13.7)	789
< 10	.85 (.36)	.74 (.44)	.53 (.50)	.57 (.50)	98.2 (15.1)	169
How long TV is on per day						
< 4 hours	.80 (.38)	.49 (.50)	.38 (.49)	.37 (.48)	104.5 (14.1)	330
> = 4 hours	.81 (.39)	.60 (.49)	.46 (.50)	.48 (.50)	102.5 (14.1)	631
Mother's highest grade in 1988						
> 12	.81 (.39)	.56 (.50)	.44 (.50)	.45 (.50)	103.2. (14)	521
< = 12	.80 (.40)	.55 (.50)	.44 (.50)	.45 (.50)	103.1 (14.1)	440
Whether child was breastfed						
Breastfed	.80 (.40)	.54 (.50)	.40 (.49)	.41 (.49)	104 (14.8)	471
Not breastfed	.81 (.40)	.59 (.49)	.45 (.50)	.47 (.50)	102.2 (13.3)	429

Table 2: The Direct Effects of Spanking on BPI in Linear Structural Models

	BPI at age 6-7					BPI at age 4-5			
	IV	IV (B)	IV (B')	LSE (B, D)	LSE (D)	IV	IV'	IV (B)	LSE (B, T)
Spanked at age 2-3	-2.87 (4.86)	-4.03 (7.20)		-4.55 (2.67)*		-4.94 (15.1)		-4.07 (19.6)	-14.88* (8.52)
Spanking # at age 2-3			-3.11 (7.85)		- 2.71* (1.53)		-8.79 (18.2)		
Spanking # at 2-3 squared			.62 (2.38)		.53* (.29)		2.31 (3.81)		
Spanked at age 4-5	-1.26 (3.41)	1.42 (6.39)		-.68 (1.88)					
Spanking # at age 4-5			-1.60 (6.12)		.34 (1.25)				
Spanking # at 4-5 squared			.78 (1.66)		-.09 (.30)				
BPI at age 4-5	.56*** (.12)	.52*** (.17)	.48** (.19)	.44*** (.07)	.50*** (.05)				
Motors Score at age 0-2						-.06 (.11)	-.05 (.13)	-.02 (.10)	
Sample Size	638	476	476	216	330	488	488	535	135
R-squared	-	-		.44	.48	-	-	-	.20

Note: *p<.1; ** p<.05; *** p<.01. Standard deviations are in the parentheses.
The sample is composed of kids spanked 3 times or less in a week at age two to three and spanked 5 times or less at age four to five. The controlled inputs include a child's race, sex, birth order, and current home inputs. The instrumental variables are earlier inputs.
B – Family backgrounds variables included (mother's AFQT score, her age at child birth, whether the child was breastfed, her marriage status, highest grade, and family income).
D – Disciplinary inputs at age six to seven are excluded to avoid endogenity problem (excluded inputs are: # of times mother grounded child, took away TV or allowance, sent child to room, praised child, showed physical affection, and said positive things).
T – Measures of child temperament are included.

Table 3: The Marginal Effects of Spanking on BPI for G-Estimation

	BPI at age 6-7 (higher than sample mean)			BPI at age 4-5 (higher than sample mean)		
	Probit (B)	Probit	Probit (D)	Probit (T)	Probit (T, B)	Probit (T, C)
Spanked at least once at age 2-3	-.35* (.21)	-.23* (.12)	-.20* (.11)	-.45* (.12)	-.44* (.11)	-.44* (.12)
Spanked at least once at age 4-5	.07 (.11)	.08 (.08)	.11 (.07)			
BPI at age 4-5 is higher than mean	.46*** (.09)	.46*** (.06)	.44*** (.06)			
Motors Score at age 0-2	-.002 (.003)	-.003 (.002)	-.003 (.002)	-.0005 (.003)	.001 (.003)	.0002 (.009)
Sample Size	200	288	301	241	207	223
Pseudo R-squared	.48	.28	.22	.23	.24	.25

Note: * p<.1; ** p<.05; *** p<.01. Standard deviations are in the parentheses.
The sample is composed of kids spanked 3 times or less in a week at age two to three and spanked 5 times or less at age four to five. The controlled inputs include a child's race, sex, birth order, current and earlier home inputs.
B – Family backgrounds variables included (mother's AFQT score, her age at child birth, whether the child was breastfed, her marriage status, highest grade, and family income).
D – Disciplinary inputs at age six to seven are excluded to avoid endogenity problem (excluded inputs are: # of times mother grounded child, took away TV or allowance, sent child to room, praised child, showed physical affection, and said positive things).
T – Measures of child temperament are included.
C – Variables on whether child attended child care in the first three years are included.

Table 4: The Effects of Spanking on BPI: Structural Nested Model

	Point Estimate of ψ_0	95% Confidence Interval for ψ_0	Estimated effects evaluated at the sample mean (105.3) (in terms of point reduction in BPI at age 6-7)
Logit	0	[-.05, .05]	0
Logit (T)	.01	[-.06, .08]	1.06
Logit (T, B)	.04	[-.06, .14]	4.3
Logit (T, B), assuming $\psi_1=c\psi_0$ where c is a constant:			
When $\psi_1=\psi_0/4$	.20	[-.20, .60]	12.64
When $\psi_1=\psi_0/3$	.135	[-.15, .42]	9.48
When $\psi_1=\psi_0/2$	.075	[-.1, .25]	5.79
When $\psi_1=\psi_0$	.04	[-.06, .14]	4.3
When $\psi_1=2\psi_0$	.02	[-.02, .06]	3.16
When $\psi_1=3\psi_0$	.015	[-.01, .04]	3.16
When $\psi_1=4\psi_0$	.01	[-.01, .03]	2.11

Note: The sample is composed of kids spanked 3 times or less in a week at age two to three and spanked 5 times or less at age four to five. The regressor in the logit models is whether a child was spanked at age four to five, while regressants in the basic specification include a child's race, sex, birth order, and detailed home inputs at age four to five and two to three.
T – Measures of child temperament are included.
B – Family backgrounds variables included (mother's AFQT score, her age at child birth, whether the child was breastfed, her marriage status, highest grade, and family income).

Table 5: The Effects of Spanking on BPI: Granger Causality

	BPI at age 6-7				
	LSE	LSE (D)	LSE (D,A)	LSE (D,B)	LSE (D,B,A,T)
Spanked at least once at age 2-3	-4.59* (2.67)	-5.22** (2.69)		-5.27* (3.13)	-6.80** (3.2)
Spanking # at age 2-3			-.57* (.33)		
Spanking # at 2-3 squared			.03** (.01)		
Spanked at least once at age 4-5	-.69 (1.79)	-1.19 (1.76)		-1.55 (2.08)	-6.03** (2.50)
Spanking # at age 4-5			-.68* (.41)		
Spanking # at 4-5 squared			.04** (.02)		
BPI at age 4-5	.48*** (.06)	.49*** (.05)	.50*** (.05)	.49*** (.07)	.53*** (.09)
Motors Score at age 0-2	-.02 (.05)	-.006 (.05)	-.03 (.05)	.01 (.06)	.003 (.08)
Sample Size	263	271	394	224	182
R-squared	.53	.49	.50	.47	.56

Note: * p<.1; ** p<.05; *** p<.01. Standard deviations are in the parentheses.
The controlled inputs include a child's race, sex, birth order, current and earlier home inputs.
D – Disciplinary inputs at age six to seven are excluded to avoid endogenity problem (excluded inputs are: # of times mother grounded child, took away TV or allowance, sent child to room, praised child, showed physical affection, and said positive things).
A – All kids with various spanking frequencies are included.
B – Family backgrounds variables included (mother's AFQT score, her age at child birth, her marriage status, her highest grade, and family income).
T – Measures of child temperament are included.

Table A: Summary Statistics of the Basic Sample (1329 children)

Variable	Mean	SD
Behavior Problem Index (BPI) total standard score at age 6-7	105.3	14.7
Behavior Problem Index (BPI) total standard score at age 4-5	104.8	14.8
Motors and Social Development Scale by age two	102.7	14.1
Spanked # last week at survey time at age 4-5	1.79	2.49
Whether a child was spanked at least once at age 4-5	.64	.48
Spanked # last week at survey time at age 2-3	3	3.77
A child was spanked at least once at age 2-3	.86	.35
(G1). Child Demographic and Health Information		
race of child: Black or Hispanic	.50	.50
sex of child: boy	.50	.50
birth order of child	1.94	1.01
Whether a child has a low birth weight	.06	.25
Child is breastfed	.52	.50
(G2) Mother's Background Information		
Mother's AFQT score taken in 1981	42.79	28.7
Mother's highest grade at 1988	12.45	2.56
Family salary in 1988	8360	9234
Family incomes in 1988	27632	20286
Mother was married in 1988	.76	.43
Mother's age at child birth	26	5.82
(G3) Current home inputs for 6-7 years olds		
child has 10 or more books	.84	.36
how often mom reads to child: at least 3 times a week	.76	.43
is there musical instrument at home	.39	.49
family gets newspaper daily	.47	.50
how often child reads for enjoyment: everyday	.76	.43
family encourages hobbies	.89	.32
child get special lessons/activities	.50	.50
how often child taken to museum: at least several times a year	.77	.42
how often child taken to performance: at least several times a year	.61	.49
how often get together with relatives/friends: at least 2-3 times/month	.60	.49
# of hours/weekday child sees TV	4.65	6.53
# hours/weekend day child sees TV	4.65	4.79
child ever sees a father figure	.94	.24
how often child w/ dad outdoors: at least once a week	.51	.50
how often child eats w/mom & dad: at least once a day	.78	.41
parents discuss TV programs w/child	.82	.39
#times past week grounded child	.65	2.36
#times past week took away TV	.63	2.01
#times past week praised child	8.05	11.66
#times past week took away allowance	.18	2.11
#times mom showed child physical affection	19.45	22.4

#times past week sent child to room	1.62	2.96
#times past week said positive things	7	11.64
(G4) Home inputs for 3-5 years olds		
how often mother read to child: at least 3 times a week	.56	.50
how many books does child have: 10 or more books	.81	.39
how many magazines does family get: 3 or more	.59	.49
Does child have record/tape player	.77	.42
amount of choice child has in food: a lot	.68	.47
# of hours TV is on per day	5.56	5.3
how often child taken on outing: at least several times a week	.55	.50
how often child taken to museum: at least several times a year	.69	.46
how often child eats w/ mom & dad: at least once a day	.75	.43
child see father(-figure) daily	.80	.40
Mom helps child learn numbers	.94	.23
Mom helps child learn alphabet	.93	.26
Mom helps child learn colors	.94	.23
Mom helps child learn shapes	.83	.37
Mom responds to hit-hit child back	.16	.36
Mom responds to hit-send to room	.51	.50
Mom responds to hit-spank child	.49	.50
Mom responds to hit-talk to child	.71	.45
Mom responds to hit-ignore it	.03	.17
Mom responds to hit-give chores	.05	.21
Mom responds to hit-take allowance	.04	.20
Mom responds to hit-hold child hands	.13	.33
(G5) Home Inputs for 0-2 years olds		
how often child gets out of the house: everyday	.63	.48
how many children's books child has: 10 or more books	.80	.40
how often mother reads to child: at least 3 times a week	.57	.49
how often mother takes child to grocery: once a week or less	.61	.49
how many cuddly or role-playing toys	16.87	14
how many push or pull toys child has	7.91	7.77
mothers attitude how child learns best: parents should always teach	.53	.50
Does child see father (-figure) daily?	.83	.37
how often child eats with both mom and dad: at least once a day	.69	.46
how often mother talks to child while working: often	.56	.50
(G6). Mother Working, Prenatal Care, and Child Care		
hours worked per week on main job 4th quarter before birth child	35.82	11.3
hours worked per week on main job 3rd quarter before birth child	35.58	11.4
sonogram done during pregnancy	.74	.44
mother took vitamins during pregnancy	.96	.20
child in regular child care during 1st year	.47	.50
child in regular child care during 2nd year	.51	.50
child in regular child care during 3rd year	.53	.50