

i got the page lines to match as close as i could to the book. you should compare to orig text to get better sense of space.

### 4.5.5 Indirect Effects and the Mediation Formula

Remarkably, the definition of the natural direct effect (4.11) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the  $do(x)$  operator, to disable the direct link from  $X$  to  $Y$  so as to let  $X$  influence  $Y$  solely via indirect paths.

The natural indirect effect,  $IE$ , of the transition from  $x$  to  $x'$  is defined as the expected change in  $Y$  affected by holding  $X$  constant, at  $X = x$ , and changing  $Z$  to whatever value it would have attained had  $X$  been set to  $X = x'$ . Formally, this reads (Pearl 2001c):

$$IE_{x,x'}(Y) \triangleq E[(Y(x, Z(x'))) - E(Y(x))], \quad (4.14)$$

which is almost identical to the direct effect (Eq. (4.11)) save for exchanging  $x$  and  $x'$ .

Indeed, it can be shown that, in general, the total effect  $TE$  of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \triangleq E(Y(x') - Y(x)) = DE_{x,x'}(Y) - IE_{x',x}(Y). \quad (4.15)$$

In linear models where reversal of transitions amounts to negating the signs of their effects, we obtain a formal justification for the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \quad (4.16)$$

In the simple case of a single unconfounded mediator, (as in Figure 11.13) the natural direct and indirect effects are estimable through the following regression equations, called the Mediation Formula:

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \quad (4.17)$$

$$IE_{x,x'}(Y) = \sum_z E(Y|x, z)[P(z|x') - P(z|x)] \quad (4.18)$$

These provide two ubiquitous measures of mediation effects, applicable to any nonlinear system, any distribution, and any type of variables (Pearl 2009b, 2010b).

Note that the indirect effect has clear policy-making implications. For example: in a hiring discrimination environment, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of  $X$  on  $Y$  through a selected set of paths (Avin et al. 2005).

A general characterization of counterfactuals that are empirically testable is given in Chapters 7, 9, 11, and in Shpitser and Pearl (2007). In all these cases, the policy

intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001c) has suggested therefore that signal sensing is more fundamental to the notion of causation than manipulation; the latter being but a crude way of stimulating the former in experimental setup. (See Section 11.4.5.)