Journal of the Royal Statistical Society
SERIES A Statistics in Society

J. R. Statist. Soc. A (2012)
175, *Part* 2, *pp.* 535–567

Finding dynamic treatment effects under anticipation: the effects of spanking on behaviour

Myoung-jae Lee

Korea University, South Korea, and Australian National University, Canberra, Australia

and Fali Huang

Singapore Management University, Singapore

[Received June 2009. Final revision June 2011]

Summary. The dynamic treatment effect literature considers multiple treatments administered over time, with some treatments affected by interim outcomes. But the literature overlooks the possibility of individuals acting in anticipation of future treatments. This lack of anticipation aspect may not matter in the drug–response relationships which motivated the literature. But human beings (or animals with some intelligence) do not just respond to current and past treatments, but also 'reflect and anticipate' future treatments. For example, a punishment or reward is likely to prompt forward looking. Even if no personal punishment or reward is involved, people may take action in anticipation of a future government policy, which would be an important concern for policy makers. The paper explores how to find dynamic treatment effects allowing for forward looking or anticipation by extending available dynamic treatment effect approaches in the literature. Then the methods proposed are applied to the effects of spanking on a child's bad behaviour where a child may act better in anticipation of future spanking, which is analogous to the relationship between punishment and crime.

Keywords: Anticipation; Dynamic models; Dynamic treatment effect; Panel data; Spanking

1. Introduction

Finding the effect of a treatment on a response variable is probably the most fundamental task in scientific research. When the treatment is one shot, the task is fairly straightforward, as the time dimension involved in causality is virtually stripped off. But, when multiple treatments are given over time for a final response variable at the end, things become more complicated, and we can think of three cases. First, no treatment is affected by any interim responses: *no feedback*. Second, some treatments are affected by some interim responses, but these interim responses do not affect the final response: *feedback without lagged response effects*. Third, some treatments are affected by some interim responses which also affect the final response: *feedback with lagged response effects*.

In the no-feedback case, the multiple temporal treatments can be regarded as a single static treatment taking on many different values, and its effect on the final response can be assessed as in the usual static case. In 'feedback without lagged response effects', we can use a dynamic panel (i.e. longitudinal) data model controlling the lagged responses, which in essence cuts off the feedback feature. In 'feedback with lagged response effects', controlling the lagged responses

Address for correspondence: Myoung-jae Lee, Department of Economics, Korea University, Seoul 136-701, South Korea.
E-mail: myoungjae@korea.ac.kr

 0964–1998/12/175535

does not work as will be shown later, but the so-called 'G-algorithm' is applicable. This algorithm was developed by Robins (1986, 1987) and has been refined since then in many ways as can be seen, for example, in Robins and Hernán (2009).

Those G-algorithm-based approaches have been motivated primarily by medical dose–response relationships where no sophisticated reactions by the individuals are involved. But human beings (or other animals for that matter) do not just respond to a stimulus (i.e. treatment) but also anticipate what will happen in the future with the treatment possibly reapplied. One such case is penalty or reward effects on behaviour, where an important justification of a penalty or reward is the deterrence or encouragement that the individuals take into account when considering the future penalty or reward for the current behaviour. That is, the expected future treatments can influence the current response, which is unthinkable for medical dose–response relationships as the expected future doses of a drug cannot influence the current disease state.

As an example, consider the effect on a child's behaviour of mild spanking, which is something that almost every child is subject to—this will be examined in the empirical part of this paper. In this example, the degree and frequency of spanking in the current period can influence the child's current behaviour not only directly but also indirectly through their effect on the future spankings that are anticipated by the child, although acting on such an anticipation may be too difficult for the child if the child is too young. More general (and important) than this kind of individual punishment might be that people act in anticipation of various government policies, which should matter greatly to policy makers in announcing as well as designing a policy.

The goal of this paper is to extend the G-algorithm-based methods to the direction of forward looking or anticipation so that the extended framework can deal with the dynamic effects of an anticipation-inducing treatment, and then apply the extended approaches to find the effects of spanking on children's behaviour. Section 2 reviews the G-algorithm and four 'marginal structural model' based approaches to dynamic treatments with feedback and lagged response effects. Section 3 explores how to allow for the anticipation aspect. Section 4 provides an empirical analysis of spanking and children's bad behaviour, and Section 5 concludes.

2. Various approaches to dynamic treatment effects

2.1. Basic framework

Suppose that

$$(X_{0i}, Y_{0i}, X_{1i}, D_{1i}, Y_{1i}, X_{2i}, D_{2i}, Y_{2i}), i = 1, \ldots, N, \text{ are observed and independent and identically distributed.}$$

We shall often omit the subscript i in view of the assumption of independent and identically distributed variables. Mostly, we shall use capital letters for random variables and lower-case letters for their realized values. In period t, X_t is the baseline covariates which can affect the treatment D_t and response Y_t, and D_t then may affect Y_t; Fig. 1. Before period 1, there are baseline variables X_0 (and Y_0). It is desired to find the effect of the treatment 'profile' (d_1, d_2) (exogenously fixing (D_1, D_2) at (d_1, d_2)) on the final response Y_2 relative to no treatment $(0, 0)$, when there is a *feedback* $Y_1 \to D_2$ and a *lagged response effect* $Y_1 \to Y_2$; see Fig. 2 that omits X_ts.

In Fig. 2, there are several arrows for various effects, and the effects of interest are

(a) *direct and indirect* effects of D_1 on Y_2 ($D_1 \to Y_2$ and $D_1 \to Y_1 \to Y_2$) and
(b) the *direct* effect of D_2 on Y_2 ($D_2 \to Y_2$).

The sum of the direct and indirect effects is the *total effect* of (d_1, d_2) on Y_2. There is no $D_1 \to D_2$ in Fig. 2, because D_1 influences D_2 only indirectly through Y_1. Without the feedback and lagged

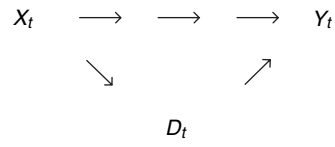

Fig. 1

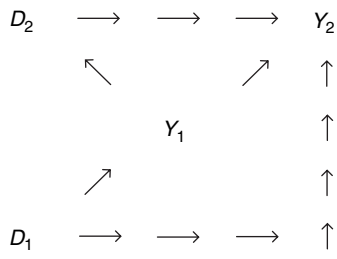

Fig. 2

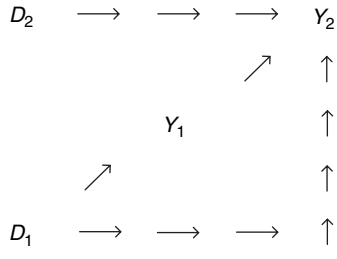

Fig. 3

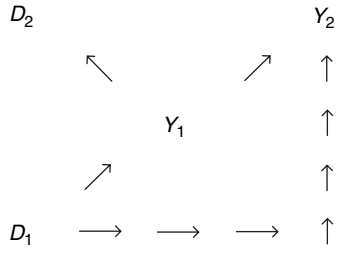

Fig. 4

response effects, we can recast D_1 and D_2 in a static framework. If D_1 and D_2 are binary taking values 0 and 1, then $D_1 = 0, 1$ and $D_2 = 0, 1$ together define a static treatment with $4 = 2 \times 2$ categories, which can be handled with static treatment effect estimators developed for multiple treatments; see Imbens (2000) and Lechner (2001).

In Fig. 2 $Y_1 \to D_2$ means that D_2 bacomes adjusted after the interim response Y_1 has been observed. This makes D_2 endogenous even if D_1 is randomized. Hence, what is essential is removing the feedback $Y_1 \to D_2$ to obtain Fig. 3, which has only the desired direct and indirect effects. The feedback does not have to be through Y_1: instead of Y_1, some 'mediating variable' that affects D_2 and helps to predict Y_2 may appear.

Examine Fig. 4 with no true direct effect $D_2 \to Y_2$. But, if Y_1 is not controlled, then Y_1 becomes a common factor for D_2 and Y_2, i.e., despite no true direct effect of D_2 on Y_2, D_2 may look influential for Y_2 because Y_1 affects both D_2 and Y_2. Suppose that we control Y_1 in Fig. 2 to

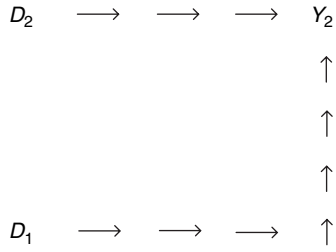

Fig. 5

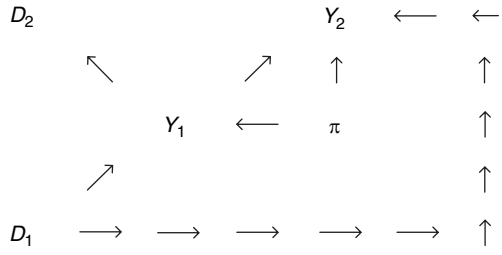

Fig. 6

avoid this kind of problem. Then we obtain Fig. 5 without the indirect effect $D_1 \to Y_1 \to Y_2$. This demonstrates the *fundamental dilemma*: control Y_1 to miss the indirect effect (in Fig. 5), or do not control Y_1 to incur the omitted variable bias (in Fig. 4). Panel data models with Y_2 on the left-hand side and D_1 and D_2 on the right-hand side will commit either mistake depending on whether Y_1 appears on the right-hand side or not.

If there is an unobserved variable affecting both Y_1 and Y_2 as π does in Fig. 6, then the situation becomes more complicated. In Fig. 6, Y_1 is the 'common effect' for D_1 and π. Hence, if Y_1 is controlled, then the relationship between D_1 and π becomes 'distorted' (and thus the relationship between D_1 and Y_2), which biases the direct effect of D_1 on Y_2, i.e. controlling Y_1 does not yield the correct direct effects of D_1 and D_2, differently from Fig. 5. The concern for π is genuine: often in panel data, Y_1 and Y_2 are postulated to share a common time constant error term π that is called a 'unit-specific effect' or 'individual-specific effect'; see for example Lee (2002).

The bias due to controlling a common effect is sometimes called a 'selection bias', but other names are used as well (see Hérnan *et al.* (2004) and Robins (2008)); also, the term selection bias is used often in different contexts. In contrast with the bias due to a common effect, another well-known source of bias is a common cause or factor as in Fig. 4. This is typically called a 'confounding factor', but other names have been used for this as well.

2.2. *G-algorithm*

This section introduces the *G*-algorithm which is non-parametric, and explains four operational versions based on linear 'marginal structural models (MSMs)'. In the literature, a somewhat confusing array of names with the prefix *G* have been used. For instance, Robins and Hernán (2009), page 554, mentioned three '*G*-methods': a '*G*-computation algorithm', an inverse probability of treatment weighting of MSMs and a '*G*-estimation' of 'structural nested models'. The *G*-computation algorithm has also been called the '*G*-algorithm', and *G*-estimation is a way of estimating dynamic treatment effects with an artificial regressor as follows.

Suppose that there is an initial binary treatment randomization ($R = 0, 1$). If the treatment effect on Y_2 due to (D_1, D_2) takes the additive form $\gamma_1 D_1 + \gamma_2 D_2$ for the effect parameters

(γ_1, γ_2), then $Y(\gamma_1, \gamma_2) \equiv Y_2 - \gamma_1 D_1 - \gamma_2 D_2$ is the untreated initial response. Since $Y(\gamma_1, \gamma_2)$ should be independent of R, if $Y(\gamma_1, \gamma_2)$ is used as an artificial regressor to explain R, it should have a zero coefficient. G-estimation tries different $Y(g_1, g_2)$s with (g_1, g_2) ranging over a parameter space to find $(\hat{g}_1, \hat{g}_2)$ that results in a zero coefficient for $Y(\hat{g}_1, \hat{g}_2)$ in the R-equation. Then $(\hat{g}_1, \hat{g}_2)$ is taken as an estimate for (γ_1, γ_2). This method was in fact tried in an earlier version of this paper but then dropped later, as it seems unsuitable when there are many treatment parameters to estimate by using only a single zero-parameter restriction. Even if there is no randomization, G-estimation can still be done by using an 'ignorability condition'.

For the three G-methods, Robins and Hernán (2009), pages 554–555, noted that

> 'If we used only completely saturated (i.e., nonparametric) models, all three methods would give identical estimates of the effect of treatment. However, in realistic longitudinal studies, the data are sparse and high-dimensional. Therefore, possibly misspecified, non-saturated models must be used. As a consequence, the three methods can provide different estimates.'

In this paper, we adopt a linear MSM to apply inverse probability of treatment weighting and two other approaches for MSM. See Toh and Hernán (2008) for an easy-to-read illustration of G-methods.

2.2.1. Non-parametric version

Define the *potential responses* for the observed responses Y_1 and Y_2:

(a) $Y_1^{d_1}$, when treatment d_1 is given exogenously at time 1;
(b) $Y_2^{d_1 d_2}$, when treatments d_1 and d_2 are given exogenously at times 1 and 2, $d_1, d_2 \in [0, \infty)$.

With D_1 and D_2 observed, we have $Y_1 = Y_1^{D_1}$ and $Y_2 = Y_2^{D_1 D_2}$, i.e. only the potential responses corresponding to the realized treatment levels are observed, and all the other potential responses—'counterfactuals'—are not. Also define the *potential treatment* for D_2:

(c) $D_2^{d_1}$, when treatment d_1 is given exogenously at time 1 (thus $Y_1^{d_1}$ realized);

with D_1 observed, $D_2 = D_2^{D_1}$. The goal is to find *the mean effect* $E(Y_2^{d_1 d_2} - Y_2^{00})$ for the intervened treatment 'profile' (d_1, d_2) *versus* no treatment at all.

Let

$$\bar{X}_2 \equiv (X_0, Y_0, X_1, X_2);$$

note that $\bar{X}_2$ does not include Y_1, as Y_1 must be integrated out differently from the variables in $\bar{X}_2$. Assume *no unobserved confounder (NUC)* for $Y_2^{d_1 d_2}$ relative to the treatments:

(a) NUC 1, $Y_2^{d_1 d_2} \amalg D_1 | \bar{X}_2$, $\forall d_1, d_2 \in [0, \infty)$;
(b) NUC 2, $Y_2^{d_1 d_2} \amalg D_2^{d_1} | (D_1 = d_1, Y_1^{d_1}, \bar{X}_2)$, $\forall d_1, d_2 \in [0, \infty)$.

Here '$a \amalg b | c$' means the conditional independence of a and b given c. The NUC condition can be also called 'sequential ignorability'.

The G-algorithm under NUCs is (Robins (1986, 1987, 1998, 1999), and the references therein)

$$E(Y_2^{d_1 d_2} | \bar{X}_2) = \int E(Y_2 | d_1, d_2, y_1, \bar{X}_2)\, f(y_1 | d_1, \bar{X}_2)\, \mathrm{d}y_1$$

where $f(y_1 | d_1, \bar{X}_2)$ denotes the conditional density of $y_1 | (d_1, \bar{X}_2)$; conditional density functions will be often denoted with $f(\cdot|\cdot)$ in the remainder of this paper. The G-algorithm holds because the right-hand side of the last display is

$$\int E(Y_2^{d_1d_2}|d_1, d_2^{d_1}, y_1, \bar{X}_2)\, f(y_1|d_1, \bar{X}_2)\, \mathrm{d}y_1$$
$$= \int E(Y_2^{d_1d_2}|d_1, y_1, \bar{X}_2)\, f(y_1|d_1, \bar{X}_2)\, \mathrm{d}y_1 \qquad \text{(owing to NUC 2)}$$
$$= E(Y_2^{d_1d_2}|d_1, \bar{X}_2) = E(Y_2^{d_1d_2}|\bar{X}_2) \qquad \text{(owing to NUC 1).}$$

In essence, the G-algorithm starts with the mean of $Y_2^{d_1d_2}$ for the subpopulation $(d_1, d_2, y_1, \bar{X}_2)$. The subpopulation components d_1 and d_2 are needed because $Y_2^{d_1d_2}$ is observed only for those with $D_1 = d_1$ and $D_2 = d_2$, and y_1 is needed to account for the dynamic feedback. Then the subpopulation is generalized to the whole population (i.e. the 'selection problem' is ruled out) as d_1 and d_2 are removed by NUC 1 and NUC 2 respectively, and y_1 is removed by integration. Obtaining $E(Y_2^{d_1d_2}|\bar{X}_2)$ and $E(Y_2^{00}|\bar{X}_2)$ and then integrating out $\bar{X}_2$, we obtain the desired $E(Y_2^{d_1d_2} - Y_2^{00})$. Theorem 2 of Gill and Robins (2001) shows that the G-algorithm works also for continuous treatments with zero probability of the conditioning event.

Even for two periods, implementing the G-algorithm requires finding $E(Y_2|d_1, d_2, y_1, \bar{X}_2)$ and $f(y_1|d_1, \bar{X}_2)$ $\forall(y_1, \bar{X}_2)$ and then integrating out y_1 and $\bar{X}_2$, which could be daunting to say the least. Later, we shall introduce practical approaches based on MSMs. As has been already seen, G-estimation for structural nested models (Robins, 1998, 1999) is available as well, although it is not further discussed in this paper. Also Lechner (2008) applied 'sequential matching' to dynamic treatment effects. Cases where a treatment timing is the main choice variable in a continuous duration set-up were examined by Abbring and van den Berg (2003) whereas discrete time cases were dealt with by Heckman and Navarro (2007). Li *et al.* (2001) proposed 'risk set matching' in duration contexts, and the matching done in Lee (2010) to address 'bundling effects' of a secondary product tied to its primary product is in essence a risk set matching. A review of (dynamic) treatment effects and causality from the econometric viewpoint can be seen in Lee (2005), Abbring and Heckman (2007, 2008) and Lechner (2011).

2.2.2. *Linear model version*

Analogously to $\bar{X}_2$, define

$$\bar{X}_1 \equiv (X_0, Y_0, X_1).$$

Suppose that a simple linear model holds:

$$\left.\begin{aligned} D_1 &= \xi_1 + \bar{X}_1'\xi_x + \varepsilon_1, \\ Y_1^{d_1} &= \alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1 + U_1 \qquad (d_1 \text{ affects } Y_1^{d_1}), \\ D_2^{d_1} &= \zeta_1 + X_2'\zeta_x + \zeta_y Y_1^{d_1} + \varepsilon_2 \qquad (Y_1^{d_1} \text{ affects } D_2^{d_1}), \\ Y_2^{d_1d_2} &= \beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_y Y_1^{d_1} + U_2 \qquad (d_1, d_2, Y_1^{d_1} \text{ affect } Y_2^{d_1d_2}) \end{aligned}\right\} \quad (1)$$

(this is a structural form (SF) in potential variables) where ξ, α, ζ and β are parameters and $(\varepsilon_1, \varepsilon_2, U_1, U_2)$ are mean 0 errors. As $Y_2^{d_1d_2}$ depends on the same period X_2 and the last period $Y_1^{d_1}$, $Y_1^{d_1}$ depends on the same period X_1 and the last period $Y_0 \in \bar{X}_1$; a similar statement can be made for $D_2^{d_1}$ and D_1. X_1 appears in expression (1) with the upper bar different from X_2, as the first period has to 'collect' all the past information.

The $Y_2^{d_1 d_2}$ *reduced form (RF)* with $Y_1^{d_1}$ substituted out is

$$\begin{aligned} Y_2^{d_1 d_2} &= \beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_y(\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1 + U_1) + U_2 \\ &= \beta_1 + \beta_y\alpha_1 + \bar{X}_1'\alpha_x\beta_y + X_2'\beta_x + (\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2 + \beta_y U_1 + U_2. \end{aligned}$$

In this model, the desired *total effect* is

$$E(Y_2^{d_1 d_2} - Y_2^{00}) = (\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2$$

consisting of the direct effect $\beta_{d\text{lag}} d_1 + \beta_d d_2$ of (d_1, d_2) on Y_2 and the indirect effect $\beta_y\alpha_d d_1$ of d_1 on Y_2 through Y_1.

For the linear model, Huang and Lee (2010) showed that the following expressions are sufficient for NUC 1 and NUC 2, and thus for the *G*-algorithm:

(a) $(U_1, U_2) \amalg \varepsilon_1 | \bar{X}_2$ for NUC 1;
(b) $U_2 \amalg \varepsilon_2 | (\varepsilon_1, U_1, \bar{X}_2)$ for NUC 2.

Huang and Lee (2010) also showed that the *G*-algorithm reduces to the above total effect under conditions (a) and (b). The linear model assumption is critical for this, as well as for the decomposition of the total effect into the direct and indirect effects. Although restrictive, the linear model shows well what is essentially going on in the *G*-algorithm: the process of removing $Y_1^{d_1}$ (by substituting or integrating it out) while conditioning on d_1 (and $\bar{X}_2$) yields $\beta_y\alpha_d d_1$, which is the key indirect effect that is difficult to capture otherwise.

2.3. Marginal structural model approaches

The linear model as above has been called an MSM. The qualifier 'marginal' in an MSM comes because only the marginal model for $Y_2^{d_1 d_2}$ is specified in the joint distribution of $\{Y_2^{d_1 d_2}, \forall d_1, d_2 \in [0, \infty)\}$, and 'structural' is because a (data-generating) structure is imposed on $Y_2^{d_1 d_2}$. This section examines four approaches to the linear MSM to apply the first three in the empirical section. Since the dynamic treatment effects that are estimated by these approaches consist of linear-model-based direct and indirect effects, we also discuss non-parametric direct and indirect effects near the end of this section, drawing on Pearl (2009, 2010).

The first approach is a *weighted least squares (WLS) estimator to the Y_2 RF* (Robins, 1999; Hérnan *et al.*, 2001; Joffe *et al.*, 2004); related weighting approaches to dynamic treatment effect analysis were also used in Lechner (2009) and Lechner and Wiehler (2012). The second is an *instrumental variable estimator (IVE) to the Y_2 RF*, which obviates weighting. The third is a *two-stage estimator (TSE)* where both Y_1 and Y_2 SFs are estimated separately (Huang and Lee, 2010). The fourth is a *'control function' type TSE to an 'error-augmented' Y_2 RF* (Almirall *et al.*, 2010).

Before we proceed, one remark for models in observed variables should be made. For estimation, we need the $Y_1^{d_1}$ and $Y_2^{d_1 d_2}$ SFs to hold in their observed variables as in

$$\begin{aligned} Y_1 &= \alpha_1 + \bar{X}_1'\alpha_x + \alpha_d D_1 + U_1, \\ Y_2 &= \beta_1 + X_2'\beta_x + \beta_{d\text{lag}} D_1 + \beta_d D_2 + \beta_y Y_1 + U_2. \end{aligned} \tag{2}$$

Huang and Lee (2010) showed that this holds if

$$\begin{aligned} E(Y_1^{d_1} | d_1, \bar{X}_1) = E(Y_1^{d_1} | \bar{X}_1) \Leftrightarrow E(U_1 | d_1, \bar{X}_1) = E(U_1 | \bar{X}_1) = 0, \\ E(Y_2^{d_1 d_2} | d_1, d_2, Y_1^{d_1}, X_2) = E(Y_2^{d_1 d_2} | Y_1^{d_1}, X_2) \Leftrightarrow E(U_2 | d_1, d_2, Y_1^{d_1}, X_2) = E(U_2 | Y_1^{d_1}, X_2) \end{aligned}$$

for which the above conditions (a) and (b) are sufficient. This kind of 'selection-on-observables' condition is pervasive in the literature, and it is also called 'conditional independence' or 'ignorability'. Substituting the Y_1-equation in expression (2) into the Y_2-equation, we obtain the Y_2 RF with D_1 and D_2 on the right-hand side:

$$Y_2 = \beta_1 + \beta_y\alpha_1 + \bar{X}_1'\alpha_x\beta_y + X_2'\beta_x + (\beta_{d\text{lag}} + \beta_y\alpha_d)D_1 + \beta_d D_2 + \beta_y U_1 + U_2. \tag{3}$$

2.3.1. Weighted least squares to Y_2 reduced form

The main idea of WLS to MSMs can be seen in the Y_2 RF: the coefficients of D_1 and D_2 are the total effect of $(d_1, d_2) = (1, 1)$ when added up. One might try to apply a least squares estimator (LSE) to the Y_2 RF. But the LSE to RF (3) is inconsistent because D_2 *is correlated with the error term* $\beta_y U_1 + U_2$ *because* Y_1 *influences* D_2. The main idea is thus to remove this endogeneity problem by weighting. Observe that

$$\begin{aligned}
E\left\{ Y_2 \frac{f(D_2|D_1, \bar{X}_2)}{f(D_2|D_1, Y_1, \bar{X}_2)} \middle| \bar{X}_2 \right\} &= \int y_2 \frac{f(d_2|d_1, \bar{X}_2)}{f(d_2|d_1, y_1, \bar{X}_2)} \\
&\quad \times f(y_2|d_1, d_2, y_1, \bar{X}_2)\, f(d_1, d_2, y_1|\bar{X}_2)\, \mathrm{d}y_2\, \mathrm{d}d_2\, \mathrm{d}y_1\, \mathrm{d}d_1 \\
&= \int y_2 \frac{f(d_2|d_1, \bar{X}_2)}{f(d_2|d_1, y_1, \bar{X}_2)} f(y_2|d_1, d_2, y_1, \bar{X}_2)\, f(d_2|d_1, y_1, \bar{X}_2) \\
&\quad \times f(y_1|d_1, \bar{X}_2)\, f(d_1|\bar{X}_2)\, \mathrm{d}y_2\, \mathrm{d}d_2\, \mathrm{d}y_1\, \mathrm{d}d_1 \\
&= \int y_2\, f(y_2|d_1, d_2, y_1, \bar{X}_2)\, f(d_2|d_1, \bar{X}_2)\, f(y_1|d_1, \bar{X}_2) \\
&\quad \times f(d_1|\bar{X}_2)\, \mathrm{d}y_2\, \mathrm{d}d_2\, \mathrm{d}y_1 \mathrm{d}d_1.
\end{aligned}$$

The weighting replaces $f(d_2|d_1, y_1, \bar{X}_2)$ with $f(d_2|d_1, \bar{X}_2)$, creating an artificial population where D_2 is not affected by Y_1 and thus uncorrelated with the error $\beta_y U_1 + U_2$ in RF (3).

With the endogeneity of D_2 absent, the LSE can be applied to the artificial population, which is merely the WLS estimator, i.e. multiply each variable in $(Y_{2i}, \bar{X}_{1i}, X_{2i}, D_{1i}, D_{2i}) = (Y_{2i}, \bar{X}_{2i}, D_{1i}, D_{2i})$ of RF (3) by

$$\omega(\bar{X}_{2i}, D_{1i}, D_{2i}, Y_{1i}) \equiv \left\{ \frac{f(D_{2i}|D_{1i}, \bar{X}_{2i})}{f(D_{2i}|D_{1i}, Y_{1i}, \bar{X}_{2i})} \right\}^{1/2}$$

to obtain the transformed variables $(Y_{2i}^*, \bar{X}_{2i}^*, D_{1i}^*, D_{2i}^*)$ and to do the least squares estimation of Y_2^* on $(\bar{X}_2^*, D_1^*, D_2^*)$.

To understand better why WLS works, observe that

$$E(D_2 Y_1|d_1, \bar{X}_2) = \int d_2 y_1\, f(d_2, y_1|d_1, \bar{X}_2)\, \mathrm{d}d_2\, \mathrm{d}y_1$$

whereas

$$E(D_2^* Y_1^*|d_1^*, \bar{X}_2^*) = \int d_2^* y_1^*\, f(d_2^*|d_1^*, \bar{X}_2^*)\, f(y_1^*|d_1^*, \bar{X}_2^*)\, \mathrm{d}d_2^*\, \mathrm{d}y_1^*$$

as $D_2^* \amalg Y_1^*|(D_1^*, \bar{X}_2^*)$. Intuitively, in the artificial population with the 'starred variables', $D_2^* \amalg Y_1^*|(D_1^*, \bar{X}_2^*)$ holds, which implies that $D_2^* \amalg U_1^*|(D_1^*, \bar{X}_2^*)$ and thus $E(D_2^* U_1^*) = 0$: the endogeneity of D_2 due to $\mathrm{corr}(D_2, U_1) \neq 0$ in the LSE to RF (3) does not matter for the WLS estimator.

Suppose that D_1 and D_2 are continuously distributed. In theory, the densities in $\omega(\bar{X}_2, D_1, D_2, Y_1)$ can be estimated non-parametrically. But, in practice, it would be difficult if the dimen-

sion of $\bar{X}_2$ is high as in our data. One practical solution is to assume that $D_2|(D_1, Y_1, \bar{X}_2)$ is normally distributed with a linear function of $(D_1, Y_1, \bar{X}_2)$ as its mean:

$$D_2 = \theta_1 + \theta_d D_1 + \theta_y Y_1 + \bar{X}_2'\theta_x + \psi_1, \qquad \psi_1 \sim N(0, \sigma_1^2) \text{ and } \psi_1 \amalg (D_1, Y_1, \bar{X}_2)$$
$$\Rightarrow f(D_2|D_1, Y_1, \bar{X}_2) = \frac{1}{\sigma_1}\phi\left\{\frac{D_2 - (\theta_1 + \theta_d D_1 + \theta_y Y_1 + \bar{X}_2'\theta_x)}{\sigma_1}\right\}$$

where ϕ denotes the $N(0,1)$ density and all θ-parameters and σ_1 can be estimated with the LSE.

From $f(D_2|D_1, Y_1, \bar{X}_2)$, average out Y_1 while fixing $(D_1, \bar{X}_2)$: with $Q_i \equiv (D_{1i}, \bar{X}_{2i}')'$,

$$f(D_{2i}|D_{1i}, \bar{X}_{2i}) = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{\sigma_1}\phi\left\{\frac{D_{2i} - (\theta_1 + \theta_d D_{1i} + \theta_y Y_{1j} + \bar{X}_{2i}'\theta_x)}{\sigma_1}\right\}K\left(\frac{Q_j - Q_i}{h}\right)$$

where K is a kernel and $h \downarrow 0$ is a bandwidth. Suppose that the dimension of Q_i is $\vartheta \times 1$ and the components of Q_i are denoted as q_{ci}, $c = 1, \ldots, \vartheta$. Then we may set

$$K\left(\frac{Q_j - Q_i}{h}\right) = \prod_{c=1}^{\vartheta}\phi\left\{\frac{q_{cj} - q_{ci}}{\nu \,\mathrm{SD}(q_c) N^{-1/(\vartheta+4)}}\right\}$$

where $\mathrm{SD}(q_c)$ is the standard deviation (SD) of q_c, and $\nu = 1$ is a good rule of thumb, although ν can be chosen more formally by using 'cross-validation' if desired.

A simpler alternative to this conditional averaging is adopting a linear model for $D_2|(D_1, \bar{X}_2)$:

$$D_2 = \kappa_1 + \kappa_d D_1 + \bar{X}_2'\kappa_x + \psi_0, \qquad \psi_0 \sim N(0, \sigma_0^2) \text{ and } \psi_0 \amalg (D_1, \bar{X}_2)$$
$$\Rightarrow f(D_2|D_1, \bar{X}_2) = \frac{1}{\sigma_0}\phi\left\{\frac{D_2 - (\kappa_1 + \kappa_d D_1 + \bar{X}_2'\kappa_x)}{\sigma_0}\right\}.$$

With the preceding linear model for $D_2|(D_1, Y_1, \bar{X}_2)$ holding, this means that Y_1 becomes included in the error term ψ_0. Since $\mathrm{corr}(D_1, Y_1) \neq 0$, we obtain $\mathrm{corr}(D_1, \psi_0) \neq 0$: the LSE to the Y_1-omitted D_2 model would be inconsistent. But the goal is not estimating κs consistently, which means that κs being different from θs does not matter so long as both ψ_1 and $\psi_0 = \theta_y Y_1 + \psi_1$ are approximately normal.

Suppose now that D_1 and D_2 are binary. Then $P(D_2 = j_2|D_1 = j_1, Y_1, \bar{X}_2)$, $j_1, j_2 = 0, 1$, can be estimated by probit or logit estimation with $(Y_1, \bar{X}_2)$ as regressors on the subpopulation with $D_1 = j_1$, and $P(D_2 = j_2|D_1 = j_1, \bar{X}_2)$ can be obtained by averaging out Y_1 (or by assuming another probit or logit model without Y_1). With these,

$$\omega(\bar{X}_{2i}, D_{1i} = j_1, D_{2i} = j_2, Y_{1i}) \equiv \left\{\sum_{j_2=0}^{1}\sum_{j_1=0}^{1}\frac{P(D_{2i} = j_2|D_{1i} = j_1, \bar{X}_{2i})}{P(D_{2i} = j_2|D_{1i} = j_1, Y_{1i}, \bar{X}_{2i})}\mathbf{1}(D_{1i} = j_1, D_{2i} = j_2)\right\}^{1/2}$$

where $\mathbf{1}(A) = 1$ if A holds and $\mathbf{1}(A) = 0$ otherwise. In both continuous and discrete treatment cases, observations with too small denominators in the weight should be removed, which is one shortcoming of the WLS estimator. For instance, we may use only the observations with

$$f(D_{2i}|D_{1i}, Y_{1i}, \bar{X}_{2i}) > 0.001$$

(or $\min_{j_1, j_2}\{P(D_{2i} = j_2|D_{1i} = j_1, Y_{1i}, \bar{X}_{2i})\} > 0.001$ for discrete D_1 and D_2).

2.3.2. Instrumental variable estimator for Y_2 reduced form, two-stage estimator for Y_1 and Y_2 structural form and two-stage estimator for augmented Y_2 reduced form

The WLS estimator for the Y_2 RF (3) can be cumbersome and unstable if some weights become inflated because of too small denominators. Screening out those observations brings arbitrariness into the procedure. Also, estimating the weight requires a distributional assumption on $D_2|(D_1, Y_1, \bar{X}_2)$ unless the dimension of $\bar{X}_2$ is small. But an IVE can avoid these. Since the motivation for WLS is to overcome the endogeneity of D_2, suppose that there is a variable that

(a) affects D_2,
(b) does not affect Y_2 directly and
(c) is uncorrelated with the Y_2 RF error $\beta_y U_1 + U_2$.

Then it is an instrumental variable (IV) for D_2 , and we can apply the IVE to the Y_2 RF. For instance, time varying components of X_1 may qualify as an IVE for D_2 if the treatment (D_2) decision is based on the entire history of the individual including X_1 which, however, does not affect directly Y_2 as X_2 is already in the Y_2-model. This IVE trades off assumptions with the WLS estimator, because the three assumptions are additional requirements, whereas the complications that are brought in by weighting are absent in the IVE.

Suppose that we estimate the Y_1 and Y_2 SFs with the LSE to find α_d, $\beta_{d\text{lag}}$, β_d and β_y. Then we can easily construct the total effect $E(Y_2^{d_1 d_2} - Y_2^{00}) = (\beta_{d\text{lag}} + \beta_y \alpha_d) d_1 + \beta_d d_2$. Since the two SFs are estimated separately and then the total effect is constructed next, this is a TSE. Depending on the model and data at hand, the LSE may not be valid. For instance, if $U_{1i} = \pi_i + V_{1i}$ and $U_{2i} = \pi_i + V_{2i}$ (recall Fig. 6) for time varying errors V_{1i} and V_{2i}, then Y_1 in the Y_2 SF is endogenous. In this case an IVE may be applied to the Y_1 and Y_2 SFs so long as appropriate IVs exist; for example, time varying components of X_t can be used as IVs as just noted. Instead of an IVE, the first-differenced model free of π_i may be estimated by an LSE, although this loses all time constant regressors along with π_i.

Recalling RF(3), the source of the D_2-endogeneity is U_1 in the error term $\beta_y U_1 + U_2$. The 'control function' approach (see Lee (2012) and the references therein) is to add a term to control or remove the source of endogeneity, and the right term in this case is U_1 so that only U_2 remains in the error term of RF (3). Although U_1 is not observed, it can be estimated by the residual $\hat{U}_1$ (from the LSE or IVE to the Y_1 SF). Hence this procedure yields another TSE: the first stage is obtaining $\hat{U}_1$, and the second stage is the LSE to the Y_2 RF with $\hat{U}_1$ as an additional regressor. The 'structural nested mean model' approach of Almirall *et al.* (2010) is a control function approach, as they added the residuals for the mediating variables (in our model, Y_1 is the only mediating variable). In their approach, the regression function part is viewed as the addition of two 'blip functions' that are the effects of D_1 and D_2 whereas the added residuals are nuisance terms that are not interesting on their own.

2.3.3. Non-parametric direct and indirect effects and linear models

Although we derived direct and indirect effects by using linear models, non-parametric definitions of those effects do exist. To simplify the discussion, consider a treatment D, a 'mediator' W and a response variable Y, where D can affect both W and Y, and W can affect Y. Let W_d be the potential W when $D = d$, and Y_{dw} the potential Y when $D = d$ and $W = w$. Our interest is in the effect of changing D from d to d' (denoted '$d \to d'$'). Note that

$$Y_d = Y_{dW_d}:$$

Y_d is Y with $D = d$ exogenously set whereas W takes whatever value it naturally takes under $D = d$.

Write the individual total effect as

$$Y_{d'} - Y_d = Y_{d'W_{d'}} - Y_{dW_d} = (Y_{d'W_{d'}} - Y_{dW_{d'}}) + (Y_{dW_{d'}} - Y_{dW_d}),$$

subtracting and adding the counterfactual $Y_{dW_{d'}}$. The first part $Y_{d'W_{d'}} - Y_{dW_{d'}}$ is the direct effect of $d \to d'$ while W is kept at the value that it would take under $D = d'$. The second part $Y_{dW_{d'}} - Y_{dW_d}$ is the indirect effect of $d \to d'$ through W while D is kept at the baseline value d. Taking $E(\cdot)$ turns the individual effects into the mean effects, and we shall not further mention this in the rest of this section.

The awkward aspect in the above decomposition is that the direct effect has $W_{d'}$, not W_d, as we would rather have W_d there with d being the baseline value of D. Prompted by this, we may decompose $Y_{d'} - Y_d$ alternatively by using $Y_{d'W_d}$:

$$Y_{d'} - Y_d = (Y_{d'W_{d'}} - Y_{d'W_d}) + (Y_{d'W_d} - Y_{dW_d}).$$

The first part is the indirect effect of $d \to d'$ with D kept at d', and the second part is the direct effect of $d \to d'$ with W kept at the value that it takes under $D = d$. In this new decomposition, the indirect effect is not satisfactory because D is kept at d', not at the baseline value d.

In view of these problems, one may define the direct and indirect effects as respectively $Y_{d'W_d} - Y_{dW_d}$ and $Y_{dW_{d'}} - Y_{dW_d}$ where d is used whenever D should be held constant. But, since $Y_{d'W_{d'}}$ does not appear at all here, these direct and indirect effects do not add up to the total effect $Y_{d'} - Y_d$. One way out of this predicament is to define the total effect as the *difference* between the above direct effect of $d \to d'$ and the indirect effect of the reverse change $d' \to d$ as in Pearl (2010), page 42:

$$Y_{d'W_d} - Y_{dW_d} - (Y_{d'W_d} - Y_{d'W_{d'}}) = Y_{d'W_{d'}} - Y_{dW_d} = Y_{d'} - Y_d.$$

Although this restores the decomposition, it is 'unpalatable' because it involves two opposite changes $d \to d'$ and $d' \to d$.

For example, consider a simple linear structural model: with parameters β_d and β_w,

$$Y = \beta_d D + \beta_w W;$$

an error U can be added, but U will have no role in the following discussion. From this,

$$Y_{d'} = \beta_d d' + \beta_w W_{d'}, \qquad Y_d = \beta_d d + \beta_w W_d \Rightarrow Y_{d'} - Y_d = \beta_d(d' - d) + \beta_w(W_{d'} - W_d);$$
$$Y_{d'W_{d'}} - Y_{dW_{d'}} = \beta_d(d' - d) = Y_{d'W_d} - Y_{dW_d}$$

(same direct effect regardless of W);

$$Y_{d'W_{d'}} - Y_{d'W_d} = Y_{dW_{d'}} - Y_{dW_d} = \beta_w(W_{d'} - W_d)$$

(same indirect effect regardless of D). The indirect effect of the reverse change $d' \to d$ is $\beta_w(W_d - W_{d'})$, which equals -1 times the indirect effect of $d \to d'$—this does not necessarily hold for non-linear models in general. Hence the total effect of $d \to d'$ is the sum of the direct and indirect effects.

Consider now a linear structural model with interaction term DW:

$$Y = \beta_d D + \beta_w W + \beta_{dw} DW.$$

From this,

$$Y_{d'} = \beta_d d' + \beta_w W_{d'} + \beta_{dw} d' W_{d'}, \qquad Y_d = \beta_d d + \beta_w W_d + \beta_{dw} d W_d$$
$$\Rightarrow Y_{d'} - Y_d = \beta_d (d' - d) + \beta_w (W_{d'} - W_d) + \beta_{dw} (d' W_{d'} - d W_d);$$
$$Y_{d'W_{d'}} - Y_{dW_{d'}} = \beta_d (d' - d) + \beta_{dw} (d' - d) W_{d'}, \qquad Y_{d'W_d} - Y_{dW_d} = \beta_d (d' - d) + \beta_{dw} (d' - d) W_d;$$
$$Y_{d'W_{d'}} - Y_{d'W_d} = \beta_w (W_{d'} - W_d) + \beta_{dw} d' (W_{d'} - W_d),$$
$$Y_{dW_{d'}} - Y_{dW_d} = \beta_w (W_{d'} - W_d) + \beta_{dw} d (W_{d'} - W_d).$$

The direct effects differ depending on W, and the indirect effects also differ depending on where D is fixed. The indirect effect of the reverse change $d' \to d$ with D fixed at d' is no longer the same as -1 times the indirect effect of $d \to d'$ with D fixed at d.

Although linear structural models are parametric and thus restrictive as such, this section demonstrates why linear models are valuable. Firstly, they illustrate well the direct and indirect effects. Although a total effect may look like a 'black box', decomposing it into the subeffects helps in understanding the black box. Secondly, when the decomposition is not unique, linear models augmented by interaction terms explain why, as has just been done. Pearl (2010), pages 46–47, provided an example where a simple linear approximation fails but a linear approximation with interaction terms works in identifying the desired indirect effect. Thirdly, when there are many covariates and confounders, linear models (augmented by interaction terms) seem to be the only operational means of estimation.

3. Dynamic effects of treatments under anticipation

3.1. Expectation-augmented model

Consider a linear model that is the expectation-augmented version of expression (1):

$$\left.\begin{aligned} Y_1^{d_1} &= \alpha_1 + \bar{X}_1' \alpha_x + \alpha_d d_1 + \alpha_{de} E(D_2^{d_1} | I_1) + U_1, \\ D_2^{d_1} &= \zeta_1 + X_2' \zeta_x + \zeta_y Y_1^{d_1} + \varepsilon_2, \\ Y_2^{d_1 d_2} &= \beta_1 + X_2' \beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_{de} E(D_3^{d_1 d_2} | I_2) + \beta_y Y_1^{d_1} + U_2 \end{aligned}\right\} \tag{4}$$

where I_t is the information that is available at period t, *after* D_t has been determined but before Y_t. The linear model SF (4) nests SF (1) as a special case when $\alpha_{de} = 0 = \beta_{de}$. In the example of spanking and child poor behaviour, $\beta_{d\text{lag}} = \beta_d = 0$ and $\beta_{de} < 0$ (as well as $\alpha_d = 0$ and $\alpha_{de} < 0$) means that spanking reduces poor behaviour only because the children are forward looking, whereas $\alpha_{de} = \beta_{de} = 0$ means that there is no anticipation effect.

In the observed Y_1-equation, $Y_1 = \alpha_1 + \bar{X}_1' \alpha_x + \alpha_d D_1 + \alpha_{de} E(D_2^{d_1} | I_1) + U_1$, suppose that $E(D_2^{d_1} | I_1) = \bar{X}_1' \lambda_x + \lambda_d D_1$ to obtain the Y_1 RF in the observed variables:

$$Y_1 = \alpha_1 + \bar{X}_1' (\alpha_x + \alpha_{de} \lambda_x) + (\alpha_d + \alpha_{de} \lambda_d) D_1 + U_1$$

where the coefficient of D_1 is the sum of α_d and the anticipation contribution $\alpha_{de} \lambda_d$. With this equation, we can estimate $\alpha_d + \alpha_{de} \lambda_d$. This may not matter, but there are cases where separating $\alpha_{de} \lambda_d$ from α_d may be of interest.

For instance, suppose that D is the rate for income tax and Y is work hours. An increase in current tax rate will affect the current work hours by α_d, but the increase in tax rate may lead to an increase in the expected future tax rate, which in turn affects the current work hours by $\alpha_{de} \lambda_d$; the latter part is the anticipation part. If the government desires to alter the work hours only by α_d, then an announcement of no further increase in tax rate can be made (assuming that the government is credible). Another example for crime and punishment is that it is interesting

to know what proportion of incarceration's effect on crime is due to the current incapacitation (no crime while in jail) and what proportion is due to prevention (anticipation). To separate α_{de} from $\alpha_d + \alpha_{de}\lambda_d$, $E(D_2^{d_1}|I_1)$ needs a variation that is independent of $\bar{X}_1$ and D_1, which calls for a variable affecting $E(D_2^{d_1}|I_1)$, but not Y_1 directly.

We may allow an expectation term such as $E(Y_2^{d_1 0}|I_1)$ in the $D_2^{d_1}$-equation, because a poor expected outcome when not treated in period 2 would affect $D_2^{d_1}$. But the equations of interest are the Y_1- and Y_2-equations, not the D_2-equation, and we can take the above D_2-equation (4) as an RF with $E(Y_2^{d_1 0}|I_1)$ substituted out, i.e. although anticipation may matter for both D and Y, we explicitly take anticipation into account only in the Y-equations. When an intervention is imposed on the treatment, the existing treatment equation becomes irrelevant while the structural Y-equations still stand; this 'autonomy' (see Pearl (2009) and the references therein) is the reason why we are interested in the structural Y-equations. Note that using expectations does not imply that the future literally affects the present—it is only the 'currently expected version of the future' based on the present variables that matters.

There are at least two ways to find $E(D_2^{d_1}|I_1)$ and $E(D_3^{d_1 d_2}|I_2)$. One is estimating them non-parametrically by using D_2, D_3 and variables in I_1 and I_2. This is an RF approach, as we do not use any information on how the expectations are generated by the individuals. The other is obtaining $E(D_2^{d_1}|I_1)$ and $(D_3^{d_1 d_2}|I_2)$ by solving expression (4); this is done in Appendix A. The latter approach needs the 'rational expectation' assumption that the individuals know the model (4) in forming $E(D_2^{d_1}|I_1)$ and $E(D_3^{d_1 d_2}|I_2)$. Appendix A shows that solving expression (4) requires specifying the $D_3^{d_1 d_2}$-equation and imposing parameter restrictions $\zeta_y \alpha_{de} \neq 1$ and $\delta_y \beta_{de} \neq 1$ where δ_y is the coefficient of $Y_2^{d_1 d_2}$ in a linear structural equation for $D_3^{d_1 d_2}$. These are the disadvantages of the system solving approach. But, if the information sets I_1 and I_2 are high dimensional, non-parametrically estimating $E(D_2^{d_1}|I_1)$ and $E(D_3^{d_1 d_2}|I_2)$ is out of the question, leaving no choice other than to use the system solving approach.

3.2. Difficulties due to expectation terms

Although α_d, α_{de}, β_d and β_{de} show the direct effects, if we want to know the total effect, then the indirect effect must be derived as in the no-anticipation case. But this task is complicated owing to $E(D_2^{d_1}|I_1)$ and $(D_3^{d_1 d_2}|I_2)$. To appreciate the difficulty, substitute the $Y_1^{d_1}$ SF in expression (4) into the $Y_2^{d_1 d_2}$ SF to obtain

$$\begin{aligned} Y_2^{d_1 d_2} = {} & \beta_1 + \beta_y \alpha_1 + \beta_y \bar{X}_1' \alpha_x + X_2' \beta_x + (\beta_{d\text{lag}} + \beta_y \alpha_d) d_1 + \beta_d d_2 + \beta_y \alpha_{de} E(D_2^{d_1}|I_1) \\ & + \beta_{de} E(D_3^{d_1 d_2}|I_2) + \beta_y U_1 + U_2. \end{aligned} \tag{5}$$

As $\beta_y \alpha_{de} E(D_2^{d_1}|I_1)$ suggests, there are the indirect effects of treatments (d_1, d_2) through the future expected treatments, which are shown in detail in Appendix A. For the spanking example, when parents spank their child, they are aware of the possibility that the current spanking reduces bad behaviour, which reduces the expected future spanking, which in turn increases the future bad behaviour and so on.

The difficulty is not just finding the direct and indirect effects due to anticipation. There might be also a 'deeper' difficulty that is associated with anticipation: an infinite series of interplays. For example, consider a government policy D. If people know that the government incorporates their expectations on the policy into the policy formation, then this would alter the government's $D_3^{d_1 d_2}$-equation, which then alters the people's expectations, which in turn alters the $D_3^{d_1 d_2}$-equation, *ad infinitum*. This warrants caution in assessing anticipation effects, and a complete solution to this complex query is left for future research.

Given these difficulties, we may as well be content with the total effect without the anticipation aspect and then try to see whether $\alpha_{de}=0$ and $\beta_{de}=0$ or not, which may be taken as a sensitivity analysis (into the direction of anticipation effects) for the conventional dynamic treatment effect analysis without anticipation.

3.3. Instrumental variables estimator and non-parametric estimation for expectations

Although accounting for all effects is complicated, estimating the linear model with the conditional expectations is not necessarily difficult, so long as there are regressors that affect treatments but not the responses. To see this, rewrite the observed versions of the $Y_1^{d_1}$- and $Y_2^{d_1d_2}$-equations in expression (4) as

$$Y_1=\alpha_1+\bar{X}_1'\alpha_x+\alpha_d D_1+\alpha_{de}D_2+[\alpha_{de}\{E(D_2|I_1)-D_2\}+U_1],$$
$$Y_2=\beta_1+X_2'\beta_x+\beta_{d\text{lag}}D_1+\beta_d D_2+\beta_{de}D_3+\beta_y Y_1+[\beta_{de}\{E(D_3|I_2)-D_3\}+U_2]$$

where the terms in square brackets are the new error terms. Consider an instrument Z_1 such that

$$E([\alpha_{de}\{E(D_2|I_1)-D_2\}+U_1]Z_1)=0.$$

A variable qualified for Z_1 is a variable in I_1 that is uncorrelated with U_1, excluded from the Y_1 SF but included in the D_2-equation; Z_1 can be used as an IV for D_2 in the last Y_1-equation. Analogously, we can think of Z_2 in I_2 that is uncorrelated with U_2, excluded from the Y_2 SF but included in the D_3-equation; Z_2 can be used as an IV for D_3 in the last Y_2-equation. In the spanking behaviour example, test scores at period t may influence the spanking D_t, but not the behaviour Y_t directly; i.e. test scores may serve as IVs.

With the IVs, the TSE of Huang and Lee (2010) estimating the last Y_1 and Y_2 SFs separately can be done; no other method such as WLS is being considered here. One may think that estimating $E(D_2^{d_1}|I_1)$ and $E(D_3^{d_1d_2}|I_2)$ non-parametrically does not require any IVs. But, even in this case, IVs are necessary because $E(D_2^{d_1}|I_1)$ and $E(D_3^{d_1d_2}|I_2)$ should have variations that are independent of the other regressors in the Y_1 and Y_2 SFs. In view of this, an IVE should be preferred to the non-parametric estimation. To dissipate any doubts on whether this idea works or not, Appendix A provides simulation evidence that this idea, as well as the no-anticipation methods, works by using simple models that are almost the same as expressions (1) and (4). The total effect under anticipation is presented in equation (9) in Appendix A that includes the total effect without anticipation as a special case when $\alpha_{de}=\beta_{de}=0$.

4. Empirical analysis: spanking and child bad behaviour

4.1. Data

The National Longitudinal Survey of Youth child sample that was first sampled in 1979 contains rich information on children born to the women respondents of the surveys in the USA. Starting from 1986, the National Longitudinal Survey of Youth collected information on the cognitive, social and behavioural development of the children as well as their family backgrounds and detailed home inputs. The surveys were conducted every 2 years, which enables us to obtain detailed information on home inputs and family backgrounds when a child was 2–3, 4–5, 6–7 and 8–9 years old.

A child's behaviour problems at ages above 4 years are measured by *behaviour problems index* standard scores, BPI, which are based on answers from the mother to 28 questions about poor

behaviours of the child in the previous 3 months. The items include difficulties interacting with other children, difficulties concentrating, being too dependent or clingy, having a strong temper and being argumentative. The questionnaire used three response categories: 'often true', 'sometimes true' and 'not true'. The responses are then dichotomized (categories often true and sometimes true are 1, and not true is 0) and then summed to produce BPI.

The survey question on spanking 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 children, although its frequency decreases as a child grows. In our sample of about 4000 children surveyed from 1986 to 1998, 87% of the mothers spanked their toddlers at least once in the past week, whereas only 68% spanked their 5-year-olds. The corresponding proportions are 46% for ages 6-7 years and 28% for ages 8–9 years. Most children, however, are spanked modestly. For example, over 96% of the children at ages 6–7 years and 91% at ages 4–5 years were spanked not more than four times. These children, including non-spanked children, will be the focus of our study, since it is the effects of modest spanking that are debated most (Baumrind *et al.*, 2002).

Our main working sample contains 2436 children who have no missing values in BPI-scores and were spanked four times or fewer a week at ages 4–5 and 6–7 years, including the non-spanked. Because all children in the main working sample were spanked not more than four times a week, and also because the answered spanking frequency may not be representative of the 'regular' spanking frequency—recall that the questionnaire is only for the past week—we also use a binary variable for spanking (1 if spanked at all and 0 otherwise). Estimation results by using both spanking frequencies and their binary versions will be presented.

Among the four age-based periods 2–3, 4–5, 6–7 and 8–9 years, the first period is ages 6–7 years, the second is ages 8–9 years, and the ages 4–5 years are then used as the base period (period 0); the earliest period, for ages 2–3 years, is not used because BPI-scores are not available. Since the spanking question refers to the week before the survey date whereas BPI comes from the past 3 months, to assure the temporal order $X_t \to D_t \to Y_t$, we set

$$X_3, X_2, X_1, X_0, \text{covariates at ages 8–9, 6–7, 4–5 and 2–3 years,}$$
$$D_3, D_2, D_1, D_0, \text{spanking at ages 8–9, 6–7, 4–5 and 2–3 years,}$$
$$Y_2, Y_1, Y_0, \text{BPI at ages 8–9, 6–7 and 4–5 years.}$$

Spanking at ages 8–9 years and the covariates at ages 8–9 years are used only for the anticipation analysis. The variables at ages 2–3 years are not used in principle, but they are used occasionally to gain more insight or to control or remove sources of endogeneity better. Also 'child temperament scores' for insecurity and compliance that are available for ages 2–3 years, which are related to BPI, will be used.

4.2. Preliminary analysis

The summary statistics of the variables in the main working sample are in Table 1 (some are also in Table 2). BPI has a mean of 106.0 (SD 14.7) for children of ages 8–9 years and 105.3 (SD 14.2) for children of ages 6–7 years. The average spanking frequency is 0.62 (SD 0.96) times a week at ages 8–9 years and 1.28 times (SD 1.22) at ages 6–7 years.

The link between spanking and behaviour problems is controversial (e.g. Gershoff (2002)). The often observed positive relationship between the two might be driven by the fact that children with more behaviour problems tend to be spanked more often, even though spanking may be helpful in reducing behaviour problems in future. The difficulty in establishing the causal link is the endogeneity of spanking, which may arise from various sources. For example, inappropriate home inputs may induce poor behaviour of children and more spanking from the parents.

Table 1. List of variables and summary statistics

Variable	*Mean (SD)*	*Size*
BPI at ages 8–9 years	106.0 (14.7)	2436
BPI at ages 6–7 years	105.3 (14.2)	2436
BPI at ages 4–5 years	105.1 (14.6)	2284
Child temperament score—insecure at ages 1–3 years	20.0 (4.37)	1049
Child temperament score—compliance at ages 1–3 years	21.9 (4.84)	1003
Spanked number of times last week at survey time at age 6–7 years	0.62 (0.96)	2436
Spanked number of times last week at survey time at age 4–5 years	1.28 (1.22)	2436
Spanked number of times last week at survey time at age 2–3 years	2.63 (3.19)	1192
A child was spanked at least once at age 6–7 years	0.38 (0.49)	2436
A child was spanked at least once at age 4–5 years	0.65 (0.48)	2436
A child was spanked at least once at age 2–3 years	0.85 (0.35)	1192
Child demographic information		
Race of child: black or Hispanic	0.51 (0.50)	2436
Sex of child: boy	0.50 (0.50)	2436
Birth order of child	1.89 (0.98)	2436
Home inputs for 8–9-year-olds		
Child has 10 or more children's books at home	0.86 (0.35)	2378
How often mother reads to child: at least 3 times a week	0.58 (0.49)	2379
How often child reads for enjoyment: every day	0.33 (0.47)	2378
Family encourages hobbies	0.93 (0.26)	2374
Child has special lessons or activities	0.61 (0.49)	2372
How often child taken to museum: at least several times a year	0.42 (0.49)	2376
How often child taken to performance: at least several times a year	0.35 (0.48)	2425
How often family meets with relatives or friends: at least 2–3 times per month	0.58 (0.49)	2372
How often child with father outdoors: at least once a week	0.48 (0.50)	2267
How often child eats with mother and father: at least once a day	0.56 (0.50)	2278
Number of hours per weekday child watches television	4.55 (5.31)	2349
Number of hours per weekend day child watches television	4.74 (3.60)	2357
Parents discuss television programmes with child	0.83 (0.38)	2359
Number of times past week grounded child	0.52 (1.72)	2346
Number of times past week took away television	0.60 (2.46)	2343
Number of times past week praised child	6.91 (10.13)	2340
Number of times past week took allowance	0.24 (2.86)	2331
Number of times mother showed child physical affection	16.74 (20.90)	2320
Number of times past week sent child to room	1.20 (3.46)	2345
Number of times past week said positive things	6.36 (10.89)	2330
Home inputs for 6–7-year-olds		
Child has 10 or more children's books at home	0.83 (0.38)	2429
How often mother reads to child: at least 3 times a week	0.72 (0.45)	2429
How often child reads for enjoyment: every day	0.70 (0.46)	2424
Family encourages hobbies	0.89 (0.31)	2427
Child has special lessons or activities	0.50 (0.50)	2426
How often child taken to museum: at least several times a year	0.76 (0.42)	2432
How often child taken to performance: at least several times a year	0.59 (0.49)	2425
How often family meets with relatives or friends: at least 2–3 times per month	0.60 (0.49)	2424
How often child with father outdoors: at least once a week	0.52 (0.50)	2354
How often child eats with mother and father: at least once a day	0.78 (0.42)	2396
Number of hours per weekday child watches television	4.28 (5.23)	2073
Number of hours per weekend day child watches television	4.48 (4.11)	2075
Parents discuss television programmes with child	0.83 (0.38)	2387
Number of times past week grounded child	0.47 (1.08)	2416
Number of times past week took away television	0.51 (1.05)	2414
Number of times past week praised child	7.70 (10.94)	2399

(*continued*)

Table 1 (*continued*)

Variable	*Mean (SD)*	*Size*
Home inputs for 6–7-year-olds		
Number of times past week took allowance	0.12 (0.73)	2391
Number of times mother showed child physical affection	17.31 (20.06)	2378
Number of times past week sent child to room	1.36 (2.77)	2419
Number of times past week said positive things	6.22 (9.29)	2395
Home inputs for 4–5-year-olds		
How many books does child have: 10 or more books	0.80 (0.40)	2431
How often mother reads to child: at least 3 times a week	0.55 (0.49)	2431
How often child taken to museum: at least several times a year	0.70 (0.46)	2422
How many magazines does family take: 3 or more	0.57 (0.49)	2423
Does child have record or tape player	0.76 (0.43)	2412
How often child taken on outing: at least several times a week	0.55 (0.50)	2418
How often child eats with mother and father: at least once a day	0.73 (0.44)	2331
Number of hours television is on per day	5.74 (5.01)	2105
Home observation measurement of the environment variable (HOME)	202.1 (36.5)	2384
Mother helps child to learn numbers	0.94 (0.23)	2435
Mother helps child to learn alphabet	0.92 (0.27)	2436
Mother helps child to learn colours	0.94 (0.24)	2435
Mother helps child to learn shapes	0.83 (0.38)	2436
Mother responds to hit—hit child back	0.15 (0.35)	2436
Mother responds to hit—send to room	0.49 (0.50)	2436
Mother responds to hit—talk to child	0.72 (0.45)	2436
Mother responds to hit—ignore it	0.02 (0.15)	2436
Mother responds to hit—give chores	0.05 (0.21)	2436
Mother responds to hit—take allowance	0.03 (0.18)	2436
Mother responds to hit—hold child's hands	0.12 (0.32)	2436
Home inputs for 2–3-year-olds		
How often child taken out of the house: every day	0.58 (0.49)	1614
How many children's books child has: 10 or more books	0.76 (0.43)	1614
How often mother reads to child: at least 3 times a week	0.56 (0.50)	1606
How often mother takes child to grocer: once a week or less	0.63 (0.48)	1604
How many cuddly or role-playing toys	16.1 (13.55)	1597
How many push or pull toys child has	7.19 (6.72)	1597
How often child eats with both mother and father: at least once a day	0.70 (0.46)	1497
How often mother talks to child while working: often	0.56 (0.50)	1608
Home observation measurement of the environment variable (HOME)	139.8 (23.1)	1878
Family background		
Mother's AFQT-score† taken in 1981	38.1 (27.1)	2350
Mother's highest grade at 1988	12.4 (2.1)	2390
Family salary income in 1988	6962.4 (8342.2)	2376
Family income in 1988	26298 (18532)	2082
Mother's age at child birth	24.5 (3.3)	2436
Child is breast fed	0.49 (0.50)	2342
Mother lived in the south at age 14 years	0.35 (0.48)	2337

†AFQT, armed forces qualification test.

If the detailed home inputs are not properly controlled, then the positive correlation between spanking and behaviour problems may be spurious.

Table 2 shows close relationships between BPI-scores, spanking frequencies, home inputs and family backgrounds. White children are spanked on average 23% less at ages 4–5 years and 35% less at ages 6–7 years than the others, and their BPI-scores are around 2 points lower than for non-white children. Boys have more behaviour problems than girls and are also spanked a little

Table 2. Spanking and child behaviour problems index BPI: summary statistics†

Statistic	*Weekly spanking frequency*		*BPI total standard scores*		*Group size*
	At age 6–7 years	*At age 4–5 years*	*At age 8–9 years*	*At age 6–7 years*	
Main sample	0.62 (0.96)	1.28 (1.22)	106.0 (14.7)	105.3 (14.2)	2436
Race					
White	0.49 (0.84)	1.11 (1.18)	105.1 (14.5)	104.3 (14.0)	1180
Non-white	0.75 (1.04)	1.44 (1.24)	106.9 (14.9)	106.3 (14.5)	1256
Sex					
Boy	0.69 (1.00)	1.34 (1.23)	107.2 (14.9)	106.4 (14.6)	1225
Girl	0.55 (0.90)	1.22 (1.21)	104.8 (14.4)	104.1 (13.8)	1211
Birth order					
First-borns	0.57 (0.89)	1.30 (1.20)	105.8 (14.1)	105.4 (13.6)	1027
Others	0.66 (1.00)	1.27 (1.24)	106.2 (15.1)	105.2 (14.7)	1409
How many children's books a child has at home at age 6–7 years					
$\geqslant 10$	0.55 (0.90)	1.18 (1.20)	105.1 (14.2)	104.4 (13.9)	2016
< 10	0.96 (1.15)	1.75 (1.24)	110.4 (16.0)	109.5 (15.1)	413
Mother reads to child at age 6–7 years					
Often	0.58 (0.92)	1.23 (1.22)	105.1 (14.6)	104.1 (14.0)	1744
Not often	0.73 (1.04)	1.41 (1.21)	108.4 (14.6)	108.2 (14.5)	685
Number of hours television on per day at ages 4–5 years					
< 5	0.51 (0.89)	1.07 (1.19)	104.6 (14.5)	103.8 (14.1)	1149
$\geqslant 5$	0.72 (1.01)	1.46 (1.21)	107.2 (14.8)	106.4 (14.3)	1205
HOME scores at ages 4–5 and 1–3 years					
Above mean	0.36 (0.71)	0.86 (1.06)	102.0 (13.4)	101.6 (13.2)	912
Below mean	0.91 (1.12)	1.67 (1.28)	109.6 (16.6)	108.7 (14.8)	563
Mother's highest grade					
$\geqslant 16$	0.45 (0.81)	0.97 (1.15)	103.5 (14.2)	103.7 (14.6)	569
< 16	0.67 (0.99)	1.38 (1.23)	106.8 (14.8)	105.8 (14.1)	1867
Mother's AFQT-score in 1981					
Above mean	0.44 (0.79)	1.04 (1.17)	104.6 (13.9)	103.4 (13.2)	1062
Below mean	0.76 (1.05)	1.47 (1.23)	107.1 (15.2)	106.8 (14.9)	1374
Mother's birth age					
$\geqslant 25$ years	0.48 (0.85)	0.94 (1.16)	104.1 (14.6)	103.3 (14.1)	1257
< 25*years*	0.77 (1.05)	1.65 (1.18)	108.0 (14.5)	107.5 (14.1)	1179
Mother's residence at age 14 years					
South	0.77 (1.02)	1.44 (1.27)	106.8 (15.2)	105.7 (14.6)	823
Not south	0.55 (0.93)	1.21 (1.19)	105.5 (14.4)	105.2 (14.1)	1514

†The entries are group means and standard deviations (in parentheses). The main sample is composed of children from the National Longitudinal Survey of Youth child sample who were spanked four times or less a week between ages 6 and 9 years.

more: their BPI-scores are 2.3 points higher than for girls and they are spanked about 20% more at ages 6–7 years. There are virtually no differences in BPI between the first-born child and the others. The small group of children with fewer than 10 children's books at home are spanked 33% more at ages 4–5 years and 43% more at ages 6–7 years than those with more books, and they have much higher BPI-scores (5.2 points higher, or 36% of 1 SD) at ages 6-9 years. Children

whose mothers read to them frequently (at least three times a week) are spanked slightly less than the others and have less behaviour problems (BPI-scores around 3.3–4.1 points lower). The number of hours that a television is on at home when a child is 4–5 years old is also related to behavioural problems; children from homes with a television switched on for more than 5 h per day are spanked 27% more and have 2.6 points higher BPIs at ages 4–5 and 6–7 years.

Children with higher quality home environments at ages 2–5 years (measured by the age-specific 'home observation measurement of the environment' variable HOME (which is often used in child development studies as an aggregate quality indicator of home environment) are spanked much less (49% less at ages 4–5 years and 60% less at ages 6–7 years) and have over 7.1 lower BPIs at ages 6–9 years (about 50% of 1 SD) than the others. In short, children with high quality home inputs are spanked less and have fewer behaviour problems.

Children with mothers having at least 16 years of schooling have lower BPI-scores (about 3.3 lower) at ages 8–9 years and 41% lower spanking frequency at ages 4–5 years. A similar gap exists between children whose mothers have above average armed forces qualification test scores (AFQT-scores) and the others. Children who were born to mothers below 25 years old are spanked about 40% more at ages 4–7 years and their BPI-scores are 4 points higher at ages 6–9 years. It is interesting to note that mothers who lived in the south at age 14 years are more likely than others to spank their children (29% more) at ages 6–7 years, even though there is little difference between BPI-scores. In short, the characteristics of mothers, including their behaviours and experiences, seem to matter.

An important message from Table 2 is that frequently spanked children have disadvantages in terms of important home inputs and family backgrounds, which must be taken into account for causal effects of spanking on BPI-scores. The strength of our data is that a rich set of home inputs up to age 9 years as well as key family background variables are available. This would greatly reduce, though not completely remove, the potential biases due to omitted variables. Most home input variables were categorical or ordinal, but not cardinal, which were thus converted to dummy variables.

The availability of a rich set of control variables in our data is also crucial for the application of our estimation methods, which rely on the credibility of the NUC assumption. As listed in Table 1, our control variables have three groups:

(a) child demographic information including race, sex and birth order;
(b) home inputs at ages 2–3, 4–5, 6–7 and 8–9 years, which cover many aspects of a child's life at home, such as the number of child books at home, how often the mother reads to the child, how often the child is taken to museums, performances, outdoor activities, meeting with relatives, having meals with parents, etc.;
(c) family backgrounds including the mother's AFQT-score, the highest grade, age at the child's birth, whether the child was breast fed and family income.

Furthermore, two child temperament scores around ages 1–3 years, indicating the degree of a child's insecurity and compliance, are also controlled in the estimation in addition to BPI-scores at ages 4–5 years. These variables collectively constitute a highly comprehensive set of controls.

An important identification variation for spanking effects, after controlling for a large set of home inputs and family backgrounds, comes from different parental beliefs about spanking effects. This belief itself does not affect child behaviour problems, but it can directly affect spanking. For example, two identical children may be spanked differently only because one child's parents believe in 'spare the rod and spoil the child', whereas the other's believe that spanking is morally wrong. The literature shows that such discrepancy in beliefs is substantial.

For instance, 61% of parents in a recent US survey viewed spanking as an acceptable form of discipline whereas 94% held this view in the late 1960s. Even among psychologists in clinical practice, about a third thought that the American Psychological Association should definitely have a policy opposing corporal punishments, whereas another third thought that the American Psychological Association should definitely not (Straus and Mather, 1996; Benjeta and Kazdin, 2003). It is useful to note that the different beliefs that were reported in these surveys may not only capture differences in people's initial beliefs, which are the source of identification here, but may also reflect views adjusted through their experiences with spanking.

4.3. Empirical results

4.3.1. Baseline results of spanking effects

The main results are in Table 3. On the basis of the upper panel estimates, the total effects of spanking once a week are shown in the lower panel. The first column 'Last lag' applies the LSE to the Y_2 RF (3); it is called last lag because Y_0 appears as part of $\bar{X}_1$ when Y_1 is substituted out. The control variables include a child's race, sex, birth order, family background variables and detailed home inputs at ages 6–7 and 8–9 years as well as indicators of earlier home environments before 5 years of age and two child temperament scores around age 3 years. The coefficients of all the spanking variables at both ages 4–5 and 6–7 years are statistically significant below the 5% level, which gives the total effect of spanking $d = (d_1, d_2)$ at ages 4–5 and 6–7 years:

$$-12.80 + 10.56d_1 - 1.91d_1^2 + 11.39 - 8.10d_2 + 1.64d_2^2.$$

The estimates in the same column in the lower panel suggest that the total effect of spanking once a week at ages 4–5 years ($d_1 = 1; d_2 = 0$) is -4.16 in BPI-score at ages 8–9 years, compared with no spanking at all, which is significant at the 10% level. The effect of spanking at ages 6–7 years, in contrast, is positive and significant, which makes the aggregate effect of spanking once a week at both ages 4–5 and 6–7 years positive. But the magnitude 0.77 is small and not significantly different from 0. As suggested earlier, the LSE may be inconsistent because of a possible correlation between spanking at ages 6–7 years and the error term that affects BPI-scores at ages 6–7 years. This endogeneity problem is addressed in the next two columns.

The second column contains the WLS results where the control variables are identical as in last lag. Between the two methods for WLS, we used the latter not using kernels, as there are too many covariates to control. The WLS effect of spanking at ages 4–5 and 6–7 years is

$$-13.91 + 11.69d_1 - 2.17d_1^2 + 12.27 - 9.70d_2 + 2.04d_2^2$$

and the total effect of spanking once a week at ages 4–5 years ($d_1 = 1; d_2 = 0$) is -4.38 (about 30% of 1 SD) at ages 8–9 years, compared with no spanking at all at both ages, which is also significant at the 10% level. The effect of spanking at age 6–7 years is again positive and significant, which makes the aggregate effect 0.23, again insignificantly different from 0. The WLS results are very similar to the last lag results, suggesting that the endogeneity problem, if any, is greatly mitigated by controlling a comprehensive set of variables.

In the third column 'LLIV', spanking at ages 6–7 years is instrumented by some key home inputs at ages 4–5 years. The 'F-statistic' for the instrument validity is 10 in the first-stage regression, which means that at least the IVs affect spanking at ages 6–7 years; the 'inclusion restriction' for the IVs holds. As for the exclusion restriction, since BPI at ages 4–5 years and the home inputs at ages 6–7 years are already controlled, it is unlikely that the IVs affect the BPI-scores at ages 6–7 years directly. As for the IVs being unrelated to the error terms of

Table 3. Spanking on child behaviour problems index BPI: baseline results†

	BPI at age 8–9 years			*BPI at 8–9 years, two stage*	*BPI at 6–7 years, two stage*
	Last lag	*WLS*	*LLIV*		
Spanked at age 4–5 years	−12.80‡ (4.92)	−13.91‡ (4.97)	−24.27 (16.09)	−5.66 (4.14)	−15.13‡ (4.66)
Spanking frequency at age 4–5 years	10.56‡ (4.87)	11.69‡ (4.94)	20.96 (20.24)	4.27 (4.27)	14.05‡ (4.62)
Spanking frequency at 4–5 years squared	−1.91 (1.04)	−2.17‡ (1.05)	−3.79 (3.82)	−0.71 (0.93)	−2.72‡ (0.99)
Spanked at age 6–7 years	11.39 (5.85)	12.27‡ (5.95)	58.45 (55.30)	13.48‡ (5.26)	
Spanking frequency at age 6–7 years	−8.10 (6.53)	−9.70 (6.63)	−55.61 (71.38)	−13.68‡ (5.96)	
Spanking frequency at 6–7 years squared	1.64 (1.44)	2.04 (1.46)	10.01 (15.37)	2.72‡ (1.32)	
BPI at age 4–5 years	0.47‡ (0.06)	0.45‡ (0.06)	0.43‡ (0.22)		0.55‡ (0.06)
BPI at age 6–7 years				0.56‡ (0.02)	
Home inputs at 8–9 years	Yes	Yes	Yes	Yes	
Home inputs at 6–7 years	Yes	Yes	Yes	Yes	Yes
Earlier home environments and child temperament	Yes	Yes	Yes	Yes	Yes
Family backgrounds	Yes	Yes	Yes	Yes	Yes
Sample size	292	289	213	289	289
R^2	0.52	0.51	0.45	0.60	0.43
	BPI-scores at age 8–9 years§				
Effects of spanked once a week versus not spanked					
Total effect of spanked at age 4–5 years	−4.16 (2.17)	−4.38 (2.29)	−7.10 (3.93)	−4.37 (2.27)	
Direct effect				−2.09 (1.96)	
Indirect effect through BPI at 6–7 years				−2.27 (1.21)	
Direct effect of spanked at age 6–7 years	4.93‡ (1.98)	4.61‡ (1.96)	12.85 (23.19)	2.52 (1.81)	
Total effect of spanked at 4–5 and 6–7 years	0.77 (2.55)	0.23 (2.70)	5.74 (22.86)	−1.84 (2.53)	

†Standard deviations are in parentheses.
‡$p < 0.05$
§Bootstrapped standard deviations are in parentheses.

the BPI-equations at ages 6–7 and 8–9 years, it is difficult to make any convincing statement unless we have some idea of what those error terms are; one can invoke the randomization argument with experimental data, but this is impossible for observational data such as ours. For the IVE, the spanking variables are still jointly significant, especially those at ages 4–5 years. The total effect of spanking once a week at ages 4–5 years is −7.10 (48% of 1 SD) at ages 8–9 years, which is again significant at the 10% level, whereas the aggregate effect is still positive but insignificant. The overall LLIV results are similar to the LSE results, and this again suggests that the endogeneity problem is minor once a large set of relevant controls are used.

The TSE in the last column 'two stage' provides the indirect effect of spanking at ages 4–5 years operating through BPI-scores at ages 6–7 years; recall that the TSE estimates the two SFs (BPI-scores at ages 8–9 and 6–7 years in expression (2)) separately. We obtain

$$\text{total effect of } d_1 = -14.13 + 12.14d_1 - 2.23d_1^2$$
$$= \underbrace{-5.66 + 4.27d_1 - 0.71d_1^2}_{\text{direct effect}} + \underbrace{0.56(-15.13 + 14.05d_1 - 2.72d_1^2)}_{\text{indirect effect}},$$
$$\text{total effect of } d_2 = 13.48 - 13.68d_2 + 2.72d_2^2,$$

where the coefficients in the total effects at both ages 4–5 and 6–7 years are very similar to those in column 'WLS'. The total effect of spanking at ages 4–5 years on BPI-scores at ages 8–9 years is negative and significant as before with a similar magnitude (−4.37), where the indirect effect −2.27 is significant and larger than the direct effect −2.09 in absolute magnitude. The effect of spanking at ages 6–7 years is still positive, but the aggregate effect is now negative, −1.84; both are insignificant. The TSE suggests that spanking at earlier ages not only reduces bad behaviour in the short run, which in turn leads to further improvement in future behaviour (indirect effect), but also directly mitigates behavioural problems in the long run.

In summary, the estimation *results are highly coherent across the different estimation methods and specifications* in Table 3, all of which suggest that spanking at ages 4–5 years reduces BPI-scores at ages 8–9 years, whereas the opposite is true for spanking at ages 6–7 years, and the total effect is insignificant. These results can be interpreted as a cautious message on using spanking to deal with children's behaviour problems: mild spanking on children below age 5 years tends to reduce poor behaviours, but spanking older children does not seem to be effective.

4.3.2. Spanking effects under anticipation

Table 4 presents the TSE results under forward looking in Section 3. The sets of control variables in the first two columns are exactly the same as those in the TSE for Table 3, except for a future spanking variable at ages 8–9 and 6–7 years respectively, in the first and second columns labelled '(1)'. The future spanking variables to assess the anticipation effects are instrumented by earlier cognitive test scores; the first-stage F-statistics for instrument validity are 5.5 and 22.9 respectively at ages 8–9 and 6–7 years. The rationale for the instruments is as follows. Given that the current BPI-score is already controlled, the mathematics and reading scores 2 years earlier are likely to be independent of the future BPI-scores (exclusion restriction), and the child's cognitive abilities measured by the test scores are likely to affect the way that the children form their anticipation about future spanking on the basis of today's experiences (inclusion restriction). Regarding the IVs being unrelated to the error terms of the BPI-equations at ages 6–7 and 8–9 years, as in the above IVE without anticipation, it is difficult to make a convincing statement as we do not have a good idea of what those error terms are.

The effects of anticipation at both ages 6–7 and 8–9 years are positive but insignificant in columns (1); their magnitudes are similar to the direct effect of spanking at ages 6–7 years. In the next two columns labelled '(2)', a dummy variable indicating the lack of earlier child temperament scores is used for those observations with missing earlier child temperament scores, which increases the sample size, particularly for the BPI-equation at ages 6–7 years. In column (2), the basic pattern is still similar, although the anticipation effect of spanking at ages 8–9 years is now insignificantly negative. It is certainly possible that these results occurred because the instruments are not satisfactory in some aspect.

In summary, somewhat disappointingly, the anticipation effects in Table 4 are insignificant, which might have been due to inadequate instruments. Viewed then just as a sensitivity analysis, since the other estimates are similar to those in the last two columns of Table 3, we may say that our baseline results are robust to the inclusion of expected future spanking.

Table 4. Spanking on child behaviour problems index BPI: future expectation†

	Results for two stage with expectation			
	(1)		*(2)*	
	BPI at 8–9 years	*BPI at 6–7 years*	*BPI at 8–9 years*	*BPI at 6–7 years*
Spanked at age 4–5 years	−6.40 (4.48)	−11.22 (26.32)	−6.89 (4.86)	−5.59 (5.24)
Spanking frequency at age 4–5 years	4.97 (4.69)	8.07 (29.45)	8.59 (5.04)	0.25 (5.43)
Spanking frequency at 4–5 years squared	−0.97 (1.07)	−1.65 (4.12)	−1.66 (1.26)	−0.37 (1.13)
Spanked at age 6–7 years	13.51‡ (5.65)		12.13 (6.33)	
Spanking frequency at age 6–7 years	−13.38‡ (5.99)	3.60 (52.31)	−10.50 (6.94)	10.75 (15.72)
Spanking frequency at 6–7 years squared	−2.50‡ (1.27)		−2.13 (1.38)	
BPI at age 4–5 years		0.50 (0.49)		0.49‡ (0.12)
BPI at age 6–7 years	62‡ (0.06)		55‡ (0.06)	
Spanking frequency at 8–9 years	2.06 (1.56)		−0.97 (6.85)	
Home inputs at 8–9 years	Yes		Yes	
Home inputs at 6–7 years	Yes	Yes	Yes	Yes
Earlier home environments	Yes	Yes	Yes	Yes
Family backgrounds	Yes	Yes	Yes	Yes
Earlier child temperament	Yes	Yes		Missing dummy
Sample size	281	229	293	293
R^2	0.60	0.50	0.53	0.36
			BPI-scores at age 8–9 years§	
			(1)	*(2)*
Effects of spanked once a week versus not spanked				
Total effect of spanked at age 4–5 years			−5.36 (9.99)	−3.13 (7.78)
Direct effect			−2.40 (2.88)	0.04 (2.71)
Indirect effect through BPI at 6–7 years			−2.96 (9.81)	−3.17 (7.37)
Total effect of spanked at age 6–7 years			4.84 (102.11)	9.72 (49.08)
Direct effect			2.63 (2.80)	3.75 (2.47)
Indirect effect through expectation			2.22 (101.83)	5.97 (48.79)
Total effect of spanked at 4–5 and 6–7 years			−0.51 (93.77)	6.58 (42.99)

†Standard deviations are in parentheses.
‡$p < 0.05$.
§Boostrapped standard deviations are in parentheses.

4.3.3. Robustness check for missing variables

To satisfy the NUC assumptions at ages 6–7 and 8–9 years, a comprehensive set of variables should be controlled. This leads to a smaller sample size even though each individual variable contains only a few missing entries; the situation of missing values can be seen from the summary statistics in Table 1. An exception is the two child temperament scores measured at age 2–3 years, which are missing for many children because the children's surveys started from 1986, implying that the children older than 3 years in 1986 do not have the scores. Also, mothers of these children are more likely to be younger owing to the age distribution in the National Longitudinal Survey of Youth. As shown in Table 2, children with younger mothers tend to be spanked more and have more behavioural problems. In this sense, the missing entries may not be 'random' and it is not clear how they should be filled in.

To check how the missing data may affect our results, two approaches are used, with the results in Table 5. The first approach is to exclude variables containing missing data, specifically, detailed home inputs at ages 8–9 and 6–7 years as well as the two child temperament scores. As shown in the second, third and fourth columns in Table 5, the results are qualitatively similar to those in Table 3, although the magnitudes of the spanking effects at ages 4–5 and 6–7 years are smaller. Varying the set of control variables in other ways and using other specifications also yield similar results. The second approach is to use the multiple-imputation method (Rubin, 1987). We used family background variables and current home inputs as predictors for missing entries (50 imputations), and the results are in the last two columns of Table 5. Note that the imputed sample size shrinks when the full set of control variables is included, but it is still much larger than that in Table 3. The regression results are again similar to those before, where the total effect of spanking at ages 4–5 years in the last lag model (column 'LLMI') is −0.61 (SD 1.30), and that at ages 6–7 years is 3.23 (SD 1.23). Despite the heavy missing data problem, we could not present these multiple-imputation results as our main results, because it was not clear how to obtain the standard errors for multistage estimates such as ours under multiple imputation; the standard errors in Table 5 under multiple imputation are *ad hoc*, being based on bootstrapping.

4.3.4. Spanking effects with three periods

As early spanking seems to be more effective than later spanking, one may wonder what the effects are of spanking even earlier than ages 4–5 years. Also, it is of interest to see how taking the earlier spanking into consideration affects our estimation results. These questions are addressed in Table 6, which is for three-period extension including ages 2–3 years spanking variables. The three-period extension of the estimators is discussed in Appendix A.

The first column 'LL1' in Table 6 controls the same set of variables as in the column last lag in Table 3 except for the additional spanking variables at ages 2–3 years, whereas the second column 'LL2' includes, in addition, detailed home inputs at ages 2–3 and 4–5 years. In both columns, the estimated total spanking effects at ages 4–5 years are again negative, and those at ages 6–7 years positive, which are the same as in Table 3. The spanking effect at ages 2–3 years is also negative, and its magnitude becomes larger when earlier home inputs are further controlled in column LL2.

The same two sets of controls in the first two columns are also used in the columns 'WLS1' and 'WLS2'. As before, including earlier home inputs greatly increases the explanatory power of the regression (from 0.46 to 0.68), although the estimates become less precise owing to the smaller sample size. The overall WLS results are similar to the last lag model. On the basis of WLS2 the total effect of spanking $d = (d_0, d_1, d_2)$ at ages 2–3, 4–5 and 6–7 years is

$$-10.34 + 2.62d_0 - 0.20d_0^2 - 27.19 + 29.26d_1 - 6.04d_1^2 + 41.90 - 42.51d_2 + 8.35d_2^2.$$

The estimates in the same column in the lower panel suggest that the total effect of spanking once a week at ages 2–3, 4–5 and 6–7 years ($d_0 = 1; d_1 = 1; d_2 = 1$) is −4.16 (28% of 1 SD) compared with no spanking at all, whereas the effects of spanking at each of the three periods are −7.92, −3.97 and 7.73 respectively. These results suggest that earlier spanking seems to be more effective in reducing a child's behavioural problems. The next three columns show the results of the three-stage estimator, which is the extension of the TSE. The effects of spanking once a week at ages 2–3, 4–5 and 6–7 are −9.95, −5.53 and 3.16 respectively, which lead to the aggregate effect of spanking at all three periods of −12.26 (83% of 1 SD).

In summary, a common pattern in Table 6 is negative effects of spanking at ages 2–3 and 4–5 years and positive effects of spanking at ages 6–7 years on BPI-scores at ages 8–9 years. Spanking

Table 5. Spanking on child behaviour problems index BPI: robustness check for missing variables†

	BPI at 8–9 years				*BPI at 8–9 years, two-stage multiple imputation*	*BPI at 6–7 years, two-stage multiple imputation*
	Last lag	*WLS*	*LLIV*	*LLMI*		
Spanked at age 4–5 years	−1.63 (2.16)	−1.34 (2.19)	−0.85 (4.54)	−2.57 (3.30)	−0.74 (2.82)	−2.68 (2.48)
Spanking frequency at age 4–5 years	1.77 (2.23)	1.46 (2.27)	−4.19 (7.79)	2.16 (3.29)	0.74 (2.97)	2.59 (2.52)
Spanking frequency at 4–5 years squared	−0.20 (0.49)	−0.13 (0.0)	0.64 (1.55)	−0.20 (0.74)	0.11 (0.68)	−0.52 (0.56)
Spanked at age 6–7 years	1.81 (2.80)	3.49 (2.81)	−15.22 (105.03)	5.41 (5.04)	7.11 (4.09)	
Spanking frequency at age 6–7 years	0.07 (3.13)	−2.01 (3.11)	63.04 (118.43)	−2.93 (5.75)	−6.65 (4.73)	
Spanking frequency at 6–7 years squared	0.24 (0.70)	0.73 (0.68)	−16.75 (27.20)	0.75 (1.30)	1.27 (1.06)	
BPI at age 4–5 years	0.57‡ (0.03)	0.56‡ (0.03)	0.50‡ (0.08)	0.50‡ (0.04)		0.53‡ (0.03)
BPI at age 6–7 years					0.61‡ (0.04)	
Home inputs at 8–9 years				Yes	Yes	
Home inputs at 6–7 years				Yes	Yes	Yes
Early child temperament				Yes	Yes	Yes
Earlier home environments	Yes	Yes	Yes	Yes	Yes	Yes
Family backgrounds	Yes	Yes	Yes	Yes	Yes	Yes
Sample size	1503	1446	1485	651	648	711
R^2	0.397	0.40	—	—	—	—
	BPI-scores at 8–9 years§					
Effects of spanked once a week versus not spanked						
Total effect of spanked at age 4–5 years	−0.06 (0.75)	−0.01 (0.86)	−4.40 (7.62)	−0.61 (1.30)	−0.27	
Direct effect					0.11 (1.08)	−0.61 (1.08)
Indirect effect through BPI at 6–7 years					−0.37	
Direct effect of spanked at age 6–7 years	2.13‡ (0.82)	2.21‡ (0.94)	31.07 (45.76)	3.23‡ (1.23)	1.73 (1.60)	
Total effect of spanked at 4–5 and 6–7 years	2.07 (1.07)	2.20 (1.23)	26.67 (40.37)	2.62 (1.60)	1.46	

†Standard deviations are in parentheses.
‡$p < 0.05$.
§Boostrapped standard deviations are in parentheses.

Table 6. Spanking on child behaviour problems index BPI: earlier periods spanking†

	BPI-scores at 8–9 years				*BPI-scores at 8–9 years, three stages*	*BPI-scores at 6–7 years, three stages*	*BPI-scores at 4–5 years, three stages*
	Last lag with three periods		*WLS with three periods*				
	LL1	*LL2*	*WLS1*	*WLS2*			
Spanked at age 4–5 years	−17.05‡ (7.92)	−22.00‡ (10.40)	−19.87‡ (8.18)	−27.19‡ (11.28)	−20.32‡ (9.18)	−17.85‡ (6.30)	
Spanking frequency at 4–5 years	16.01‡ (7.74)	23.86‡ (10.18)	18.56‡ (7.97)	29.26‡ (10.66)	21.12‡ (9.16)	17.80‡ (6.07)	
Spanking frequency at 4–5 years squared	−3.23 (1.72)	−4.68‡ (2.23)	−3.78‡ (1.74)	−6.04‡ (2.31)	−4.34‡ (1.98)	−3.85‡ (1.33)	
Spanked at age 6–7 years	21.10‡ (8.75)	30.87‡ (11.90)	26.58‡ (9.43)	41.90‡ (11.75)	23.63‡ (9.78)		
Spanking frequency at 6–7 years	−20.12‡ (10.08)	−29.22‡ (12.96)	−26.48‡ (10.87)	−42.51‡ (12.66)	−25.11‡ (10.77)		
Spanking frequency at 6–7 years squared	3.98 (2.17)	5.29 (2.83)	5.47‡ (2.28)	8.35‡ (2.74)	4.64‡ (2.33)		
Spanked at age 2–3 years	−1.83 (9.66)	−9.14 (11.11)	−2.46 (9.75)	−10.34 (10.14)	−9.42 (8.71)	8.10 (6.49)	−16.47 (12.05)
Spanking frequency at 2–3 years	1.49 (5.82)	−0.69 (9.34)	0.03 (5.87)	2.62 (9.12)	−1.58 (7.13)	2.95 (4.79)	1.58 (6.74)
Spanking frequency at 2–3 years squared	−0.32 (1.21)	0.44 (1.89)	−0.00 (1.22)	−0.20 (1.85)	0.58 (1.42)	−0.14 (1.01)	−0.38 (1.30)
BPI score at age 6–7 years					0.51‡ (0.09)		
BPI score at age 4–5 years						0.65‡ (0.08)	
Home inputs at 8–9 years	Yes	Yes	Yes	Yes	Yes		
Home inputs at 6–7 years	Yes	Yes	Yes	Yes	Yes	Yes	
Home inputs at 4–5 years		Yes		Yes	Yes	Yes	Yes
Home inputs at 2–3 years		Yes		Yes	Yes	Yes	Yes
Earlier home environment and temperament	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Family backgrounds	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample size	192	165	183	159	163	159	159
R^2	0.45	0.66	0.46	0.68	0.77	0.70	0.36

	BPI-scores at age 8–9 years§				
	LL1	*LL2*	*WLS1*	*WLS2*	*Three stage*
Effects of spanked once a week versus not spanked					
Total effect of spanked at age 4–5 years	−4.27 (3.71)	−2.82 (9.61)	−5.09 (3.95)	−3.97 (11.24)	−5.53 (16.11)
Direct effect of spanked at age 6–7 years	4.96 (3.08)	6.95 (8.18)	5.57 (3.23)	7.73 (10.01)	3.16 (18.18)
Total effect of spanked at age 2–3 years	−0.66 (12.32)	−9.39 (20.99)	−2.44 (11.12)	−7.92 (27.59)	−9.95 (25.44)
Total effect of spanked at 2–7 years	0.03 (13.01)	−5.27 (24.53)	−1.95 (11.74)	−4.16 (30.26)	−12.16 (37.92)

†Standard deviations are in parentheses.
‡$p < 0.05$.
§Bootstrapped standard deviations are in parentheses.

at the very young ages (2–3 years) has the largest effect in mitigating behaviour problems at ages 8–9 years, with the magnitude about twice that of spanking at ages 4–5 and 6–7 years. So the overall evidence suggests that spanking at younger ages (2–5 years old) can be more effective than spanking at later ages.

4.4. Related literature and summary of empirical analysis

Many parent would have wondered: does spanking work, in the sense of reducing a child's behaviour problems? The causal link between spanking and behaviour problems is complicated, and still hotly debated after many years of investigation in disciplines such as education and psychology (Gershoff, 2002; Larzelere, 2000; Benjeta and Kazdin, 2003). Although the question may sound irrelevant to economics, it is not, because early child development results are important determinants of later schooling and socio-economic success (Keane and Wolpin, 1997; Heckman, 2000; Heckman and Rubinstein, 2001; Heckman *et al.*, 2006; Currie and Stabile, 2006). McLeod and Kaiser (2004), for example, showed that children who have behaviour problems at early ages (6–8 years) are less likely to graduate from high school or to attend college.

There are various difficulties in establishing the causal link between spanking and behaviour. One is that the definition of spanking (or physical punishment in general) is delicate. For example, spanking with an open hand with composure can be quite different from spanking with visible anger (Benjeta and Kazdin, 2003). Second, spanking may be endogenous for various reasons; for example, low income may induce poor child behaviour and more spanking, which may result in a spurious positive relationship between spanking and child behaviour. Third, the relationship is likely to be dynamic with complicated feedback; for example, an initial spanking D_1 affects an interim child behaviour Y_1, which in turn affects both future spanking D_2 and future behaviour Y_2 and so on. Among these difficulties, this paper focused on the last (dynamic aspect), which seems to have been completely neglected in the literature, and went one step further by incorporating anticipation effects. This paper also took as much care in controlling for the endogeneity problem as any non-randomized data could.

A coherent pattern across different samples and specifications is that spanking once a week at ages 2–3 and 4–5 years is effective in reducing child behaviour problems, whereas spanking older children is not. In the baseline results of Table 3, the effect of spanking once a week at ages 4–5 years ranges from 30% to 50% of 1 SD reduction in BPI-scores at ages 8–9 years, whereas the corresponding estimates at ages 2–3 years range from 55% to 70% under the more comprehensive specification in Table 6. Also, the spanking effect is dynamic, as a substantial part of the spanking effect at ages 4–5 years on BPI-scores at ages 8–9 years is indirectly transmitted through the intermediate BPI-scores at ages 6–7 years. A similar dynamic pattern also holds for spanking at ages 2–3 years.

These results, especially the dynamic effects, are new to the spanking literature. Broadly speaking, they are in line with the general findings of more causally relevant studies where predominantly beneficial outcomes of physical punishment are found for children under 6 years old whereas detrimental effects are found for older children (Larzelere, 2000). The dynamic nature of spanking effects is also consistent with the control system theory in psychology (Patterson, 1988; Granic and Patterson, 2006), which emphasizes the recursive, bidirectional nature between child behaviour and parental reactions. Somewhat disappointingly, we did not find significant evidence for anticipation (i.e. deterrence) effects, which is possibly because of weak forward looking skills of children as well as because of inadequate instruments.

A shortcoming of our empirical analysis is that our data are observational, but this seems unavoidable as it would be impossible to do experiments on spanking. Another shortcoming is

that mothers report both spanking and outcome variables, which may cause underestimation of spanking and overestimation of bad behaviour for those 'spankers'. This problem may be overcome by having the children answer on spanking as well for a double check (or the fathers on the child behaviour), although this was not done in our data. Yet another shortcoming is that only the frequency of spanking is asked, whereas the severity and specific method of spanking are not. Given that there are so many elements possibly affecting a child's behaviour problems (Granic and Patterson, 2006), even after controlling for a very comprehensive set of covariates, one can never be sure of covering all omitted variables; the problem, however, seems fairly minor because our estimates are similar across different methods that vary in ability to deal with such problems. Further corroborating evidence is that the explanatory powers of the regressions are reasonably large.

5. Conclusions

Treatment effect analysis is possibly the most important area in science, because an important reason to do science is to change a cause to improve the outcome. Much of the literature is concerned with 'one-shot' static treatment. But many treatments are repeated over time in reality and, when this is done, they are often modified depending on the interim outcomes. This is natural, as people always try to do better given the information at hand that accumulates over time, and the interim outcomes are part of the information.

When some treatments are affected by interim outcomes that also affect the final outcome of interest, finding the total effect of the treatment profile is complicated. An approach to handle this situation was devised a long time ago in the form of the *G*-algorithm. But this non-parametric approach is difficult to implement in practice, and, several variants have appeared. A popular one is the MSM approach, which applies WLS to the RF equation of the final response, and the total effect can be found directly from the RF. But there are other alternatives as well such as applying an IVE to the RF, or the LSE to the SFs of the final and interim responses; the latter can decompose the total effect into the direct and indirect effects. These practical approaches and the decomposition of the total effect rely heavily on linear model specifications.

This paper reviewed the literature and made an extension to an important direction: anticipation effects. Differently from a simple dose–response relationship where there is no scope for future expected dosages having any effect on the current response, forward looking matters greatly when human beings are involved as they act on anticipation. Causal analysis under anticipation is a complicated endeavour, because one party's anticipation may depend on what the other party anticipates, which in turn depends on what the first party anticipates, *ad infinitum*. It will take a while before a coherent dynamic causal framework can be built up allowing for anticipation effects. Nevertheless, at a minimum, our proposal of using expected values to find anticipation effects may be taken as a sensitivity analysis in that direction.

After presenting the analytic framework on how to extend dynamic treatment analyses to incorporate anticipation, we then applied the method to children's bad behaviour and spanking. Using National Longitudinal Survey of Youth data, we found that early spanking (at ages 4–5 years) is more effective than later spanking (at ages 6–7 years) in reducing future bad behaviour (at ages 8–9 years). However, no evidence has been found that anticipation matters in this case.

Admittedly, our empirical analysis of spanking effects on children's behaviour is unsatisfactory in two regards. Firstly, the treatment effect methods are geared up for randomized treatments, or at least randomizable treatments. Although spanking can be randomized for poorly behaving children, it cannot be randomized for perfectly well behaving children; an assumption

is thus needed that spanking is mild and any child can be spanked as no child behaves perfectly. Secondly, spanking is a delicate and complicated matter with no experiment possible, and some assumptions put on our data are not easy to justify. Nevertheless, the issues that we addressed—dynamic spanking effects when there are feedback and lagged response effects along with anticipation effects—are important and have never been addressed in the psychology and education literature as far as we know. We doubt whether there will ever be completely satisfactory data or answers to the question of a spanking effect. At least we showed what it would take, in terms of methods and assumptions, to address the question properly.

Acknowledgements

The authors are grateful to the Joint Editor, Associate Editor and two reviewers for providing detailed comments and directing the authors' attention to relevant references in the literature.

Myoung-jae Lee's work was supported by a National Research Foundation of Korea grant funded by the Korean Government (NRF-2010-330-B00060).

Appendix A

A.1. Effects under anticipation

Recall expression (4) and substitute the $Y_1^{d_1}$ SF into the $D_2^{d_1}$-equation to obtain

$$\begin{aligned} D_2^{d_1} &= \zeta_1 + X_2'\zeta_x + \zeta_y\{\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1 + \alpha_{de}\, E(D_2^{d_1}|I_1) + U_1\} + \varepsilon_2 \\ &= \zeta_1 + X_2'\zeta_x + \zeta_y(\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1) + \zeta_y\alpha_{de}\, E(D_2^{d_1}|I_1) + \zeta_y U_1 + \varepsilon_2. \end{aligned}$$

Take $E(\cdot|I_1)$ on this equation to obtain, under $E(\zeta_y U_1 + \varepsilon_2|I_1) = 0$,

$$E(D_2^{d_1}|I_1) = \zeta_1 + E(X_2'|I_1)\zeta_x + \zeta_y(\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1) + \zeta_y\alpha_{de}\, E(D_2^{d_1}|I_1).$$

Assuming $\zeta_y\alpha_{de} \neq 1$, solve this for $E(D_2^{d_1}|I_1)$:

$$E(D_2^{d_1}|I_1) = \frac{1}{1-\zeta_y\alpha_{de}}\{\zeta_1 + E(X_2'|I_1)\zeta_x + \zeta_y(\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1)\}. \tag{6}$$

Substitute this back into the $Y_1^{d_1}$ SF to obtain

$$Y_1^{d_1} = \alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1 + \frac{\alpha_{de}}{1-\zeta_y\alpha_{de}}\{\zeta_1 + E(X_2'|I_1)\zeta_x + \zeta_y(\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1)\} + U_1.$$

Note, as it will be needed below, that

$$Y_1^{d_1} - Y_1^0 = \left(\alpha_d + \frac{\alpha_{de}\zeta_y\alpha_d}{1-\zeta_y\alpha_{de}}\right) d_1. \tag{7}$$

Turning to $E(D_3^{d_1 d_2}|I_2)$, to remove $E(D_3^{d_1 d_2}|I_2)$ in the $Y_2^{d_1 d_2}$ SF, we need a model for $D_3^{d_1 d_2}$. Suppose that

$$D_3^{d_1 d_2} = \delta_1 + X_3'\delta_x + \delta_y Y_2^{d_1 d_2} + \varepsilon_3$$

which is analogous to the $D_2^{d_1}$-equation. Substitute the $Y_2^{d_1 d_2}$ SF to obtain

$$\begin{aligned} D_3^{d_1 d_2} &= \delta_1 + X_3'\delta_x + \delta_y\{\beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_{de}\, E(D_3^{d_1 d_2}|I_2) + \beta_y Y_1^{d_1} + U_2\} + \varepsilon_3 \\ &= \delta_1 + X_3'\delta_x + \delta_y(\beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2) + \delta_y\beta_{de}\, E(D_3^{d_1 d_2}|I_2) + \delta_y\beta_y Y_1^{d_1} + \delta_y U_2 + \varepsilon_3. \end{aligned}$$

Take $E(\cdot|I_2)$ on this and solve for $E(D_3^{d_1 d_2}|I_2)$ under $E(\delta_y U_2 + \varepsilon_3|I_2) = 0$ and $\delta_y\beta_{de} \neq 1$:

$$E(D_3^{d_1 d_2}|I_2) = \frac{1}{1-\delta_y\beta_{de}}\{\delta_1 + X_3'\delta_x + \delta_y(\beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2) + \delta_y\beta_y Y_1^{d_1}\}. \tag{8}$$

Observe that the $Y_2^{d_1 d2}$ RF with $Y_1^{d_1}$ removed is

$$
\begin{aligned}
Y_2^{d_1d_2} &= \beta_1 + X_2'\beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_{de}\, E(D_3^{d_1d_2}|I_2) \\
&\quad + \beta_y\{\alpha_1 + \bar{X}_1'\alpha_x + \alpha_d d_1 + \alpha_{de}\, E(D_2^{d_1}|I_1) + U_1\} + U_2 \\
&= \beta_1 + \beta_y\alpha_1 + \bar{X}_1'\alpha_x\beta_y + X_2'\beta_x + (\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2 \\
&\quad + \beta_y\alpha_{de}\, E(D_2^{d_1}|I_1) + \beta_{de}\, E(D_3^{d_1d_2}|I_2) + \beta_y U_1 + U_2.
\end{aligned}
$$

From this,

$$
\begin{aligned}
E(Y_2^{d_1d_2} - Y_2^{00}) &= (\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2 + \beta_y\alpha_{de}\, E\{E(D_2^{d_1}|I_1) - E(D_2^0|I_1)\} \\
&\quad + \beta_{de}\, E\{E(D_3^{d_1d_2}|I_2) - E(D_3^{00}|I_2)\}.
\end{aligned}
$$

From equations (6)–(8),

$$
\begin{aligned}
E(D_2^{d_1}|I_1) - E(D_2^0|I_1) &= \frac{\zeta_y\alpha_d}{1-\zeta_y\alpha_{de}} d_1, \\
E(D_3^{d_1d_2}|I_2) - E(D_3^{00}|I_2) &= \frac{1}{1-\delta_y\beta_{de}}(\beta_{d\text{lag}} d_1 + \beta_d d_2) + \frac{\delta_y\beta_y}{1-\delta_y\beta_{de}}(Y_1^{d_1} - Y_1^0) \\
&= \frac{1}{1-\delta_y\beta_{de}}(\beta_{d\text{lag}} d_1 + \beta_d d_2) + \frac{\delta_y\beta_y}{1-\delta_y\beta_{de}}\left(\alpha_d + \frac{\alpha_{de}\zeta_y\alpha_d}{1-\zeta_y\alpha_{de}}\right) d_1.
\end{aligned}
$$

Therefore, the *total effect under anticipation* is (set $\alpha_{de} = \beta_{de} = 0$ for no anticipation)

$$
\begin{aligned}
E(Y_2^{d_1d_2} - Y_2^{00}) &= (\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2 + \frac{\beta_y\alpha_{de}\zeta_y\alpha_d}{1-\zeta_y\alpha_{de}} d_1 + \frac{\beta_{de}}{1-\delta_y\beta_{de}}(\beta_{d\text{lag}} d_1 + \beta_d d_2) \\
&\quad + \frac{\beta_{de}\delta_y\beta_y}{1-\delta_y\beta_{de}}\left(\alpha_d + \frac{\alpha_{de}\zeta_y\alpha_d}{1-\zeta_y\alpha_{de}}\right) d_1. \qquad (9)
\end{aligned}
$$

A.2. Simulation demonstration for linear models (1) and (4)

Recall expression (1) and consider (no X_0 and Y_0 for simplification) an IV-augmented version:

$$
\begin{aligned}
D_1 &= \xi_1 + \xi_z Z_1 + \xi_x X_1 + \varepsilon_1 \qquad & Y_1 &= \alpha_1 + \alpha_x X_1 + \alpha_d D_1 + U_1, \\
D_2 &= \zeta_1 + \zeta_z Z_2 + \zeta_x X_2 + \zeta_y Y_1 + \varepsilon_2, \qquad & Y_2 &= \beta_1 + \beta_x X_2 + \beta_{d\text{lag}} D_1 + \beta_d D_2 + \beta_y Y_1 + U_2;
\end{aligned}
$$

all ξ, α, ζ and β parameters are 1, $N = 500$ and all regressors and errors as well as the IVs Z_1 and Z_2 are independent identically distributed $N(0, 1)$. The total effect for (1, 1) is $(\beta_{d\text{lag}} + \beta_y\alpha_d)d_1 + \beta_d d_2 = 3$.

Table 7 shows the Monte Carlo result estimating the total effect 2000 times. For each estimator, the average and SD are shown from the 2000 estimates. Whereas the TSE from Huang and Lee (2010) and the IVE for the Y_2 RF need no explanation, the WLSs do. WLS1 'conditionally' averages $f(D_2|D_1, Y_1, \bar{X}_2)$ to obtain $f(D_2|D_1, \bar{X}_2)$ by using the rule-of-thumb bandwidth, but WLS2 uses a linear model for D_2 on $(D_1, \bar{X}_2)$ as explained in the main text. The numbers 0.0001 and 0.001 are trimming constants to remove observations with too small $f(D_2|D_1, Y_1, \bar{X}_2)$. Table 7 demonstrates that the estimators work as they are supposed to, but the SDs of the TSE and IVE are about half the SDs of the WLSs. WLS2 performs better than WLS1 although WLS1 is more theoretically sound. The larger trimming constant 0.001 gives slightly better results than the smaller one, 0.0001.

Table 7. Monte Carlo results for the no-anticipation linear model

	TSE	*IVE*	*WLS1 (trimming constant 0.0001)*	*WLS2 (trimming constant 0.0001)*	*WLS1 (trimming constant 0.001)*	*WLS2 (trimming constant 0.001)*
Average	3.000	3.000	3.001	3.001	3.001	3.000
SD	0.043	0.045	0.096	0.072	0.091	0.069

Table 8. Monte Carlo results for the anticipation linear model

	α_x	α_d	α_{de}	β_x	$\beta_{d\text{lag}}$	β_d	β_{de}	β_y
Average	1.001	1.000	0.500	1.001	1.000	1.000	0.500	1.001
SD	0.050	0.038	0.016	0.051	0.057	0.040	0.016	0.045

Recall expression (4) and consider an IV-augmented version: again with $N = 500$,

$$Y_1 = \alpha_1 + \alpha_x X_1 + \alpha_d D_1 + \alpha_{de} E(D_2|I_1) + U_1, \qquad D_3 = \delta_1 + \delta_z Z_3 + \delta_x X_3 + \delta_y Y_2 + \varepsilon_3,$$
$$Y_2 = \beta_1 + \beta_x X_2 + \beta_{d\text{lag}} D_1 + \beta_d D_2 + \beta_{de} E(D_3^{d_1 d_2}|I_2) + \beta_y Y_1 + U_2, \qquad \alpha_{de} = \beta_{de} = 0.5,$$

where the D_1- and D_2-equations have been omitted as they are the same as above whereas D_3 is new; the δ-parameters are all 1, and Z_3, X_3 and ε_3 are independent identically distributed $N(0, 1)$. We set $\alpha_{de} = \beta_{de} = 0.5$ for the condition $\zeta_y \alpha_{de} \neq 1$ and $\delta_y \beta_{de} \neq 1$ under which equations (6) and (8) exist; in generating the simulation data, we used $E(D_2|I_1)$ and $E(D_3^{d_1 d_2}|I_2)$ based on equations (6) and (8). Table 8 is the TSE result using the IVE for the Y_1 and Y_2 SFs. 2000 repetitions were done to compute their average and SD: the results for the intercepts α_1 and β_1 have been omitted. Table 8 shows that the TSE for expression (4) works and, if desired, one can obtain the total effect (9).

A.3. Three-period case without anticipation

Linear SFs for three periods are

$$Y_1^{d_1} = \alpha_1 + \bar{X}_1' \alpha_x + \alpha_d d_1 + U_1,$$
$$Y_2^{d_1 d_2} = \beta_1 + X_2' \beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_y Y_1^{d_1} + U_2,$$
$$Y_3^{d_1 d_2 d_3} = \gamma_1 + X_3' \gamma_x + \gamma_{d\text{lag}2} d_1 + \gamma_{d\text{lag}1} d_2 + \gamma_d d_3 + \gamma_y Y_2^{d_1 d_2} + U_3.$$

The $Y_3^{d_1 d_2 d_3}$ RF with $Y_2^{d_1 d_2}$ and $Y_1^{d_1}$ removed is

$$\begin{aligned} Y_3^{d_1 d_2 d_3} &= \gamma_1 + X_3' \gamma_x + \gamma_{d\text{lag}2} d_1 + \gamma_{d\text{lag}1} d_2 + \gamma_d d_3 + \gamma_y (\beta_1 + X_2' \beta_x + \beta_{d\text{lag}} d_1 + \beta_d d_2 + \beta_y Y_1^{d_1} + U_2) + U_3 \\ &= \gamma_1 + \gamma_y \beta_1 + X_2' \beta_x \gamma_y + X_3' \gamma_x + (\gamma_{d\text{lag}2} + \gamma_y \beta_{d\text{lag}}) d_1 + (\gamma_{d\text{lag}1} + \gamma_y \beta_d) d_2 + \gamma_d d_3 \\ &\quad + \gamma_y \beta_y Y_1^{d_1} + \gamma_y U_2 + U_3 \\ &= \gamma_1 + \gamma_y \beta_1 + \gamma_y \beta_y \alpha_1 + \bar{X}_1' \alpha_x \gamma_y \beta_y + X_2' \beta_x \gamma_y + X_3' \gamma_x + (\gamma_{d\text{lag}2} + \gamma_y \beta_{d\text{lag}} + \gamma_y \beta_y \alpha_d) d_1 \\ &\quad + (\gamma_{d\text{lag}1} + \gamma_y \beta_d) d_2 + \gamma_d d_3 + \gamma_y \beta_y U_1 + \gamma_y U_2 + U_3. \end{aligned}$$

This shows the total effect consisting of various direct and indirect effects. Each SF can be estimated by the TSE and the total effect can be constructed.

As for estimating the Y_3 RF with (d_1, d_2, d_3) replaced by (D_1, D_2, D_3), the LSE is inconsistent because D_2 is related to U_1 in the error term through Y_1 and D_3 is related to U_2 through Y_2. With IVs Z_2 and Z_3 for D_2 and D_3, IV estimation is easy. As for WLS, with $\bar{X}_3 \equiv (X_0, Y_0, X_1, X_2, X_3)$, the weight is the square root of the density product in the first expression of

$$\begin{aligned} &E\left\{ Y_3 \frac{f(D_3|D_2, D_1, \bar{X}_3)}{f(D_3|Y_2, D_2, Y_1, D_1, \bar{X}_3)} \frac{f(D_2|D_1, \bar{X}_3)}{f(D_2|Y_1, D_1, \bar{X}_3)} | \bar{X}_3 \right\} \\ &\quad = \int y_3 \frac{f(d_3|d_2, d_1, \bar{X}_3)}{f(d_3|y_2, d_2, y_1, d_1, \bar{X}_3)} \frac{f(d_2|d_1, \bar{X}_3)}{f(d_2|y_1, d_1, \bar{X}_3)} f(y_3|d_3, y_2, d_2, y_1, d_1, \bar{X}_3) \\ &\qquad \times f(d_3, y_2, d_2, y_1, d_1|\bar{X}_3)\, \mathrm{d}y_3\, \mathrm{d}d_3\, \mathrm{d}y_2\, \mathrm{d}d_2\, \mathrm{d}y_1\, \mathrm{d}d_1 \\ &\quad = \int y_3 \frac{f(d_3|d_2, d_1, \bar{X}_3)}{f(d_3|y_2, d_2, y_1, d_1, \bar{X}_3)} \frac{f(d_2|d_1, \bar{X}_3)}{f(d_2|y_1, d_1, \bar{X}_3)} f(y_3|d_3, y_2, d_2, y_1, d_1, \bar{X}_3)\, f(d_3|y_2, d_2, y_1, d_1, \bar{X}_3) \\ &\qquad \times f(y_2|d_2, y_1, d_1, \bar{X}_3)\, f(d_2|y_1, d_1, \bar{X}_3)\, f(y_1|d_1, \bar{X}_3)\, f(d_1|\bar{X}_3)\, \mathrm{d}y_3\, \mathrm{d}d_3\, \mathrm{d}y_2\, \mathrm{d}d_2\, \mathrm{d}y_1\, \mathrm{d}d_1 \end{aligned}$$

$$= \int y_3 \, f(y_3|d_3, y_2, d_2, y_1, d_1, \bar{X}_3) \, f(d_3|d_2, d_1, \bar{X}_3) \, f(y_2|d_2, y_1, d_1, \bar{X}_3) \, f(d_2|d_1, \bar{X}_3)$$
$$\times \, f(y_1|d_1, \bar{X}_3) \, f(d_1|\bar{X}_3) \, \mathrm{d}y_3 \, \mathrm{d}d_3 \, \mathrm{d}y_2 \, \mathrm{d}d_2 \, \mathrm{d}y_1 \, \mathrm{d}d_1.$$

Since the removal of the endogeneity is done by $D_3 \amalg Y_2|(D_2, Y_1, D_1, \bar{X}_3)$ and $D_2 \amalg Y_1|(D_1, \bar{X}_3)$ in the artificial population, it may be better to replace $f(D_3|D_2, D_1, \bar{X}_3)$ with $f(D_3|D_2, Y_1, D_1, \bar{X}_3)$ to obtain $f(d_3|d_2, y_1, d_1, \bar{X}_3) \, f(y_2|d_2, y_1, d_1, \bar{X}_3)$ for $D_3 \amalg Y_2|(D_2, Y_1, D_1, \bar{X}_3)$ where the two conditioning sets are the same. This is analogous to $f(d_2|d_1, \bar{X}_3) \, f(y_1|d_1, \bar{X}_3)$ for $D_2 \amalg Y_1|(D_1, \bar{X}_3)$ where the two conditioning sets are the same as well.

References

Abbring, J. H. and van den Berg, G. J. (2003) The nonparametric identification of treatment effects in duration models. *Econometrica*, **71**, 1491–1517.

Abbring, J. H. and Heckman, J. J. (2007) Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In *Handbook of Econometrics*, vol. 6B (eds J. J. Heckman and E. E. Leamer), ch. 72, pp. 5145–5303. Amsterdam: North-Holland.

Abbring, J. H. and Heckman, J. J. (2008) Dynamic policy analysis. In *The Econometrics of Panel Data* (eds L. Mátyás and P. Sevestre), ch. 24, pp. 795–863. New York: Springer.

Almirall, D., Ten Have, T. and Murphy, S. A. (2010) Structural nested mean models for assessing time-varying effect moderation. *Biometrics*, **66**, 131–139.

Baumrind, D., Larzelere, R. E. and Cowan, P. A. (2002) Ordinary physical punishment: is it harmful? *Psychol. Bull.*, **128**, 580–589.

Benjeta, C. and Kazdin, A. E. (2003) Spanking children: the controversies, findings, and new directions. *Clin. Psychol. Rev.*, **23**, 197–224.

Currie, J. and Stabile, M. (2006) Child mental health and human capital accumulation: the case of ADHD. *J. Hlth Econ.*, **25**, 1094–1118.

Gershoff, E. (2002) Corporal punishment by parents and associated child behaviors and experiences: a meta-analytic and theoretical review. *Psychol. Bull.*, **128**, 539–579.

Gill, R. and Robins, J. M. (2001) Causal inference for complex longitudinal data: the continuous case. *Ann. Statist.*, **29**, 1785–1811.

Granic, I. and Patterson, G. R. (2006) Toward a comprehensive model of antisocial development: a dynamic systems approach. *Psychol. Rev.*, **113**, 101–131.

Heckman, J. J. (2000) Policies to foster human capital. *Res. Econ.*, **54**, 3–56.

Heckman, J. J. and Navarro, S. (2007) Dynamic discrete choice and dynamic treatment effects. *J. Econmetr.*, **136**, 341–396.

Heckman, J. J. and Rubinstein, Y. (2001) The importance of noncognitive skills: lessons from the GED testing program. *Am. Econ. Rev.*, **91**, 145–149.

Heckman, J. J., Stixrud, J. and Urzua, S. (2006) The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Lab. Econ.*, **24**, 411–482.

Hérnan, M. A., Brumback, B. and Robins, J. M. (2001) Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Am. Statist. Ass.*, **96**, 440–448.

Hérnan, M. A., Hernández-Díaz, S. and Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Huang, F. and Lee, M. J. (2010) Dynamic treatment effect analysis of TV effects on child cognitive development. *J. Appl. Econmetr.*, **25**, 392–419.

Imbens, G. W. (2000) The role of the propensity score in estimating dose-response functions. *Biometrika*, **87**, 706–710.

Joffe, M. M., Ten Have, T. R., Feldman, H. I. and Kimmel, S. E. (2004) Model selection, confounder control, and marginal structural models: review and new applications. *Am. Statistn*, **58**, 272–279.

Keane, M. and Wolpin, K. (1997) Career decisions of young men. *J. Polit. Econ.*, **105**, 473–522.

Larzelere, R. E. (2000) Child outcomes of nonabusive and customary physical punishment by parents: an updated literature review. *Clin. Chld Famly Psychol. Rev.*, **3**, 199–221.

Lechner, M. (2001) Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labor Market Policies* (eds M. Lechner and F. Pfeiffer), pp. 43–58. Heidelberg: Physica.

Lechner, M. (2008) Matching estimation of dynamic treatment models: some practical issues. In *Advances in Econometrics*, vol. 21, *Modelling and Evaluating Treatment Effects in Econometrics* (eds D. Millimet, J. Smith and E. Vytlacil), pp. 289–333. Oxford: JAI.

Lechner, M. (2009) Sequential causal models for the evaluation of labor market programs. *J. Bus. Econ. Statist.*, **27**, 71–83.

Lechner, M. (2011) The relation of different concepts of causality used in time series and microeconometrics. *Econmetr. Rev.*, **30**, 109–127.
Lechner, M. and Wiehler, S. (2012) Does the order and timing of active labor market programs matter? *Oxf. Bull. Econ. Statist.*, to be published.
Lee, M. J. (2002) *Panel Data Econometrics: Methods-of-moments and Limited Dependent Variables*. New York: Academic Press.
Lee, M. J. (2005) *Micro-econometrics for Policy, Program, and Treatment Effects*. New York: Oxford University Press.
Lee, M. J. (2010) Measuring the usage effects of tying a messenger to Windows: a treatment effect approach. *J. R. Statist. Soc.* A, **173**, 237–253.
Lee, M. J. (2012) Semiparametric estimators for limited dependent variable (LDV) models with endogenous regressors. *Econmetr. Rev.*, to be published.
Li, Y. P., Propert, K. J. and Rosenbaum, P. R. (2001) Balanced risk set matching. *J. Am. Statist. Ass.*, **96**, 870–882.
McLeod, J. D. and Kaiser, K. (2004) Childhood emotional and behavioral problems and educational attainment. *Am. Sociol. Rev.*, **69**, 636–658.
Patterson, G. R. (1988) Stress: a change agent for family process. In *Stress, Coping, and Development in Children* (eds N. Garmezy and M. Rutter), pp. 235–264. Baltimore: Johns Hopkins University Press.
Pearl, J. (2009) Causal inference in statistics: an overview. *Statist. Surv.*, **3**, 96–146.
Pearl, J. (2010) An introduction to causal inference. *Int. J. Biostatist.*, **6**, article 7.
Robins, J. M. (1986) A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Modllng*, **7**, 1393–1512.
Robins, J. M. (1987) A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect—errata. *Comput. Math. Applic.*, **14**, 917–921; addendum, 923–945; errata to addendum, **18** (1989), 477.
Robins, J. M. (1998) Structural nested failure time models. In *Encyclopedia of Biostatistics*, vol. 6, *Survival Analysis* (eds P. Armitage and T. Colton), pp. 4372–4389. Chichester: 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* (eds M. E. Halloran and D. Berry), pp. 95–134. New York: Springer.
Robins, J. M. (2008) Causal models for estimating the effects of weight gain on mortality. *Int. J. Obesty*, **32**, suppl., S15–S41.
Robins, J. M. and Hernán, M. A. (2009) Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis* (eds G. Fitzmaurice, M. Davidian, G. Verbeke and G. Molenberghs), ch. 23, pp. 553–599. Boca Raton: Chapman and Hall–CRC.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Hoboken: Wiley.
Straus, M. A. and Mather, A. K. (1996) Social change and change in approval of corporal punishment by parents from 1968 to 1994. In *Family Violence against Children: a Challenge for Society* (eds D. Frehsee, W. Horn and K. D. Bussman), pp. 91–105. Berlin: deGruyter.
Toh, S. and Hernán, M. A. (2008) Causal inference from longitudinal studies with baseline randomization. *Int. J. Biostatist.*, **4**, article 22.