Analyzing Individual and Average Causal Effects via Structural Equation Models

Rolf Steyer

University of Jena, Germany

Abstract. Although both individual and average causal effects are defined in Rubin's approach to causality, in this tradition almost all papers center around learning about the average causal effects. Almost no efforts deal with developing designs and models to learn about individual effects. This paper takes a first step in this direction. In the first and general part, Rubin's concepts of individual and average causal effects are extended replacing Rubin's deterministic *potential-outcome variables* by the stochastic *expected-outcome variables.* Based on this extension, in the second and main part specific designs, assumptions and models are introduced which allow identification of (1) the variance of the individual causal effects, (2) the regression of the individual causal effects on the true scores of the pretests, (3) the regression of the individual causal effects on other explanatory variables, and (4) the individual causal effects themselves. Although random assignment of the observational unit to one of the treatment conditions is useful and yields stronger results, much can be achieved with a nonequivalent control group. The simplest design requires two pretests measuring a pretest latent trait that can be interpreted as the expected outcome under control, and two posttests measuring a posttest latent trait: The expected outcome under treatment. The difference between these two latent trait variables is the individual-causal-effect variable, provided some assumptions can be made. These assumptions – which rule out alternative explanations in the Campbellian tradition – imply a single-trait model (a one-factor model) for the untreated control condition in which no treatment takes place, except for change due to measurement error. These assumptions define a testable model. More complex designs and models require four occasions of measurement, two pretest occasions and two posttest occasions. The no-change model for the untreated control condition is then a single-trait–multistate model allowing for measurement error *and* occasion-specific effects.

Keywords: Rubin's approach to causality, structural equation modeling, latent difference variables

Individual and average causal effects are the basic concepts in Rubin's approach to causality. According to Rubin (e.g., 1974, 1978), an *individual causal effect* on the unit u is the difference between the *potential outcome* $Y_1(u)$ of u if treated and its potential outcome $Y_0(u)$ if not treated (control), while the *average causal effect* is the average of the individual causal effects in a population of units. Unfortunately, the individual causal effect is hard to determine in practice, because usually only the potential outcome $Y_0(u)$ under control *or* the potential outcome $Y_1(u)$ under treatment can be assessed. For example, if 'treatment' is teaching mathematics with a new method and 'control' is teaching it in the traditional way, we may assess $Y_0(u)$ *or* we may assess $Y_1(u)$, but we cannot assess both potential outcomes for the same student u. (Teaching mathematics with both methods would constitute a third method.) The same problem will be faced in psychological or medical treatments. This is what Holland (1986) called the *fundamental problem of causal inference*.

Facing this problem, most efforts have been spent in Rubin's tradition in developing procedures to estimate average causal effects. While this is simple and straightforward in a perfect randomized experiment (Rubin, 1974, 1978; Holland, 1986), it needs more sophistication and assumptions in nonrandomized studies (Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984; Rubin, 1973). For an overview see West (2000) or Winship and Morgan (1999).

In this paper we will show that Holland's fundamental problem of causal inference is less fundamental than it seems at first sight and that *there are* designs and models which allow examining the individual effects as well, although strong assumptions are necessary, restricting the range of possible applications.

This paper draws on ideas developed in several traditions in methodology: Rubin's approach to causality, the Campbellian tradition of quasi-experimentation and internal validity (e.g., Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002),

DOI 10.1027/1614-1881.1.1.39

structural equation modeling (e.g., Bentler, 2004; Jöreskog & Sörbom, 1996; Muthén & Muthén, 2004), especially latent state-trait modeling (Dumenci & Windle, 1996; Eid & Hoffmann, 1998; Steyer, Ferring, & Schmitt, 1992; Steyer, Schmitt, & Eid, 1999), latent change modeling (McArdle, 2001; Raykov, 1999, Steyer, Eid, & Schwenkmezger, 1997; Steyer, Partchev, & Shanahan, 2000b), and latent growth curve modeling (e.g., McArdle & Epstein, 1987; Meredith & Tisak, 1990; Tisak & Tisak, 2000; Willett & Sayer, 1996).

In the first and general part of the paper Rubin's concept of potential outcomes is replaced by the more general concept of expected outcomes, the expected values of a unit under treatment and under control. Expected outcomes are similar to true scores in classical psychometric test theory (CTT). However, instead of one single true score for each unit, we consider (at least) two for a single outcome variable: Its true score under treatment and its true score under control. In the second and more specific part, we will then present latent variable models that allow us to analyze and identify individual *and* average causal effects under feasible assumptions and designs. We show that much can be learned about the individual effects even if there is no random assignment of units to treatment conditions, provided that there are specific designs in which other assumptions can made.

We will discuss the designs and models presented by the following example: Assume that a training program is offered to persons who lost their job. The program aims at improving their social skills. Everybody is free to participate in the training program or not, so that there is self-selection to the treatment. However, in a *pretest*, the social skills of both participants and nonparticipants, are assessed via a test in which the behavior of a person in social situations (presented in short video sequences) is judged. In a *posttest*, the same test or a parallel form of this test is given 3 months after the training, or 3 months after the person could have ended the training if he or she had participated.

The basic idea of the new models to be introduced is to *use the units as their own control* in repeated measurements and assess the outcomes both in pretests and after treatment assuming that the pretests yield the same results as would be observed in the control condition. This strategy is not new (see, e.g., Winship & Morgan, 2000) and the associated problems of valid causal inference have been discussed extensively by Campbell and Stanley (1963), Cook and Campbell (1979), as well as Shadish et al. (2002). From this literature it is well-known that such pretest-posttest designs need strong assumptions and that there are several threats to the validity of causal inference in these designs.

In fact, many things may go wrong in our example. If, for instance, we compare the posttest means between the treated and the untreated group, the difference between the two means may not reflect the average causal effect of the training, because the persons with the higher skills may be more inclined to take the training so that the treatment group would have a higher mean already in the pretest. Several methods have been proposed to cure this problem: Comparing pre-post differences between the two groups (Campbell & Stanley, 1963), controlling for the pretest differences via analysis of covariance (Overall & Woodward, 1977; Rogosa, 1980; Winship & Morgan, 2000), propensity scores analysis (see, e.g., Rosenbaum & Rubin, 1983; Rubin, 2001), and matching (Ho, Imai, King, & Stuart, 2004; Winship & Morgan, 2000). All these methods indeed serve to estimate the average causal effect if certain assumptions hold. Even if there are no systematic pretest differences between the treatment and groups, there are a number of other alternative explanations for the mean differences: Historical or life events may have affected the persons in the training group to larger extent than those that did not take the training, and these events might also affect the social skills.

Instead of comparing the means between a treatment and a control group we might compare the means in the pre- and posttests of the treatment group. However, the pre-posttest mean change can also be misleading, i.e., it may not reflect the average causal effect of the treatment. Campbell and Stanley (1963) identified a number of alternative explanations for the mean change between pre- and posttest such as maturation, test-retest effects, historic or live events, etc.

In this paper it will be shown that all possible alternative explanations that might invalidate a causal interpretation of pretest-posttest changes can be tested and possibly excluded in appropriate designs with an untreated control group using specific structural equation models with latent variables. Furthermore, in these designs causal inferences can be made not only about the average causal effect but also about the individual causal effects. Latent state-trait models presented by Steyer et al. (1992) or Steyer et al. (1999) will play an important role as well as models with latent difference variables (McArdle, 2001; Raykov, 1999; Steyer et al., 1997, 2000c; Steyer, Krambeer, & Hannöver, 2004). We will explicate the specific assumptions under which individual causal effects can be represented as values of such latent difference variables. Note that these latent difference variables can be endogenous or dependent variables in more complex structural equation models that aim at explaining the interindividual differences in individual causal effects of a treatment or intervention – a goal that has been traditionally pursued by growth curve models (McArdle & Epstein, 1987; Meredith & Tisak, 1990; Tisak & Tisak, 2000; Willett & Sayer, 1996).

Table 1. An illustration of the concepts of expected outcomes as well as individual and average causal effects. Note: The symbol represents an intraindividual distribution.

Unit	$P(U{=}u)$ sampling probability	$E(Y \mid X = 1, U = u)$ Expected outcome under treatment	$E(Y \mid X = 0, U = u)$ Expected outcome under control	$E(Y \mid X = 1, U = u) -$ $E(Y \mid X = 0, U = u)$ Individual causal effect	Gender	$P(X = 1 \mid U{=}u)$ treatment probability in experiment 1	$P(X = 1 \mid U{=}u)$ treatment probability in experiment 2
u_1	1/8	82	68	14	male	1/2	2/3
u_2	1/8	89	81	8	male	1/2	2/3
u_3	1/8	101	89	12	male	1/2	2/3
u_4	1/8	108	92	16	male	1/2	2/3
u_5	1/8	118	112	6	female	1/2	1/3
u_6	1/8	131	119	12	female	1/2	1/3
u_7	1/8	139	131	8	female	1/2	1/3
u_8	1/8	152	148	4	female	1/2	1/3
			Average causal effect	10			

The organization of the paper is as follows: We first introduce and generalize Rubin's concepts of individual and average causal effects replacing Rubin's potential-outcome variables by the expected-outcome variables. Summarizing previous results (Steyer, Gabler, von Davier, & Nachtigall, 2000a), we then show how to identify (1) the variance of the individual causal effects, (2) the regression of the individual causal effects on the true scores of the pretests, (3) the regression of the individual causal effects on other explanatory variables, and (4) the individual causal effects themselves.

Extending the Theory of Individual and Average Causal Effects

Why Extend Rubin's Concepts?

As mentioned previously, according to Rubin, an *individual causal effect* of the unit u is the difference between the potential outcome $Y_1(u)$ of u if treated and its potential outcome $Y_0(u)$ under control, while the *average causal effect* is the average of the individual causal effects in a population of units. Both for theoretical and for practical purposes, we favor a slight generalization of these concept as presented in Steyer, Gabler, von Davier, Nachtigall, and Buhl (2000b). We assume an intraindividual distribution within each unit-treatment combination, the expected value of which will be called the expected outcome of unit u in treatment x (see Table 1).

Such a stochastic approach was already used in the papers of Neyman (1923/1990, see also Neyman, Iwaszkiewicz, & Kolodziejczyk, 1935) and is shared by others (see, e.g., Pearl, 2000; Greenland, Robins, & Pearl, 1999; Robins & Greenland, 2000). The theoretical reason for this extension is that we consider an outcome of a treatment for a unit u not to be a simple unique number, because this view would be too deterministic (see also Dawid, 2000). Are we really willing to call the death of a patient 5 years after his heart surgery the "outcome of his surgery," even if we know that this patient died in a car accident? Or, less dramatically, should we really consider Jack's mark in his mathematics examination the outcome of taking our course, knowing that he did not have enough time for his home work because he had to earn his living before the examination? Or, referring to our social skills training example, are the observed social skill scores in the posttest really due exclusively to the treatment condition and the interindividual differences, or is it more reasonable to assume that there are also measurement errors and situational effects in the test scores? In all these examples, there are clearly several causal factors, aside from the treatment conditions and the interindividual differences, which affect the outcomes actually observed.

Rubin uses the term *potential* outcomes in the sense that either the outcome under treatment or the outcome under control will be observed, depending only on whether or not the unit is treated. Hence, the uncertainty is only in whether or not there is treatment. In contrast, the arguments presented in the last paragraph invoke an additional stochastic component, because they imply that there is a *set* of potential outcomes for each unit under treatment and a set of potential outcomes for each unit under control. Under treatment and given a unit u, each potential outcome has an (unknown) probability to occur and the same holds true for control. Which of these potential outcomes actually occurs for a given person depends not only on whether or not the person is treated but also on the other causal factors. Hence treatments and interventions can *affect the intraindividual distribution* of an outcome variable for a given unit, but they do not *determine* the outcome to be a fixed value. Hence, instead of a fixed potential outcome $Y_1(u)$ of the treatment, we assume that there is an intraindividual distribution for each unit u (its *potential outcome distribution*), and we replace the potential outcome $Y_1(u)$ by the *expected outcome* $E(Y \mid X = 1, U = u)$ of the treatment $X = 1$ for the unit u (see Table 1). The same applies, of course, to the control condition and its associated potential outcome $Y_0(u)$.

The practical reason for this stochastic extension is that the expected values $E(Y \mid X = 1, U = u)$ can be considered the true scores of the outcome variable Y in the treatment condition $X = 1$, and the same applies to the expected values $E(Y \mid X = 0, U = u)$, the expected outcomes for the control condition. And, true-score variables are easily modeled as *latent variables* (see, e.g., Jöreskog, 1971; Steyer, 2001; Steyer & Eid, 2001). Hence, while Rubin's deterministic concept of a potential outcome may have the advantage of greatest simplicity and may be useful in calling for missing data techniques, the concept of an expected outcome will prove useful, because it invites latent variable modeling.

The Single-Unit Trial

The concepts mentioned above may be better understood if we make explicit the kind of random experiment used for defining these concepts. This kind of random experiment will be referred to as the *single-unit trial*: Draw a unit u (e.g., a person) out of a set of units (the population of units), observe the numerical value z of a pretest Z (or several pretests Z_i), assign the unit (or observe its self-assignment) to one of two experimental conditions and register the numerical value y of the *outcome variable* Y (or several outcome variables Y_i). This single-unit trial is not the sample with which statistical models are usually dealing. In a sample, the single-unit trial is repeated many times. Specifically, the single-unit trial does not allow treating problems of parameter estimation and hypothesis testing. However, it is sufficient for the purpose of introducing the concepts of individual and average causal effects. (For a more detailed description see Steyer et al., 2000a).

A fundamental assumption in the single-unit trial is that each unit u has a probability of being assigned to the treatment condition which is between 0 and 1, excluding these two bounds:

$$0 < P(X = 1 \mid U = u) < 1, \text{ for each unit } u. \tag{1}$$

Without this assumption there would be units which either have a zero probability of being treated or of being in the control condition. In this case it would neither make sense to define a potential-outcome distribution nor the expected outcome in the treatment *and* in the control condition for such units.

Each possible outcome ω of the single-unit trial can be described by the quadruple $\omega = (u, z, x, y)$. The non-numerical random variable U defined by $U(\omega) = u$, called the *observational-unit variable*, maps each possible outcome of the single-unit trial onto the individual u. The pretest Z is defined by $Z(\omega) = z$. The random variable X defined by $X(\omega) = x$ is called the *treatment variable* and maps each possible outcome onto the treatment condition. For simplicity, let $X = 1$ for treatment and $X = 0$ for control. Finally, there is the outcome variable Y defined by $Y(\omega) = y$. We presuppose that the pretest Z and the outcome variable Y are real-valued variables with finite expected values. Note that all random variables mentioned in this paragraph refer to this single-unit trial. Figure 1 shows that each combination of unit, treatment condition, and score of the outcome variable may be an outcome of such a single-unit trial. Therefore, the variables U, Z, X, and Y defined above have a *joint distribution*.

Individual and Average Causal Effects

In the single-unit trial, $E(Y \mid X, U)$ denotes the (unknown) regression (or conditional expectation) of the outcome variable Y on X and U with values $E(Y \mid X = x, U = u)$, the *individual expected values of Y given X = x and U = u*. For example, $E(Y \mid X = 1, U = u_7) = 139$ is

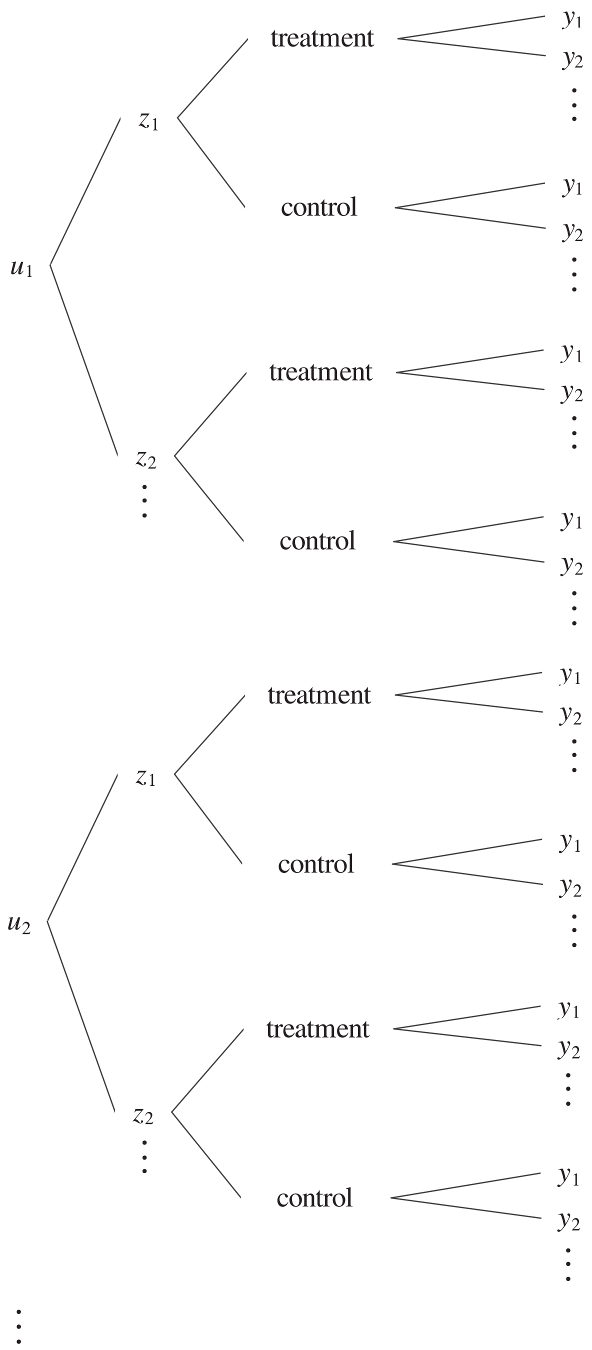

Figure 1. The set of potential outcomes of the single-unit trial.

the expected outcome if Unit 7 is drawn and assigned to the treatment condition $X = 1$ (see Table 1). Hence, $E(Y \mid X = 1, U = u)$ is called the *expected outcome of unit u under treatment*, $E(Y \mid X = 0, U = u)$ the *expected outcome of unit u under control,* and $E(Y \mid X = 1, U = u) - E(Y \mid X = 0, U = u)$ the *individual causal effect*.

We may now define the functions $f_0(U)$ and $f_1(U)$ by

$$f_0(u) = E(Y \mid X = 0, U = u), \qquad (2)$$

$$f_1(u) = E(Y \mid X = 1, U = u) - E(Y \mid X = 0, U = u), \qquad (3)$$

for each unit u, respectively. Hence, $f_0(U)$ is the *expected-outcome-under-control variable* and $f_1(U)$ the *individual-causal-effect variable*.

Presupposing a dichotomous treatment variable X with values 0 and 1, the following equation always holds:

$$E(Y \mid X, U) = f_0(U) + f_1(U) \cdot X. \qquad (4)$$

According to this equation, the regression of Y on X given a unit u is a linear regression with unit-specific constant $f_0(u)$ and unit-specific slope $f_1(u)$. Hence, we neither presume homogeneity of the expected outcomes under control nor homogeneity of the differences $ICE(u)$: = $f_1(u)$, the individual causal effects. The *average causal effect* (*ACE*) of treatment variable X (i.e., treatment X = 1 vs. $X = 0$) on the (expected value of) the response variable Y is now defined by $ACE := E[f_1(U)]$.

Another way to write Equation (4) is to add and subtract the expectation $E[f_0(U)]$ and the term $E[f_1(U)] \cdot X$:

$$E(Y \mid X, U) = E[f_0(U)] + E[f_1(U)] \cdot X + f_0(U) - E[f_0(U)] + (f_1(U) - E[f_1(U)]) \cdot X, \qquad (5)$$

which decomposes the regression $E(Y \mid X, U)$ into a *fixed part* $E[f_0(U)] + E[f_1(U)] \cdot X$ and a *random part* $f_0(U) - E[f_0(U)] + (f_1(U) - E[f_1(U)]) \cdot X$ (cf. Bryk & Raudenbush, 1992). Learning about the parameter $E[f_1(U)]$ in the fixed part means learning about the average causal effect in the population, whereas learning about the random variables $f_0(U) - E[f_0(U)]$ and $f_1(U) - E[f_1(U)]$ means learning about the expected outcomes under control and the individual causal effects, or at least about parameters characterizing their marginal and/or joint distributions.

The path diagram in Figure 2 represents Equation (4). It shows that X may depend on U and that U modifies the effect of X on Y.

Equation (4) is represented in a different way in Figure 3 with a separate path diagram for each of the two experimental conditions with the latent variable $f_0(U)$ in the control condition and the two latent variables $f_0(U)$ and $f_1(U)$ in the treatment condition. Whereas the individual causal effect function $f_1(U)$ appears as a modifying function at the path from X to Y in Figure 2, it occurs as a latent variable in Figure 3.

Although we will restrict the discussions in this paper to the case in which X is dichotomous, the generalization of Equation (4) to $q + 1$ treatment conditions is obvious:

$$E(Y \mid X_1, ..., X_q, U) = f_0(U) + f_1(U) \cdot X_1 + ... + f_q(U) \cdot X_{q,} \qquad (6)$$

where each X_k, $k = 1, ..., q$, is a dichotomous variable indicating with the values 1 and 0 whether or not a unit is assigned to treatment condition k.

Note that, in defining these concepts, we take the *prefactual perspective* from which the single-unit trial de-

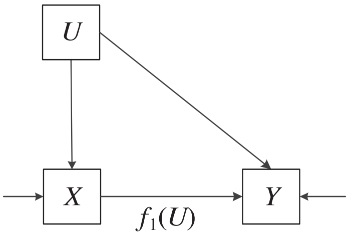

Figure 2. Path diagram of the dependencies between the three variables *X*, *U,* and *Y*.

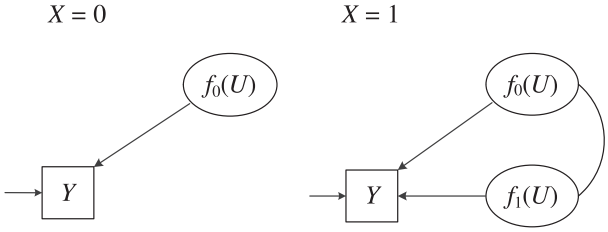

Figure 3. Path diagram with the latent variables $f_0(U)$ and $f_1(U)$ in the two experimental conditions.

scribed above is still to be conducted. It is not a counterfactual perspective implying that we look at data resulting from such a single-unit trial. Considering the expected-outcome-under-control variable $f_0(U)$ instead of Rubin's potential-outcome variable Y_0 and the individual-causal-effect variable $f_1(U)$ instead of Rubin's $Y_1 - Y_0$, we also allow for other effects on *Y* that are caused neither by *X* nor *U* (such as the effects of the car accident in the surgery example, the fact that the student had to earn his living before the examination in the teaching example discussed earlier, and effects caused by measurement error and situational effects in the social skills example).

Designs and Models

Randomized Single-Posttest Design

So far, the individual causal effect is a theoretical concept just like the true score in Classical Test Theory. Without introducing assumptions (defining models) there is no way to identify the individual causal effects or at least their expected value or their variance. "To identify" these parameters simply means to show that they can be computed from the parameters of the joint distribution of the observable or "manifest" variables, such as *X* and *Y*. If this is the case, then these parameters can also be estimated via the corresponding sample analogs.

Steyer et al. (2000a) showed how to identify the ACE defined above. Their crucial assumption is that the person-variable *U* and the treatment-variable *X* are *independent*, which can be made true if the person is randomly assigned (e.g., by a coin toss) to the treatment or the control condition. Since we only assume a single posttest we may call this first design the randomized single-posttest design. The next to last column in Table 1 characterizes this design: The treatment assignment probability is the same for all units. Hence, it does not depend on the variable *U* nor on the expected outcomes under treatment and control. (Later we will deal with designs that do not require this independence assumption.) The last column in Table 1 shows a design in which the treatment assignment probability is not independent of the units. In this design the covariate *gender* determines the assignment probability: Given gender, there is conditional independence between *X* and *U*. Steyer, Nachtigall, Wüthrich-Martone, and Kraus (2002) showed how to identify the average causal effect in such a case and much of the work related to propensity scores deals with this problem (see, e.g., Rosenbaum & Rubin, 1983; Rubin, 2001).

Multiple Pre- and Posttests Designs

We will now introduce new designs and models that do not rely on random assignment of the unit to one of the treatment conditions. For simplicity, we will denote the individual-causal-effect variable $f_1(U)$ by f_1 and the expected-outcome-under-control variable $f_0(U)$ by f_0 in the sequel. In these new designs it will be possible to identify not only the average causal effect on the treated, i.e., $E\,[f_1 \mid X = 1]$, but also the variance $Var(f_0 \mid X = 1)$ of the expected outcomes under control of the treated, the variance $Var[f_1 \mid X = 1]$ of the individual causal effects of the treated, and the covariance $Cov(f_0,\ f_1 \mid X = 1)$ between the expected outcomes under control and the individual causal effects of the treated. These parameters provide valuable information. $Var(f_0 \mid X = 1)$ informs us how the expected outcomes under control differ among the treated, and $Var[f_1 \mid X = 1]$ how different the individual causal effects are from each other among the treated. $Cov(f_0,\ f_1 \mid X = 1)$ describes how the individual causal effects correlate with the expected outcomes under control among the treated. Speaking in terms of Equation (5) we will now show how to identify both the fixed and the random part of the regression $E(Y \mid X, U)$. In a later section we will add the assumption that *U* and *X* are independent (e.g., created by random assignment). In this case all these parameters will not only describe the group of the treated but the total population.

Of course, identifying the variance of the individual-causal-effect variable f_1 and the other theoretical parameters mentioned above will only be possible if we introduce some assumptions. Exactly the same problem also pertains to the true-score variables in CTT: There is nothing we can learn about the variance of the true-score variable in CTT

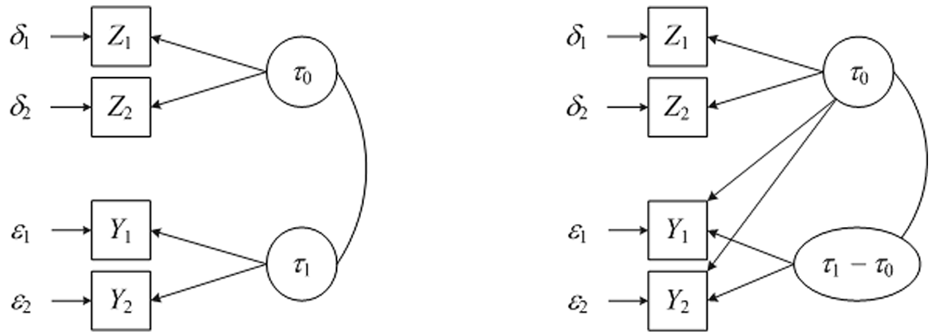

Figure 4. Path diagrams of a pretest-posttest model with two τ-equivalent tests and uncorrelated errors: The latent state version is on the left-hand side, the latent change version on the right-hand side.

unless we introduce a model, such as the models of parallel or congeneric tests. What are the necessary assumptions under which the conditional expected value $E(f_1 \mid X = 1)$, the conditional variance $Var(f_1 \mid X = 1)$, and the conditional covariances of f_1 with other variables (such as f_0) are identifiable?

First of all, let us emphasize that Holland's fundamental problem of causal inference does not *always* apply. If, for example, we compare a teaching method against a control in which there is no teaching at all, it seems possible that a measurement of the outcome variable before the students are treated (i.e., a pretest) would yield the same results as if these students were assigned to the untreated control condition and the outcome variable would be measured without a foregoing treatment. Hence, if the treatment is to be compared to an untreated control condition and if alternative explanations such as history, test-retest effects, etc. (Campbell & Stanley, 1963) can be ruled out, a pretest might be useful as a first design element.

Second, the expected-outcome-under-control variable f_0 and the individual-causal-effect variable f_1 are latent variables and we know from latent variable modeling that each latent variable needs at least two manifest measures in order to identify its variance. Hence, as a second design element, two parallel pretests and two parallel posttests seem necessary, at least. Assuming uncorrelated measurement errors will imply identifiability of the variances and covariances of these latent variables

We proceed in several steps. First, we will introduce a latent variable model containing a *latent difference variable*, which is the individual-causal-effect variable f_1 if certain additional assumptions are made. In the next steps, we will introduce these additional assumptions leading to several models in which the individual-causal-effect variable f_1 occurs as a latent variable.

True-Change Model

Latent variable models containing latent difference variables have been introduced by Steyer et al. (1997), Raykov (1999), McArdle, (2001), and Steyer et al. (2000c). In these models there are at least two variables Z_1 and Z_2 (here: two pretests) measuring the same latent variable at a first occasion, and two variables Y_1 and Y_2 (here: two posttests) measuring a common latent variable at a second occasion. For the time being there may or may not be a treatment between pre- and posttests, although, in the next section, we will presume that there is a treatment. The path diagrams in Figure 4 depict such a model in two versions: The *latent state version* and the *latent change version*.

In classical test theory, the two assumptions mentioned above are called the τ-equivalence of the two pretest and two posttest variables:

$$E(Z_1 \mid U) = E(Z_2 \mid U) =: \tau_0 \text{ and } E(Y_1 \mid U) = E(Y_2 \mid U) =: \tau_1. \quad (7)$$

These assumptions have to be supplemented by assuming uncorrelated errors:

$$Cov(\delta_1, \delta_2) = Cov(\varepsilon_1, \varepsilon_2) = Cov(\delta_i, \varepsilon_j) = 0, \quad i, j = 1, 2. \quad (8)$$

The path diagram depicted at the left-hand side of Figure 4 shows this model, which is well-known (see, e.g., Steyer et al., 1997, 2000c).

Considering the trivial equation

$$\tau_1 = \tau_0 + (\tau_1 - \tau_0) \quad (9)$$

and inserting it into the second of the Equations (7) yields

$$E(Y_1 \mid U) = E(Y_2 \mid U) =: \tau_0 + (\tau_1 - \tau_0), \quad (10)$$

which can be translated into the path diagram (see the right-hand side of Figure 4) of a structural equation model containing a true-change or latent difference variable, the difference $\tau_1 - \tau_0$ between the common true-score variable of the posttests and the common true-score variable of the pretests.

Additional Assumptions for Causal Inference

We now assume that there is a treatment between pre- and posttests. Two additional assumptions have to be added that are necessary for a causal interpretation of the

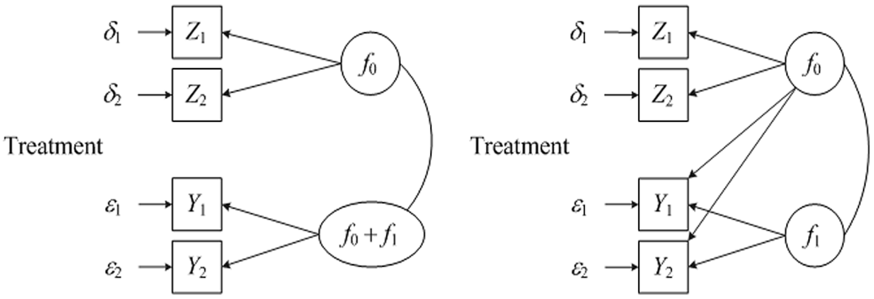

Figure 5. Two equivalent path diagrams in the treatment group if assumptions (11) and (12) hold. On the left-hand side is the latent state version, and on the right-hand side the latent change version with the individual-causal-effect variable f_1.

latent difference variable $\tau_1 - \tau_0$ as the individual-causal-effect variable f_1:

$$\tau_0 = f_0, \tag{11}$$

the *equivalence of the true-score variable of the pretests with the expected-outcome-under-control variable* f_0, and

$$\tau_1 = f_0 + f_1, \tag{12}$$

the *equivalence of the true-score variable of the posttests with the expected-outcome-under-treatment variable* $f_0 + f_1$. If these two assumptions hold, then the path diagram of Figure 4 can be changed as depicted in Figure 5.

If these two assumptions, (11) and (12), are correct, then (1) the values of $\tau_1 - \tau_0$ are the individual causal effects of the treatment (which is given between pre- and posttests), (2) the expected value of $\tau_1 - \tau_0$ is the average causal effect, (3) the variance of $\tau_1 - \tau_0$ tells us how much the individual causal effects deviate from the average causal effect, and (4) the regression coefficient β_1 describes how strongly the individual causal effects depend on the expected-outcome-under-control variable $\tau_0 = f_0$. Furthermore, additional variables could be introduced explaining the interindividual differences in the individual causal effects.

Testing the Assumptions Allowing for Causal Inferences

All these causal interpretations rely on the assumptions (11) and (12). If they are not fulfilled, the causal interpretations listed in the last paragraph are no longer valid and the difference variable $\tau_1 - \tau_0$ is only an ordinary latent change variable, the scores of which may also be caused, at least in part, by other factors than the treatment. According to Campbell and Stanley (1963), (historical or life) events between pretest and posttest may occur, which can also cause the change from τ_0 to τ_1. In this case the change is not caused by the treatment alone, and in the worst case, not at all. Furthermore, change may be caused by the passing time or by associated processes such as maturation, aging, change in hormone levels, and, in shorter time intervals, fatigue, hunger, thirst, etc. In our social skills example, a true change between the pre- and posttests of some subjects could be due to maturation if the time interval is large enough and/or the subjects are of an age at which there is still fast maturation. Personal life events or historical events such as a TV series showing “models” with high social skills might also be responsible for (some of) the true change between the pre- and posttests of some individuals.

Hence, there is good reason to think about how to test the assumptions (11) and (12). How can we rule out the alternative explanations for the change from τ_0 to $\tau_{1,}$ which invalidate the interpretation of $\tau_1 - \tau_0$ as the individual causal effect variable f_1? According to assumption (11), the common true-score variable τ_0 of the pretests is equivalent to the expected-outcome variable f_0 in the control condition $X = 0$. Only in this case the difference between the true-score variables of the posttests and the pretests is due to the treatment.

Of course, there is no way to verify assumption (11), because this would require subjects who are treated and are not treated at the same time. However, if the assumption (11) is true, then the true-score variables of the pre- and posttests should be identical in a control condition, in which there is no treatment (see Figure 6). This requirement would not be fulfilled, e.g., if one or more of the factors mentioned in the last paragraph would affect the posttests, or if there were a test-retest effect such that the mere fact of having a pretest has an effect on the posttest variables and if this testing effect would be different for each unit. Assumption (11) would only be fulfilled if there are fluctuations between pre- and posttests that are entirely due to error.

Hence, in order to rule out all the alternative explanations for the true change, we need a new experiment (design) in which there is also an untreated control condition as a third design element. In other words, after the pretests, there is a treatment $X = 1$ with probability

$$0 < P(X = 1 \mid U = u) < 1, \text{ for each unit } u, \tag{13}$$

and no treatment with probability $P(X = 0 \mid U = u) = 1 - P(X = 1 \mid U = u)$. (In the previous experiment we as-

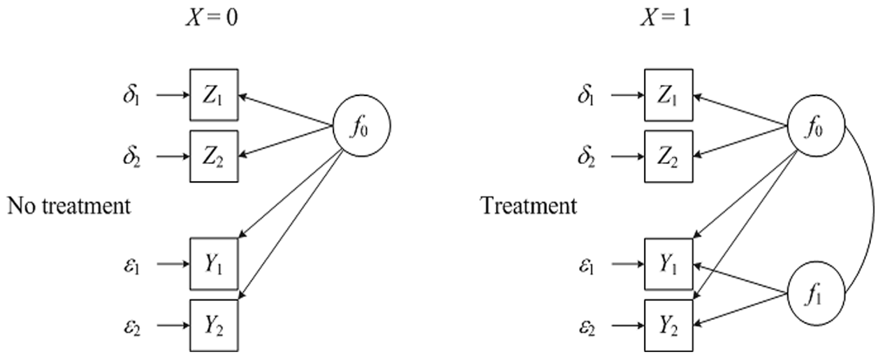

Figure 6. Path diagrams in the treatment and control conditions if assumption (11) holds.

sumed that there is a treatment with probability 1 for each unit u.) Referring to this new experiment, we can formulate the necessary assumptions mentioned above:

$$E(Z_i \mid X = 0, U = u) = E(Y_j \mid X = 0, U = u) = f_0(u),\quad \text{for each unit } u, \quad i, j = 1, 2. \tag{14}$$

Hence, in the (untreated) control condition not only the true-score variables *within* the pretests and those *within* the posttests are identical, but also the true-score variables *between* pre- and posttests. In this control condition ($X = 0$) there is no treatment and no other systematic change between pre- and posttest. All the observable change is due to uncorrelated error variables (see Figure 6). Hence, the four true-score variables are also identical with the expected-outcome-under-control variable f_0 in the control condition of the single-unit trial.

What about assumption (11) in the treatment condition? This assumption only postulates the equivalence of the true-score variables of the pretests with the expected-outcome-under-control variable,

$$E(Z_i \mid X = 1, U = u) = f_0(u), \quad \text{for each unit } u, \quad i = 1, 2. \tag{15}$$

Hence, this assumption seems relatively unproblematic, especially if we can assume that there is no systematic change between pre- and posttests under no-treatment, and this is exactly what is tested in the control condition.

What remains is assumption (12), according to which the true-score variable τ_1 of the posttests is identical with the expected-outcome-under-treatment variable $f_0 + f_1$. Here again we can rely on what we assume – and in applications also test – for the control condition: Why should there be, aside from the treatment effect, another systematic effect if such an effect does not exist in the control condition?

In a sampling experiment in which the single-unit trial is repeated many times, one could argue that there are such other effects such as maturation, life events etc. for the units in the treatment condition but not for the units in the untreated control condition. In fact, this could be a threat to the validity of our causal inferences if the units treated were of a different kind than the units in the control condition. However, this case is not very likely as long as assumption (13) holds, according to which each unit has a probability between 0 and 1 of being treated, *excluding* 0 and 1. If in fact the population of units u could be partitioned into a set with $P(X = 1 \mid U = u) = 1$ for all its members u and a set with $P(X = 1 \mid U = u) = 0$ for all its members u, then we would not be able to preclude that different factors other than treatment would exert their effects in the treatment group, but not in the control group. With assumption (13), however, each unit has a chance to be treated and a chance to remain untreated. Hence, under this assumption, it is unlikely that an alternative factor exerts its effect in the treatment but not in the control condition.

Would it be possible to test assumption (13)? As far as I can see, the answer is “no.” Whether or not it holds is rather a matter of judgment about the procedure of assignment or selection to treatment. Did the persons in the treatment group also have a real chance of being in the control and vice versa?

So far, we have not assumed randomized assignment of the unit to one of the treatment conditions. Therefore, it is very possible that alternative factors such as maturation, history, life events, etc. exert their effect to a greater extend in the treatment than in the control group. This is a serious threat to validity of causal inference in ordinary designs and models in which we compare the mean in the treatment to the mean in the control condition. However, in the design and model proposed above we rule out this threat by postulating and testing that, except for error, there is no change at all between pre- and posttests in the (untreated) control group: No change in the means, but also no change in the true-scores of the units. With this postulate we rule out *any alternative explanation* of the change observed in the treatment group. If this postulate fails and has to be rejected in a specific application, the causal interpretation of the latent-change variable in the treatment group as the individual-causal-effect variable is invalidated.

Testing this postulate of no true change in the control group, traditional analysis of variance techniques would fall too short, because they only allow for testing the hypothesis of no change in the group means between pre- and posttest, but not of no change on the level of the units. However, even if there is no change in the group means, there can be true individual change if the gains for some units are equalized by the losses for other units. Hence, this kind of hypothesis calls for latent variable modeling techniques that not only allow for testing the hypothesis of no change in the group means between pre- and posttests, but also in the true-scores of the units.

Additional Constraints and Causal Inferences Under Randomization

So far, we have not used the assumption of *U* and *X* being independent

$$P(X = 1 \mid U = u) = P(X = 1), \text{ for each unit } u, \qquad (16)$$

which could be created via random assignment of the unit to one of the treatment conditions. We do not necessarily need this assumption, because the unit in the treatment condition serves, via the pretest, as its own control. This is what the assumptions discussed above are about. The only assumption we need is that each unit *u* has a probability between 0 and 1 of being assigned to the treatment condition [see Eq. (13)]. Otherwise one could argue that there are different processes for the units that are assigned to the treatment condition as compared to the units assigned to the control condition. In such a case we would not be able to infer that f_1 is the individual-causal-effect variable. In other words, in the design and model proposed above we do not compare treatment to control. Instead our inferences are based on the pretest-posttest comparisons in the treatment condition. The control condition is only used to rule out alternative explanations for these pretest-posttest differences.

If we additionally can assume that *X* and *U* are independent (created, e.g., by random assignment), a number of additional testable consequences follow and some additional causal inferences can be drawn. First, the expected values and the variances of the pretest true-score variables in treatment and control conditions are identical:

$$E(f_0 \mid X = 0) = E(f_0 \mid X = 1), \qquad (17)$$

$$Var(f_0 \mid X = 0) = Var(f_0 \mid X = 1). \qquad (18)$$

Second, the variances of the error variables of the pretests in the treatment and control conditions are identical:

$$Var(\delta_i \mid X = 0) = Var(\delta_j \mid X = 1), \qquad i, j = 1, 2. \qquad (19)$$

Finally, we also expect the equality of the variances of the error variables of the posttests in the treatment and control conditions:

$$Var(\varepsilon_i \mid X = 0) = Var(\varepsilon_j \mid X = 1), \qquad i, j = 1, 2. \qquad (20)$$

Note, however, that there might also be an effect of the treatment on the error variance. Hence, in contrast to Equations (17) to (19), the last equality is *not* a logical consequence of randomization.

Aside from these additional possibilities of testing the assumptions, there are other – and maybe more important – benefits from randomization. These concern the additional causal inferences. The first is:

$$E(f_1 \mid X = 1) = E(f_1) = ACE, \qquad (21)$$

i.e., under randomization, the expected value of f_1 in the treatment condition is not only the average causal effect on the treated, but also the average causal effect in the total population. The second additional inference is:

$$Var(f_0 \mid X = 1) = Var(f_0). \qquad (22)$$

Hence, under randomization, the variance of f_0 in the treatment condition is not only the variance of the true-score variable of the pretests of the treated, but also the variance of the true-score variable of the pretests in the total population. Similarly,

$$Var(f_1 \mid X = 1) = Var(f_1), \qquad (23)$$

i.e., under randomization, the variance of f_1 in the treatment condition is not only the variance of the individual causal effects of the treated, but also the variance of the individual causal effects in the total population. And finally,

$$Cov(f_0, f_1 \mid X = 1) = Cov(f_{0,} f_1), \qquad (24)$$

i.e., under randomization, the covariance of f_0 and f_1 in the treatment condition is not only the covariance between the true-score variable of the pretests and the individual-causal-effect variable of the treated, but also the covariance between the true-score variable of the pretests and the individual-causal-effect variable in the total population. Furthermore, all other parameters such as correlations or regression coefficients that can be computed from the parameters mentioned above for the treatment condition will be valid for the total population as well. All this follows from the fact that under randomization *U* and *X* are independent.

These interpretational benefits may be considered the most important ones of a randomized design with latent individual-causal-effect variables, but of course only the last three equations provide additional knowledge not available in an ordinary randomized experiment. These additional inferences become possible through the pretest-posttest comparisons of the true-score variables and the assumption that all the pretest-posttest change is caused by the treatment, an assumption which can only be true if there is no systematic pretest-posttest change

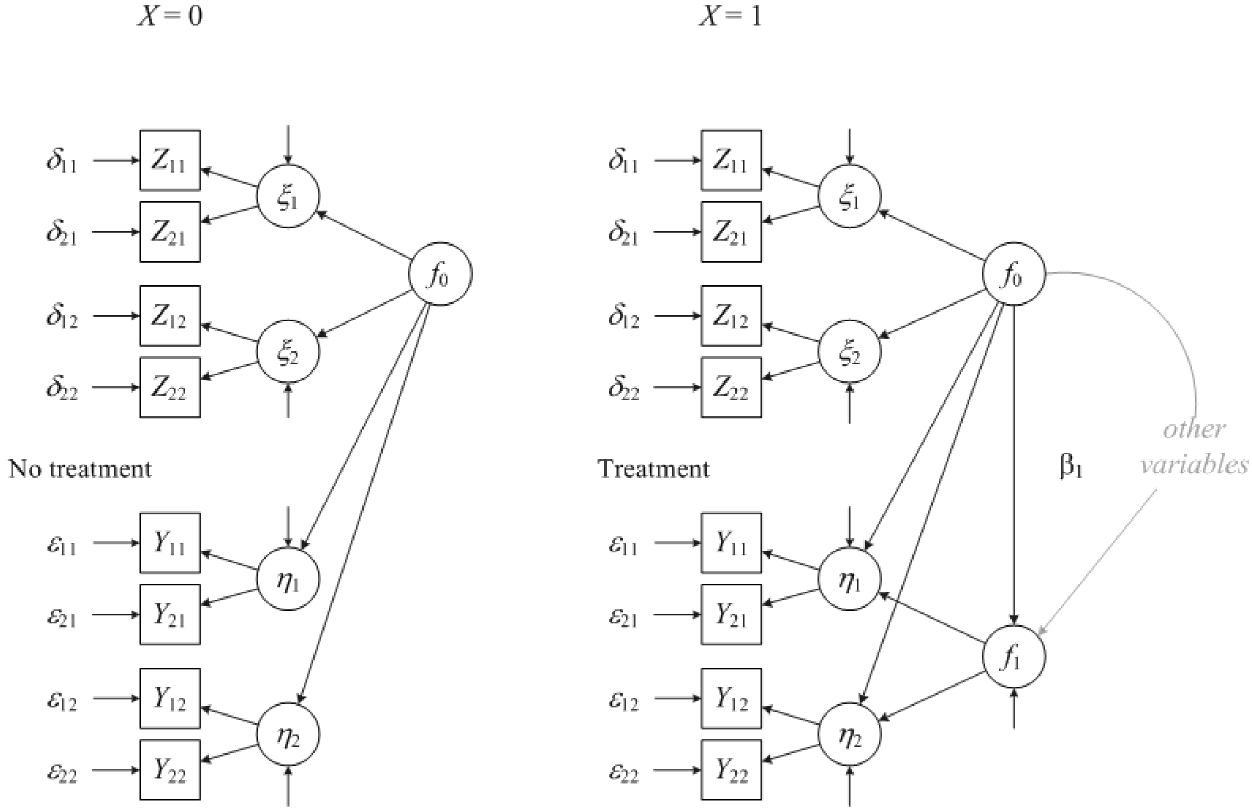

Figure 7. Single-trait–multistate model in the control condition and the additional trait change variable in the experimental condition.

in the control group, neither in the means nor in the individual true scores.

More General Designs and Models

It should be noted that the model treated in the previous section is only the simplest one containing an individual-causal-effect variable. In many applications this model will not hold, if the two pretests are assessed at a first occasion of measurement and the two posttests at a second one. The reason is that our measurements do not take place in a situational vacuum, i.e., there are usually occasion-specific effects (due to the specific situations in which the persons are at the occasion of measurement) that determine the observable measures to some degree, even if we intend to measure traits such as neuroticism or intelligence (see, e.g., Deinzer et al., 1995), and these occasion-specific effects affect both measures assessed within the same occasion. Such person-specific situational effects are also to be expected in our social skills example, because the social skills scores are, to some extent, also due to the mood states of the tested persons, for instance. If this is the case, models with a single latent trait variable in the control condition [see Figure 6 and Eq. (14)] will not hold in designs with two occasions of measurement, because they do not allow for these occasion-specific effects within the two occasions.

Hence, different designs are required for the model presented in Figure 6 to be valid in a concrete application. The first is a design in which both pre- and posttests are assessed in the same occasion of measurement. Second, instead of assessing the two pretests within one occasion and the two posttests within another, the two pretests could be assessed at two different occasions *before*, and the two posttests at two different occasions *after* the treatment. In this case, the occasion-specific effects will be different at each of the four occasions of measurement and, hopefully, be uncorrelated. In this case they will be part of the error terms δ_i and ε_i and would not be separated from the original measurement error variables.

Latent Trait Change Model

Our next design is similar to the previous one. However, instead of assessing only one single measure at each of the four time points, we now take at least two. This will disentangle occasion-specific effects and measurement errors as well and replace the single-trait model for the control condition (see the left path diagram in Figure 6

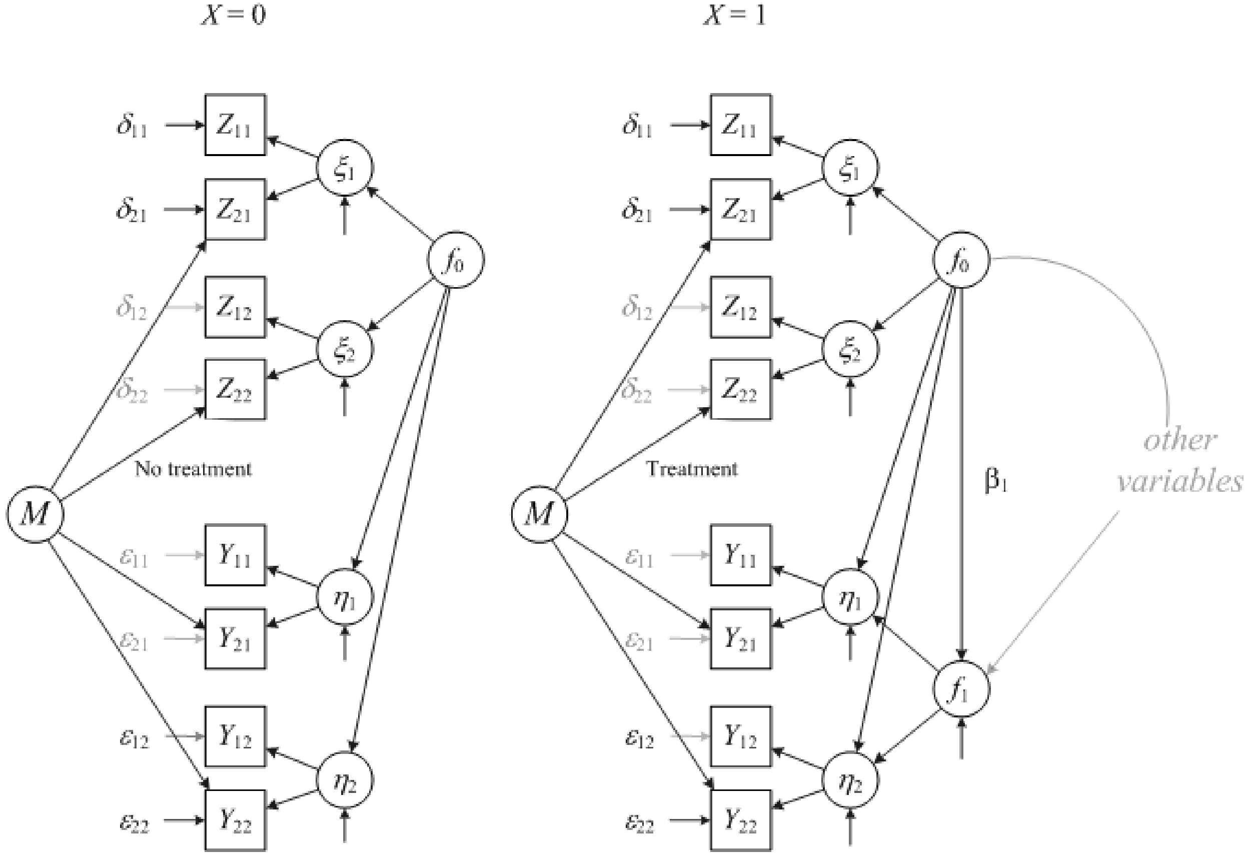

Figure 8. Single-trait–multistate model in the control condition and the additional trait change variable in the experimental condition with method factor.

by the single-trait–multistate model (see Figure 7; Steyer et al., 1992).

Comparing Figure 6 to Figure 7 shows that the first order structure depicted in Figure 6 is the second order structure in Figure 7. Hence, the only difference between the two models is that we can now separate the measurement error variances from the occasion-specific variances: The covariance between two observable variables within one of the four occasions of measurement yields the variance of the corresponding latent state variable, which can be subtracted from the variance of the observed variable yielding the variance of the measurement error variable. Similarly, all other parameters in the model can be identified (see Steyer et al., 1999, for an overview of latent state-trait theory, its models, and applications).

If the two-group model depicted in Figure 7 is correct, then (1) the values of f_1 are the individual causal effects of the treatment (exerted between pre- and posttests), (2) the expected value $E(f_1 \mid X = 1)$ is the average causal effect on the treated, (3) the variance $Var(f_1 \mid X = 1)$ tells how much the individual causal effects deviate from the average causal effect of the treated, and (4) the regression coefficient β_1 describes how strongly the individual causal effects depend on the true-score variable f_0 of the pretests. Furthermore, other variables could be introduced explaining the difference in the individual causal effects.

Of course, under random assignment of the unit to one of the treatment conditions, the validity of these parameters can be extended from the treated group to the total population, just in the same way as in the previous model.

An important extension of the model presented in Figure 7 is obtained if we introduce *method factors*. Method factors are often necessary, because we rarely have strictly parallel measures within the occasions. Therefore, the observable variables measured with the same method (test form, item) have a method-specific component. Following Eid, Lischetzke, Trierweiler, and Nußbeck (2003), we can extend the single-trait–multistate model introducing a method factor for the second measure at each time point and set free the loadings of the second measures (see Figure 8). The first observable variable in each occasion of measurement serves as a reference with respect to which the other observable variable may have a method-specific component which is constant over time: The method factor. This defines the model presented in Figure 8.

Latent Trait Change Models with Stabilizing Base Lines

In psychology, the process of taking a test often changes the attribute to be measured. This is called the test-retest effect. Assessing ability in mathematics has also a training effect and/or may cause the subject to think about and learn what he or she did not master in the test. Similarly, assessing the quality of a marriage via a questionnaire may remind the subject that he/she could devote more time to the spouse. Watching and rating the videos involving social behavior might improve the social skills of the subject. In all these cases, the single-trait–multistate model in the control condition depicted in Figure 7 and Figure 8 will not hold. Our last design and model allow for initial test-retest effects that fade out after a certain time. The basic idea is: There is a time point in the pretest phase at which there are no further test-retest effects, so that testing itself does not change the attribute to be assessed any more.

This design requires at least five occasions of measurement: Three pretest occasions and two posttest occasions. In this design we allow for trait change between occasions one and two (e.g., due to testing effects), but we assume that there is no further trait change between occasions two to five in the untreated control condition (see Eid & Hoffmann, 1998, for a model with changing traits). In the treatment condition, the observable pretests at occasions two and three measure a common first latent trait and the observable posttests at occasions four and five measure a second one. The difference between the two can be interpreted as the individual-causal-effect variable.

Estimating Individual Causal Effects

So far we only mentioned the identification of the average causal effect, of the variances of the individual causal effects, and of the effects of other variables on the individual causal effects. What about the individual causal effects themselves? In the model depicted in Figure 7, for instance, it is clear that the difference between the average of the two posttests and the average of the two pretests is an unbiased estimator of the individual causal effect. The general answer is: Individual causal effects can be estimated in the same way as factor scores can be estimated in factor analysis models, because they are the values of a latent variable. Programs for structural equation models provide the necessary tools.

In this context it should be noted that the problem of *factor score indeterminacy* does not hold for the models we proposed. In these models the latent variables are introduced in a constructive way: Assumptions are made about the conditional expectations of the pre- and posttests given the unit-variable U, for instance. From these assumptions, the existence and uniqueness of the latent variables can be mathematically derived (see Steyer, 1989; Steyer, 2001). Hence, there is no doubt about the existence of the factor scores, their meaning, and uniqueness. Nevertheless, the factor score estimates are only estimates of the true factor scores. Hence, they will vary from one sample to the other in the same way as all estimates of parameters in statistics.

Discussion

In this paper it has been argued that Holland's (1986) fundamental problem of causal inference is not ubiquitous, and that there are phenomena in Psychology and related disciplines for which it is possible to use subjects as their own control in pretests. This strategy is attractive because equivalent control groups are often difficult or impossible to create. However, Campbell and Stanley (1963; see also Cook & Campbell, 1979; Shadish et al., 2002) have warned us of the threats to validity of causal interpretations of pretest-posttest differences. As a safeguard against all the alternative explanations of the pre-post difference we need an untreated control condition that allows us to argue that all alternative explanations would also be true for this control condition. This is where the single-trait–multistate model comes into play: It is a liberal and realistic formulation of the no-change hypothesis in the untreated control condition. If this model is true for the untreated control condition and we can assume that assumption (13) holds, all alternative explanations are invalidated and we can infer that the only reason for the change in the treatment condition is the treatment.

This kind of reasoning is well-known in the Campbellian tradition. However, in this tradition, this reasoning was only applied to the comparison of *expected values*: If there is no change in the expected values of the pretest as compared to the posttest in the untreated control group, the change in the corresponding expected values in the treatment condition are due to the treatment, and all alternative explanations (such as such as testing effects, history, maturation, etc.) are considered invalid. Using latent variable modeling we extended this kind of reasoning to the scores of latent trait variables and the individual true scores they represent: If there is no trait change in the untreated control condition, the trait change observed in the treatment condition is attributed to the person-specific treatment effects. In this way trait changes in the treatment condition can be interpreted as the individual causal effects of the treatment.

The designs and models discussed in this paper allow for the identification of (1) the expected value of the individual effects on the treated (i.e., the average causal effect on the treated), (2) the variance of the individual effects of the treated, (3) the individual effects themselves, (4) the effect of the expected outcome under control on the individual causal effects, and (5) the effects of further variables explaining the interindividual differences in the individual causal effects of the treatment. Under randomization, the validity of these parameters can also be extended to the total population.

With points (2) to (4) we go beyond what has been achieved in Rubin's tradition so far: The identification of the average causal effect in various designs and via various techniques. Only a slight generalization of Rubin's concepts was necessary: Replacing the potential-outcome variables by the expected-outcome variables. They seem to be the key opening the world of latent variable modeling for the analysis of Rubin's individual causal effects. From a formal point of view, expected-outcome variables in the treatment and in the control condition are just like true-score variables in CTT. Hence, all the techniques of CTT models and their extensions, latent state-trait models, become available for the analysis of individual causal effects in Rubin's sense. With the extensions of latent state-trait models for ordinal variables (Eid, 1996), the individual-causal-effect models can also be extended to binary and/or ordinal outcome variables.

Where are the limitations? Well, if there is real development and not just situation-driven fluctuations, the single-trait–multistate model postulating no trait change in the untreated control condition will not hold, provided that we have enough time points (at least four) so that it is over-identified. Via the models proposed, we will not be able in such a case to separate the natural developmental processes that we can see in the control condition from the processes due to the treatment that are seen in the treatment condition together with those natural developmental processes. For instance, wherever there is learning, spontaneous healing, effects of historical or life events, or maturation without treatment, the models proposed will not be valid and they will be rejected in the structural equation model testing procedures. Hence, the limitation is: There must be no true trait change between the last pretest phase and the posttests in the untreated control condition. What to do if there is such a true trait change? Well, in this case the pretest true-score variable is just an ordinary latent pretest variable and other techniques may be applied (see, e.g., Rosenbaum, 2002; Rosenbaum & Rubin, 1983, 1984; Rubin, 1973; Steyer et al., 2005) that allow inferences at least for the average causal effects and/or the conditional causal effects given the pretests true-score variable.

Aside from the limitations mentioned above it should be emphasized that almost all considerations made in this paper refer to the population level. No problems were discussed that pertain to fact that statistical inferences from sample to population have to be made. For example, in applications of the models presented in Figures 6 to 8 we have to accept the null hypothesis that the model holds for the control group before we can interpret the variable f_1 as the individual-causal-effect variable. Discussing the associated problems of power of the test and all other problems of statistical inference and decision making is beyond the scope of this contribution.

In a way, this paper may be considered a synthesis of different traditions in methodology: Rubin's approach to causality; the Campbellian tradition of quasi-experimentation and internal validity; and structural equation modeling, especially latent state-trait modeling, latent change modeling, and latent growth curve modeling. Rubin's approach provides the conceptual foundation: Individual and average causal effects. The Campbellian tradition gives the inspiration to introduce the untreated control group in which there is no trait change, so that we can rule out alternative interpretations of the latent change variable in the treatment condition. Structural equation modeling, especially latent state-trait modeling, provides a realistic formulation of the hypothesis that there is no trait change in the untreated control condition. Latent change models teach us how to include latent state and latent trait change variables as latent variables in structural equation models in such a way that they can depend on other variables explaining the interindividual differences in the individual causal effects of the treatment. Note that explaining interindividual differences in intraindividual change has also been the main purpose of latent growth curve models. Hence, they can be included in the list of the traditions of methodology from which this paper emerged, as well.

Finally, some aspects should be emphasized, which seem to be relevant for psychology as a whole discipline. Traditionally, experimental psychology argues with group means in experiments when it comes to testing hypotheses. However, hypotheses in experimental psychology should usually refer to *all* individuals in a population if laws of general psychology are investigated. This is the meaning of "general" and this is why *individual* causal effects and not only *average* causal effects should be of interest to general psychology. The models presented in this paper allow not only identifying the variance of individual causal effects but also modeling the interindividual differences in the individual causal effects of the treatment. If experimental psychologists start considering individual causal effects, they will tear down the barriers to differential psychology and we can hope that nature will answer questions "she will never answer until our two disciplines ask it in a single voice" (Cronbach, 1957, p. 683).

Acknowledgments

This contribution derives from my Presidential Address given at the joint conference of the *Society of Multivariate Analysis in the Behavioural and Social Sciences* and the *European Association of Methodology* in Jena, July 18–21, 2004. This address, as well as most other contributions to the conference, are also available as videos-demand at www.smabs.org. I would like to thank Donald B. Rubin, Safir Yousfi, and Benjamin Nagengast for helpful comments on a first version of this paper.

References

Bentler, P.M. (2004). *EQS 6 Structural Equations Program manual*. Encino, CA: Multivariate Software, Inc.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook on research on teaching* (pp. 171–246). Chicago, IL: Rand McNally.

Cook, T.D.C., & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Dawid, A.P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 407–448.

Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., Ostendorf, F., & Neubauer, A. (1995). Situational effects in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality, 9,* 1–23.

Dumenci, L., & Windle, M. (1996). A latent trait-state model of adolescent depression using the Center for Epidemiologic Studies-Depression Scale. *Multivariate Behavioral Research, 31,* 313–330.

Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: Model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online, 1,* 69–91.

Eid, M., & Hoffmann, L. (1998). Measuring variability and change with an item response model for polytomous variables. *Journal of Educational and Behavioral Statistics, 23,* 193–215.

Eid, M., Lischetzke, T., Trierweiler, L.I., & Nußbeck, F.W. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8,* 38–60.

Greenland, S., Robins, J.M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science, 14,* 29–46.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2004). *Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference*. Unpublished manuscript.

Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association, 81,* 945–970.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*. Chicago: SSI.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 30,* 409–426.

McArdle, J.J. (2001). A latent difference score approach to longitudinal dynamic structural analysis. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 341–380). Lincolnwood: Scientific Software International.

McArdle, J.J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development, 58,* 110–133.

Meredith, M., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107–122.

Muthén, L.K., & Muthén, B.O. (2004). *Mplus User's Guide. Third Edition*. Los Angeles, CA: Muthén & Muthén.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (reprint 1990). *Statistical Science, 5,* 465–472.

Neyman, J., Iwaszkiewicz, K., & Kolodziejczyk, S. (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society, 2,* 107–180.

Overall, J.E., & Woodward, J.A. (1977). Common misconceptions concerning the analysis of covariance. *The Journal of Multivariate Behavioral Research, 12,* 171–185.

Pearl, J. (2000). *Causality – models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement, 23,* 120–126.

Robins, J.M., & Greenland, S. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association, 95,* 477–482.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307–321.

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70,* 41–55.

Rosenbaum, P.R. (2002). *Observational studies (2nd ed.)*. New York: Springer.

Rosenbaum, P.R., & Rubin, D.B. (1984). Comment: Estimating the effects caused by treatments. *Journal of the American Statistical Association, 79,* 26–28.

Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29,* 185–203.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6,* 34–58.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services Outcome Research Methodology, 2,* 169–188.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Steyer, R. (1989). Models of classical psychometric test theory as stochastic measurement models: Representation, uniqueness, meaningfulness, identifiability, and testability. *Methodika, 3,* 25–60.

Steyer, R. (2001). Classical test theory. In C. Ragin & T. Cook (Eds.), *International encyclopedia of the social and behavioral sciences. Logic of inquiry and research design* (pp. 481–520). Oxford: Pergamon.

Steyer, R., & Eid, M. (2001). *Messen und Testen* [Measurement and testing]. Berlin: Springer-Verlag.

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2,* 21–34. (www.mpr-online.de)

Steyer, R., Ferring, D., & Schmitt, M.J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8,* 79–98.

Steyer, R., Flory, F., Klein, A., Parchev, I., Yousfi, S., Müller, M., & Kröhne, U. (2005). Testing average effects in regression models with interactions. Submitted for publication.

Steyer, R., Gabler, S., von Davier, A.A., & Nachtigall, C. (2000a). Causal regression models II: Unconfoundedness and causal unbiasedness. *Methods of Psychological Research Online, 5,* 55–86. (www.mpr-online.de)

Steyer, R., Gabler, S., von Davier, A.A., Nachtigall, C., & Buhl, T. (2000b). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online, 5,* 39–71. (www.mpr-online.de)

Steyer, R., Krambeer, S., & Hannöver, W. (2004). Modeling latent trait-change. In K. van Montfort, J.H.L. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models: Theory and applications* (pp. 337–357). Dordrecht: Kluwer.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online, 7,* 41–68. (www.mpr-online.de)

Steyer, R., Partchev, I., & Shanahan, M.J. (2000c). Modeling true intraindividual change in structural equation models: The case of poverty and children's psychosocial adjustment. In T.D. Little & K.U. Schnabel (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 109–126). Mahwah, NJ: Erlbaum.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13,* 389–408.

Tisak, J., & Tisak, M.S. (2000). Permanency and ephemerality of psychological measures with application to organizational commitment. *Psychological Methods, 5,* 175–198.

West, S.G.B. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H.T.J. Reis (Ed.), *Handbook of research methods in social and personality psychology* (pp. 40–84). New York: Cambridge University Press.

Willett, J.B., & Sayer, A.G. (1996). Cross-domain analysis of change over time: Combining growth modeling and covariance structure analysis. In G.A. Marcoulides & R.E. Schumacker (Eds.), *Advanced structural equation modeling. Issues and techniques* (pp. 125–157). Mahwah, NJ: Erlbaum.

Winship, C., & Morgan, S.L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–706.

Winship, C., & Morgan, S.L. (2000). The estimation of causal effects from observational data. *Annual Review of Sociology, 25,* 659–706.

Address for correspondence

Rolf Steyer
Institute of Psychology
University of Jena
Am Steiger 3 Haus 1
D-07743 Jena
Germany
Tel. +49 3641 945231
Fax +49 3641 945232
E-mail rolf.steyer@uni-jena.de