27

Intent to Treat and Per Protocol Modeling

Far better an approximate answer to the right question
than an exact answer to the wrong question

- JOHN TUKEY

INTRODUCTION

This chapter considers analytic strategies for dealing with people dropping out of treatment or not adhering to treatment protocols. Dropping out of treatment in a randomized trial *is* a form of treatment non-adherence because by dropping out, one prematurely ceases intervention exposure. People drop out of treatment for many reasons. Sometimes dropping out is random but other times it is related to variables in ways that

can undermine causal inference. In a multi-session intervention to reduce anxiety, people might drop out of treatment mid-treatment if the treatment is not working for them. If we do not assess and include anxiety scores at posttest for these dropouts, the mean posttest anxiety scores will be biased in ways that make the treatment seem more effective than it really is. In general, it is good practice to obtain posttreatment data on treatment dropouts. This can be challenging in some contexts but it greatly simplifies statistical analyses for effect size estimation and causal inference. In my initial discussion of dealing with treatment dropouts and non-adherence in this chapter, I assume we have access to posttreatment data for non-compliers (including dropouts). Later, I discuss how to deal with the simultaneous occurrence of dropping out of treatment and missing data.

Some people assume that ignoring treatment dropouts leads to overestimation of treatment effects. However, treatment dropouts also can lead to underestimates of treatment effects. In a multi-session intervention for anxiety, some people drop out of treatment because they think they have recovered sufficiently and perhaps that they have even been "cured." By excluding such people from posttest analyses, treatment effects will be underestimated.

It also is possible that both of the above dynamics operate, canceling each other out. This can render dropout status uncorrelated with treatment response, the result being an unbiased estimate of treatment effectiveness despite the presence of dropouts.

Dallal (2012) describes similar dropout dynamics in the context of weight loss programs. Suppose two diet-based weight loss programs are compared, one that is effective and the other that is ineffective. People on the effective diet will lose weight and stay in the study. For those on the ineffective diet, some will lose weight regardless and will stay in the study; those who fail to lose weight will be more likely to drop out of the program and lost to measurement. The result will be that the ineffective diet will look better than it is when weight loss posttest scores of treatment completers are computed. The effective diet will look less effective compared with the ineffective diet because the only people who remain in the study on the ineffective diet are those losing weight.

To deal with problems like these, many researchers use what is known as **intent-to-treat** (ITT) analysis. Data are collected on all individuals who are randomized to the different treatment conditions and all of them are included in the analysis of posttest scores (e.g., weight loss) regardless of whether they drop out of the treatment condition to which they were assigned and regardless of protocol adherence. If we measure and use posttest scores for everyone, the biasing effects discussed by Dallal will be taken into account. Because of this, many researchers argue that ITT analyses are the best way to analyze randomized trials and, indeed, insist on it. As I show in this chapter, I disagree with this orientation because it often fails to answer the question I seek to answer.

ANSWERING THE POSED QUESTION

As noted in Chapter 4, **efficacy trials** seek to determine if an intervention affects an outcome given the intervention is properly implemented and individuals receive the "full dose" of the intervention, per protocol. **Effectiveness trials**, by contrast, seek to determine if an intervention affects an outcome based on how the intervention occurs in real-world settings where patient populations and clinic variables cannot be rigorously controlled. These two types of trials address different questions. Efficacy trials ask how well a treatment works if the treatment protocol is adhered to; effectiveness trials ask what happens to people who are prescribed or offered a treatment. As Dallal (2012) states more colloquially in the context of medication trials, one question asks "what happens if people take this stuff?" and the other asks "what happens if you hand this stuff to people and tell them to take it?" These are different questions.[1] If I *require* inmates to take a four week program while in prison to help them acquire employment skills, I might ask how this affects recidivism in the population of inmates. If I *offer* inmates the option to take the four week program, I might ask how extending such an offer to members of the prison community affects recidivism. These are different questions.

I find it useful to think of treatment effectiveness heuristically as a multiplicative function of (a) how efficacious a treatment is for changing an outcome when the treatment is executed per protocol, multiplied by (b) implementation fidelity, i.e., the extent to which the treatment is, in fact, executed per protocol:

$$\text{Effectiveness} = (\text{Efficacy})\ (\text{Implementation Fidelity}) \qquad [27.1]$$

The idea is that if a treatment is not efficacious (e.g., it takes on a value of zero in Equation 27.1), then it is not going to be effective in real life settings no matter how conscientious people or clinics are about implementing it. Similarly, even if a treatment is fully efficacious when applied per protocol, its impact will be diminished if it is not implemented correctly by clinics or patients/clients, at least in terms of those facets of implementation that matter. The multiplicative function in Equation 27.1 is an oversimplification, but it conveys the gist of the general operative dynamic and it sets the stage for many of the points I make in this chapter.

One implication of Equation 27.1 is that effectiveness trials and the ITT analyses associated with them confound two questions, namely (1) is a treatment efficacious, and (2) do people/staff adhere to treatment protocols. As I discuss in Chapter 4, it is not uncommon for trialists to separate these questions and study them independently. For

[1] In some cases, the nature of an intervention is simple and full fidelity is guaranteed upon treatment administration, in which case fidelity is moot. In such cases, effectiveness equates to efficacy.

example, in early stage clinical trials, the focus on determining the efficacy of a vaccine, taken properly, at preventing COVID infections might be explored. Once the efficacy of the vaccine has been established, researchers then address implementation fidelity, i.e., what happens if the vaccine is made available to people in community settings - will people take the vaccine and take it properly per protocol (two shots spaced by a month). Effectiveness trials typically are carried out in the later stages of intervention development once we have a good handle on how to maximize efficacy and how to maximize protocol adherence from earlier trials. Having gained this knowledge, we then seek to determine if the intervention will work given the noise in real world settings

As noted, many researchers state that ITT analyses to test for intervention effectiveness are the "correct" method of analysis for RCTs and RETs. I disagree. The appropriate method of analysis depends on the question being asked (see my quote at the opening of this chapter). If you are studying treatment efficacy and the determinants or generalizability of efficacy across settings and populations, then trial design and analyses should be driven by per protocol considerations. A common strategy for per protocol analysis is to eliminate individuals who do not achieve a pre-specified standard of adherence or a prespecified "dose" of the treatment when analyzing data. An objection to this practice is that such elimination can undermine randomization to the intervention and control conditions, producing imbalance in them that biases causal inference. ITT analyses avoid this problem by including all randomly assigned individuals in the analysis. Such preservation of randomization is why many researchers insist on ITT analyses.

Having said that, ITT analyses are oblivious to post-randomization events, such as treatment discontinuation, patient use of concomitant therapies prohibited by study protocols, and treatment protocol non-adherence. Such factors can bias estimates of treatment efficacy, hence they are unwelcome in efficacy trials. The solution to the problem of compromised randomization in an efficacy trial is not to shift the analytic strategy to ITT because doing so changes the question being addressed, i.e., it shifts the question away from treatment efficacy. Such a practice lets method dictate the questions we ask rather than vice versa. The solution to violation of randomization due to dropping out is not to change the question you seek to answer; it instead is to use modern methods of per protocol analysis that address compromised randomization (Dunn et al., 2003).

WHY EFFICACY STUDIES ARE IMPORTANT

Given perfect or near-perfect levels of adherence and implementation fidelity, differences between treatment and control groups on means or proportions of an outcome/mediator

probably can be attributed to the efficacy of the intervention, absent confounds. During the COVID pandemic, early trials focused on vaccine efficacy to determine if a vaccine candidate sufficiently killed the COVID virus in humans. To evaluate vaccine efficacy at this stage, studies ensured that people in the "vaccine" treatment condition were, in fact, exposed to the vaccination per protocol. In this case, one certainly would not want to include in the treatment condition people who failed to take the vaccine or who, say, took only half the dose they were supposed to take. If you did so, you run the risk of rejecting a viable, life-saving vaccine. In the behavioral sciences, researchers conducting efficacy studies often adopt practices that are unrealistic to use in applied, real-world settings but that permit a cleaner evaluation of treatment efficacy by maximizing implementation fidelity. Researchers might pay clinics to participate in their study; they might cover clinic costs of extra staff time; they might pay patients to be in their study; they might monitor clinicians in ways that are not possible in practice to ensure that clinicians implement the intervention correctly; they might send patient texts as reminders to come to sessions or to take their medications, reminders that might not be possible in situ; and so on. To the extent that such steps are taken, the data become increasingly ill-suited to evaluating effectiveness if the goal of the effectiveness trial is to estimate treatment response in real-life settings.

ITT analyses embrace non-adherence and implementation infidelity when evaluating interventions because, as noted, they ignore post-randomization phenomena. The primary result of using ITT analyses in an efficacy trial is to add noise to the efficacy signal that a per protocol (PP) analysis provides. As Dallal (2012) notes "if there's a moderate signal coming out of the PP analysis, an ITT analysis usually sprinkles in some random non-adherent subjects to attenuate it." ITT analyses are inappropriate for efficacy trials unless one is confident that adherence/implementation fidelity are uniformly high in the trial.

One other characteristic about effectiveness trials should be noted. In real life settings, adherence and implementation fidelity are often context dependent. An ITT analysis estimates the effect of intervention versus control treatment assignment on the outcome *in the context that the trial is conducted*. A treatment can be equally efficacious across different populations or settings but may show weaker effects in one context because of context differences in implementation fidelity. By glossing over the distinction between efficacy and implementation fidelity, ITT analyses lessen our ability to gain insight into the causes of differential effects of treatments across populations and settings. Is it because the treatment is differentially efficacious for Blacks versus non-Latino Whites, because of differential protocol adherence for these two groups, or both? If two medications are compared head-to-head in a randomized trial but one has lower

adherence due to, say, an easily remediated side effect, an ITT analysis might show a beneficial effect of the less efficacious medication purely because of what are easily reversible adherence dynamics. Such results might lead clinicians to recommend the less efficacious medication despite the availability of simple remediation of the sources of infidelity for the more efficacious medication. Informed medical and policy related decision making typically requires knowledge of both efficacy *and* effectiveness.

The latter point raises ethical issues associated with focusing our science on effectiveness trials without concern for separating efficacy and implementation dynamics. If we focus exclusively on effectiveness per ITT analyses, we essentially withhold efficacy information from consumers. Suppose a patient has the choice of two options for treating a serious, life threatening disease. One treatment is nearly 100% efficacious in preventing death as long as it is adhered to per protocol. However, perfect adherence for the treatment is challenging, resulting in an overall effectiveness rate of about 60%. The other treatment is 75% efficacious but is much easier to adhere to, yielding an overall effectiveness rate near 68%. Because of the latter treatment's superior performance in effectiveness trials analyzed using ITT, clinicians might decide it is the preferred method for treatment and steer patients to it accordingly. However, if you were the patient facing death, would you not want to at least know that there is the possibility of a sure cure as long as you take extra care to adhere to the treatment protocol that many people have difficulty with? If we rely only on ITT analyses, such information is lost.

Many health care professionals as well as consumers want effect information that is not colored by adherence dynamics (Hernán & Robins, 2017). When young adults choose a contraceptive method, perfect use effectiveness of male condoms, one of the few contraceptive methods that protect against HIV, is about 98% in preventing pregnancies. However, typical use effectiveness rates are closer to 85%. Both the former per protocol efficacy information and the latter effectiveness information might be relevant to decisions people make about using condoms. Lodi et al. (2016) reanalyzed data from a well-known clinical trial that reported ITT estimates of the benefits of immediate versus delayed initiation of antiretroviral therapy (ART) for treating HIV. It turns out that about 30% of individuals assigned to the deferred initiation condition in the randomized trial violated the instructions they were given and started ART earlier than they were supposed to. Using a modern method of per protocol analysis to correct for imbalance due to such infidelity, Lodi et al. found the ITT analyses significantly underestimated the benefit of immediate ART initiation by 23%, a result they argued could have policy implications for clinical practice.

As a final example, suppose I offer a reemployment program in a community for people who have recently suffered job loss. I conduct a randomized trial in which I select

a sample of individuals who have recently lost a job. I first secure their agreement to participate in a longitudinal survey of employment. I then randomly assign half of the participants to a treatment condition in which they are invited to voluntarily participate in four weekly seminars on reemployment. The other half of the participants serve as a control group who merely complete a baseline and follow-up survey in parallel to the intervention group. The outcome variable is employment status 3 months after the intervention. Suppose I find that only about 50% of individuals in the intervention condition who agreed to complete surveys attended the seminars. In an ITT analysis, I would compare the posttest reemployment rates for all individuals who were invited to come to the seminars with rates for those in the control group, including those who never took the seminars despite being invited to take them. Suppose I find no statistically significant difference in these rates such that the program does not appear to be effective. However, suppose I also conduct a per protocol analysis and find that for those individuals who actually attended the weekly seminars, there was a notable effect on reemployment relative to the control group. What would you do in this scenario? Would you walk away from the program based on the ITT analysis or would you instead put future efforts into getting people to participate in the program per protocol? My own bias would be to do the latter and such a decision is informed by the per protocol analysis results. Ultimately, I might end up with a truly beneficial intervention that others might give up on based on a simplistic ITT analysis. How many promising programs have we walked away from by insisting on significant ITT results during intervention development rather than conducting ITT analyses after efficacy trials have run their course and efficacy has been thoroughly evaluated and maximized?

In sum, despite exhortations that ITT analyses are the "proper" way to analyze RCTs or RETs, the best way to analyze such data depends on the questions you are trying to answer. ITT analyses answer questions about treatment effectiveness but not treatment efficacy or adherence/implementation fidelity in their own right. If we seek to understand mechanisms that account for treatment efficacy, ITT analyses are not appropriate unless adherence/implementation fidelity is high. Building an intervention science based solely on effectiveness criteria is ill advised, sometimes unethical, and can lead scientists down blind alleys. When I conduct evaluation research for clients, my clients usually are interested in the effectiveness of their programs. However, clients also want to know how to improve their programs. Teasing out the contributions of treatment efficacy and implementation infidelity to effectiveness can be key to doing so. If I know a program is reasonable in terms of efficacy but weak on implementation fidelity, my recommendations to the client for program improvement might differ than if the program is weak in efficacy but reasonable in implementation fidelity. Also important is

identifying program active ingredients that drive efficacy as well as active ingredients that drive implementation fidelity. RETs using strong ITT *and* per protocol analytic strategies can help accomplish these goals.

IMPLEMENTATION TRIALS

There is a growing literature that focuses on the use of RCTs and RETs to understand implementation fidelity. This literature adopts a broad conceptualization of infidelity to include more than just patient/client adherence to a treatment protocol. Patient/client adherence is but one part of a broader system of implementation at multiple levels of an organization or system. For example, to maximize implementation fidelity for a medication, clinics must ensure that doctors are aware of the medication; doctors must make correct diagnoses for purposes of prescribing it; the medication must be available at relevant dispensaries; patients must acquire the medication; once acquired, patients must take the medication as prescribed; and so on. An implementation randomized trial is one in which some facet of implementation is explicitly manipulated with individuals being randomized to the different conditions for purposes of making causal statements about effects of implementation facets on substantive outcomes. One might seek to increase session attendance of a multi-session program for depression by presenting information to people at intake about multiple reasons or "positive nudges" to attend every session. This might be compared to people in a control condition who do not receive such information. The outcome is the number of sessions attended by the patient, a facet of information fidelity. As another example, one might examine the effect of clinicians using a checklist for key questions to ask patients as part of an anxiety therapy. The treatment and control conditions are the provision of the checklist to physicians versus no such provision. The outcome is patient anxiety at the conclusion of therapy. Note that the analyses for this implementation trial also might properly use per protocol rather than ITT analyses if one seeks to determine the *efficacy* of using a checklist, i.e., only include in the analysis physicians in the treatment condition who were actually given the checklist.

In some effectiveness trials, we design the study to address both efficacy and implementation fidelity. These designs are called **hybrid designs** and would use both per protocol *and* ITT analyses. Mediators might include some that are key to understanding adherence/implementation and others that are key to understanding efficacy. Such designs can be complex.

ESTIMANDS

A relatively recent concept introduced into the literature on RCT design is that of an **estimand**, which refers to a specific statistical parameter that one seeks to estimate in the context of a randomized trial. One type of estimand is for the *effectiveness* of a treatment, which is the mean difference between individuals randomly assigned to the treatment and control groups independent of adherence. Another type of estimand is for the *efficacy* of a treatment, which maps onto a per protocol analysis of mean differences between treatment and control groups. In 2020, the International Council for Harmonisation published an addendum on the use of estimands in randomized trials (ICH, 2020) encouraging researchers to be explicit about the estimand one seeks to evaluate and why. According to the document, this requires, among other things, being explicit about the population of interest, the treatment conditions to be compared, the outcome measure, the parameter that will be used to represent the estimand, the statistical method used to estimate that parameter, and how nuisance factors (e.g., treatment discontinuation, non-adherence, missing data) are addressed. In essence, ICH asks researchers to be clear about the questions they seek to answer, to justify the meaningfulness of the questions, and then to ensure that trial design and data analysis maps onto the questions. The concept of an estimand in RCTs has led to an articulation of many different questions beyond those associated with effectiveness and efficacy (Greifer & Stuart, 2021).

A useful framework for specifying estimands in clinical trials is one that evolves from what is known as **complier average causal effect** (CACE) analysis. The framework takes into account the concepts of efficacy, effectiveness and adherence. It begins by making distinctions between patient/client compliers and non-compliers in a study. Compliers are defined as people who faithfully adhere to or comply with the prescribed treatment protocol; non-compliers are those who do not do so. Suppose I dichotomize compliance and score people as 1 = compliers and 0 = non-compliers. People with a score of 1 adhere to the protocol enough so that we believe the treatment can have its effect. People with a score of 0 fail to comply with the treatment protocol to the point that we believe the treatment will fail to have an impact. In a 12 session treatment for depression, a non-complier might be someone who, say, attends 6 or fewer sessions. This category represents people who are functionally treatment dropouts. In a medication or biologic trial, complier status might be defined as the extent to which people have taken the required dosages of the medication with an *a priori* threshold value specified to define sufficient dosage. Let T represent if someone has been randomly assigned to the treatment condition (a score of 1) or the control condition (a score of 0) and C represent his or her complier status. The CACE framework works with outcome means (or proportions) of four groups of individuals, per Table 27.1:

Table 27.1: The CACE Framework

	Complier	Non-Complier	Marginal Mean
Treatment	μ|T=1, C=1 [*a*]	μ|T=1, C=0 [*c*]	μ|T=1 [*e*]
Control	μ|T=0, C=1 [*b*]	μ|T=0, C=0 [*d*]	μ|T=0 [*f*]

The symbol | is read as "given that" to indicate a conditional mean and μ is the population mean (or covariate adjusted mean) of the outcome of interest. The expression μ|T=1 refers to the mean outcome for those randomly assigned to the treatment condition ignoring complier status and μ|T=0 is the mean outcome for those randomly assigned to the control condition ignoring complier status, i.e., they are marginal means. I add labels *a* to *f* to the cells of the table to simplify later explication and notation.

There is a subtle distinction in the CACE framework that is important to keep in mind. The variable C refers to complier status when an individual is exposed to treatment. As such, cell *b* in Table 27.1 refers to control individuals who would be treatment compliers *if they had been assigned to the treatment condition*. Similarly, cell *d* refers to control individuals who would not be treatment compliers *if they had been assigned to the treatment condition*. As I will show later, methodologists have evolved creative ways of estimating the means (or proportions) in cells *b* and *d* for purposes of defining estimands that are rarely used in traditional RCTs.

In Chapter 8, I introduced the concept of an average causal effect (ACE), which is the mean posttest score for individuals in the intervention condition minus the mean posttest score for people in the control condition. In the CACE framework, classic ITT analysis calls an ACE and average treatment effect and defines an ATE estimand for purposes of program evaluation as cell *e* minus cell *f* in Table 27.1, or

$$ATE_{ITT} = e - f = (\mu|T=1) - (\mu|T=0)$$

Thus, ATE_{ITT} is the mean difference between the treatment and control conditions ignoring peoples' complier status. Traditional per protocol analysis, by contrast, defines an estimand for the average treatment effect as cell *a* minus cell *f*, or

$$ATE_{PP} = a - f = (\mu|T=1, C=1) - (\mu|T=0)$$

ATE_{PP} is the mean difference between treatment compliers and everyone in the control condition irrespective of their complier status.

The CACE framework also suggests an intriguing alternative possibility for an

estimand tied to per protocol effects, namely the comparison of those people in the treatment condition who complied with the treatment protocol (cell *a*) with those in the control condition *who would have complied with the treatment protocol if they had been assigned to the treatment condition*, or

$ATE_{CACE} = a - b = (\mu \mid T=1, C=1) - (\mu \mid T=0, C=1)$

I explore the CACE approach later in this chapter. In some literatures, the CACE effect sometimes is referred to as the **local average treatment effect** (LATE).

A final estimand sometimes used in randomized trials is known as the **as treated average treatment effect**. It compares treatment compliers (cell *a*) with all other groups combined (cells *b*, *c* and *d*), treating the latter group as individuals who have not been "treated" either because they were not offered the treatment (cells *b* and *d*) or because they failed to sufficiently adhere to the treatment protocol (cell *c*):

$ATE_{AS\ TREATED} = (\mu|\text{cell } a) - = (\mu|\text{cell } b, c \text{ or } d)$

No one of these estimands (ATE_{ITT}, ATE_{IPP}, ATE_{CACE}, $ATE_{AS\ TREATED}$) is "correct." Rather, when computed, they represent different answers to different questions. The point of the International Council for Harmonisation is that different estimands require different approaches to study design and analysis and that researchers need to be explicit about the estimands they seek to document. The concept of estimands that has recently been introduced into the randomized trial literature is an important one and will have a central role in trial design and analysis in the future.

TREATMENT CONFOUNDS

Treatment dropouts and protocol non-adherence are often not problematic for obtaining unbiased estimates of ATE_{ITT} because randomization to condition is impervious to them and the focus of the investigator is on effectiveness not efficacy. If one fails to obtain posttest measures for treatment dropouts, then ITT analyses can suffer because of the loss of statistical power and because one sometimes must make untestable assumptions about the missing data (see Chapter 26 and my discussion of ITT analyses below). For traditional per protocol analyses, treatment dropouts and non-compliance can lead to estimation problems for ATE_{PP} not only because of loss of statistical power when non-compliers are eliminated from the data but also because randomization between treatment and control conditions can be compromised. For example, if females are more likely to drop out of the intervention arm of a study so as to create imbalance for biological sex between the intervention and control conditions, then ATE_{PP} estimates can be biased if

sex impacts the outcome.

In Chapter 4, I discussed the concept of imbalance between treatment and control conditions when imbalance arises by chance during random assignment. Such imbalance usually is not consequential but it is an annoyance. By contrast, imbalance due to eliminating non-compliers for purposes of per protocol analysis can be insidious. To be sure, if such imbalance occurs on variables irrelevant to the target outcome, then the imbalance will not bias efficacy estimates. However, if the imbalance occurs on treatment confounds that influence the target outcome, then inferences about treatment efficacy can be distorted.

Figure 27.1a presents an influence diagram illustrating a case where a treatment has no impact on an outcome (depression) but where biological sex determines self-selection into the treatment condition in a per protocol analysis, with females being more likely to drop out of or be dropped from the treatment condition due to adherence violations than the control condition (path *a*). This dynamic creates an imbalance of biological sex for the intervention and control group with the intervention group having more males relative to the control group. (This imbalance, of course, would be detected by testing for sex differences as a function of treatment condition in the per protocol sample). If males have lower depression than females in general (e.g., at baseline), then the imbalance can bias results towards overestimating the efficacy of the intervention because of the preponderance of males in the intervention condition. The amount of imbalance that occurs in the intervention versus control conditions is essentially a function of the strength of path *a*.

The consequences of imbalance for making faulty inferences about the effects of a treatment also depends on the strength of path *b* in Figure 27.1a. If this path is weak, the imbalance may not be consequential. If it is strong, then the estimated intervention efficacy can be distorted given sufficient imbalance reflected in path *a*. I conducted a small scale simulation in which I varied the strength of paths *a* and *b* to evaluate the effect of such variation on Type I errors for treatment efficacy. I used an alpha level of 0.05 and a two tailed test. I set up a scenario where the true mean difference between the treatment and control groups on depression is zero in the population. As such, the Type I error rate for the treatment versus control mean difference should be 0.05, the alpha level.

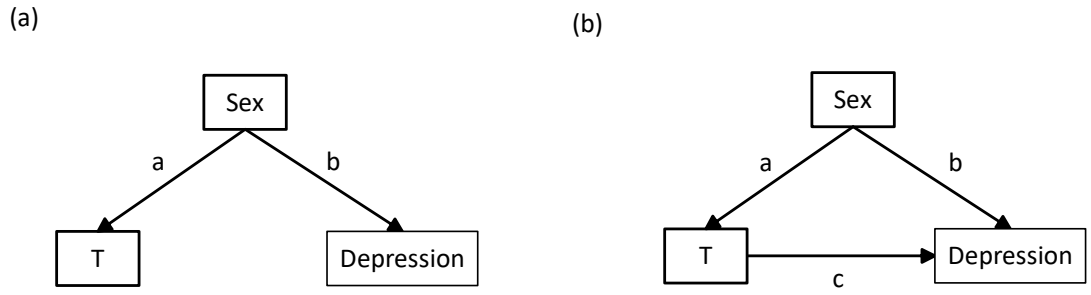

FIGURE 27.1. Effect of treatment dropout on treatment effect estimation

For path *a*, a zero effect of sex on self-selection into the treatment condition occurs when the probability of being in the intervention group given one is male, P(T=1|sex = male), equals the probability of being in the intervention group given one is female, P(T=1| sex = female). In the simulation, I set each of these probabilities to 0.50. In addition to this scenario, I also evaluated a 60-40 conditional probability difference biased towards males being more prevalent in the intervention group as well as a 70-30 conditional probability difference. These reflect stronger effects for path *a*. I also varied the strength of path *b* in terms of the population Cohen's *d* for the effect of sex on depression with males having lower depression than females. I evaluated *d* values of 0.00, -0.30, -0.50, and -0.80 based on subtracting the female mean from the male mean. Finally, I varied the total sample size after treatment non-compliant participants were eliminated, using sample sizes of 100, 200, and 300. Table 27.2 presents the Type I error rates for falsely concluding an intervention effect exists in each simulation condition.

Table 27.2: Results of Simulation for Imbalance and Treatment Dropout

	Effect of Sex on Posttest Depression			
	d = 0.00	d = 0.20	d = 0.50	d = 0.80
50-50 diff				
N = 100	0.052	0.058	0.057	0.057
N = 200	0.052	0.053	0.054	0.046
N = 300	0.048	0.051	0.051	0.055

60-40 diff				
N = 100	0.060	0.061	0.086	0.118
N = 200	0.043	0.051	0.105	0.168
N = 300	0.046	0.062	0.132	0.245
70-30 diff				
N = 100	0.051	0.085	0.177	0.336
N = 200	0.051	0.083	0.282	0.575
N = 300	0.053	0.110	0.393	0.736

When either path *a* or path *b* is zero or weak, the imbalance has minimal effect on the Type I error rate. The error rate inflates non-trivially when both paths *a* and *b* are sizeable. Clearly, RCTs can withstand some confound induced imbalance in per protocol analyses as long as it is not too strong and the confounding variables exert trivial outcome impact.

Figure 27.1b illustrates a case where the treatment has an effect on the outcome (path *c*) but the efficacy estimate will again be biased depending on the strengths of paths *a* and *b* (see my discussion of confounding in Chapter 2). Thus, meaningful confounding is of concern not only because of its impact on Type I error rates but also because of its potential to distort effect sizes. Note that the confounding can be in either direction, namely the effect size can be biased upward or downward.

Another form of confounding is shown in Figure 27.2. Here, the confounder causes the outcome but it does not formally cause imbalance (see Chapter 4 for examples). To be sure, it is correlated with the treatment condition a person is in, but it is not the cause of it. In this case, the correlated confounder represents a proxy for the causal confounder and it needs to be controlled as well, depending on the strengths of paths a and b.

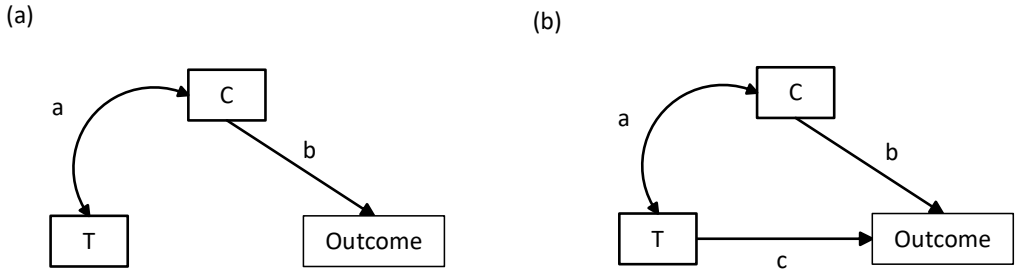

FIGURE 27.2. Another form of imbalance

In sum, defining a per protocol sample in the traditional way of eliminating treatment non-compliers can introduce imbalance into the randomization process. This imbalance will be of little consequence if it is trivial in magnitude or if the variable on which the imbalance occurs has no or only a trivial causal impact on the outcome. If the imbalance is reasonably strong and if the variable on which the imbalance occurs has a non-trivial impact on the outcome, then corrective steps are required. The more modern methods of per protocol analysis that I describe below seek to take such corrective actions.

NUMERICAL EXAMPLE

Throughout the remainder of this chapter, I illustrate ways of dealing with treatment dropouts and noncompliance using data from the JOBS studies published by Price, van Ryn, and Vinokur (1992) and Vinokur, Price and Shul (1995). These studies sought to determine the effect of a job search intervention on depression and reemployment of unemployed workers. The program consisted of five 4 hour seminars that taught job search strategies coupled with the provision of a job search pamphlet. The control group received only the pamphlet. Potential participants were sent a baseline questionnaire and those who responded were randomized to one of the two conditions. In my analyses, depression measured at 6 months post intervention was the primary outcome. To parallel analyses reported by Price et al. (1992) and numerous re-analyses of the data (e.g., Little & Yau, 1998; Yau & Little, 2001; Jo 2002), I focus on a subgroup of individuals who were a priori classified as being at high risk for depression (N = 502). Depression was measured using the Hopkins Symptom Checklist, with scores ranging from -3 to +3 and higher scores indicating greater levels of depression. Possible covariates/predictors I make use of include baseline measures of age, years of education completed, marital status (0 = married, 1 = otherwise), race (1 = Black, 0 = otherwise), a measure of economic hardship, a baseline measure of depression (scored on a different metric from that of posttest depression), motivation to engage in job searches, and a measure of assertiveness. Non-compliance occurred in the study by virtue of a substantial number of no-shows (about 45%) to the seminar sessions in the intervention condition.

To keep matters simple, I explicate statistical methods for addressing efficacy and effectiveness estimands by focusing on outcome only analyses. After doing so, I discuss how they can be extended to randomized explanatory trials.

EFFICACY (PER PROTOCOL) FOCUSED ANALYSES

In this section, I consider approaches to conducting per protocol like analyses. These

approaches are distinct from what is sometimes called **naïve per protocol analysis** which consists of dropping non-compliant individuals from the analysis and then analyzing data using traditional methods with no correction for imbalance or confounding that might result from participant deletion. The approaches I consider are direct covariate methods, inverse probability treatment weighting methods, instrumental variable/CACE methods, G and targeted maximum likelihood estimation, and formal dosage analysis. Keep in mind that each method is designed to estimate treatment efficacy, not effectiveness. Also, the "naïve" approach to per protocol analysis is not necessarily "naïve" if there is no consequential imbalance due to confounds in the per protocol population. In my opinion, it is somewhat unfair to refer to the approach as naïve because there are scenarios where this is not the case and the method works just fine.

All of the strategies I discuss make use of measures of confounders that are sources of imbalance. Unmeasured confounders are the bane of clinical trials which is why you need to heed the messages from Chapter 2; anticipate as best you can when planning a study the likely major sources of imbalance or confounding, in this case due to dropping out of treatment, and then measure them to the best of your ability.

The Direct Covariate Approach

One strategy for addressing imbalance resulting from elimination of treatment non-compliers is to include a measure of the confound associated with the imbalance as a covariate in one's model. This is probably the most straightforward approach to dealing with imbalance. I refer to it as the **direct covariate approach**. The strategy is grounded in analysis of covariance (ANCOVA) logic for RCTs and most of the statistical assumptions of ANCOVA apply to it.

The direct covariate approach uses the per protocol estimand, ATT_{PP}. Interest is in comparing a population of treatment compliers with a population of individuals who have not been exposed to the treatment irrespective of what people in the latter population would do relative to adherence if they were exposed to the intervention. This is not a conceptually unreasonable estimand, although as elaborated below in my discussion of the CACE approach, some methodologists argue otherwise.

Choosing Covariates

The choice of covariates in per protocol analyses is important and must be made carefully. One should be cautious about including long laundry lists of covariates that are not theoretically screened, in part, because estimation efficiency can deteriorate with increasing numbers of covariates and because some covariates can amplify rather than reduce coefficient bias (see my discussion of "good" and "bad" covariates in Chapter 2).

For example, research suggests that including what is known as an instrumental variable as a covariate and treating it as a confound can create coefficient bias (Ding, Vanderweele & Robins, 2017; Myers et al., 2011a, b; Pearl, 2011) because such variables take on suppressor characteristics. Typically, we want to select covariates that have non-trivial imbalance *and* that have non-trivial causal effects on the outcome.

For the JOBS study, Table 27.3 presents a comparison of those in the treatment condition with those in the control condition on baseline variables after eliminating individuals in the treatment condition who were non-compliers. If I choose covariates for per protocol analyses based on those covariates that show statistically significant differences between the treatment compliers and controls, I would select baseline depression, age, motivation, and education. However, both age and motivation were correlated near zero with posttest depression (r = 0.01 and -0.03, respectively), suggesting caution about using these variables as covariates because of potential bias amplification.

Table 27.3: Treatment Condition Comparisons on Possible Covariates

	All Randomized to Control	Only Treatment Compliers
Baseline depression	2.49	2.42*
Age	36.17	39.68*
Motivation	5.32	5.50*
Education	13.34	13.79*
Assertiveness	3.03	2.96
Marital status	0.58	0.65
Economic hardship	3.47	3.60
Non-white	0.18	0.15

Note: * treatment compliers different from controls, $p < 0.05$

Some methodologists question data-driven covariate selection (Austin & Stuart, 2015). The idea is that significance tests are influenced by sample size and that studies with small sample size may mistakenly suggest omitting a covariate that is bias reducing due to lack of statistical power. In addition, small effects for many covariates can cumulate to create a non-trivial collective effect. Sometimes empirics and theory or past research conflict in what they suggest one do. For example, although the proportion of non-Whites for treatment compliers and controls were fairly close in the JOBS study (0.15 versus 0.18), there is compelling theory and research to suggest that adherence

differences may be larger than what was observed for these groups. Reputable methodologists differ in their recommendations about how to handle such situations (see Myers et al., 2011a, b and Pearl, 2011). One strategy is to perform sensitivity analyses with and without the controversial covariates to determine if doing so affects conclusions. Hopefully, the results of the analyses will converge but if not, one moves forward one way or the other but with caution.

Some methodologists also caution against covariate control of post-randomization variables and state you should restrict the pool of covariates to baseline variables. Hesitations about such controls derive from the potential of biasing effects of collider covariates (see Chapter 2 and Chapter 11). However it is difficult to anticipate the biases that such variables will produce in complex models. In my view, there is nothing inherently wrong with using post-randomization controls as long as they are substantively appropriate and you empirically explore collider bias issues for covariates that meet collider criteria, per Chapter 11. For example, if I want to know the overall per protocol effect of a treatment on an outcome, it makes little sense to control a post-randomization mediator because the mediator carries with it a mechanism by which the treatment affects the outcome. Statistically holding constant such a post randomization mediator would be shooting myself in the foot by controlling the very mechanism through which the treatment affects the outcome. However, if I seek to answer the question of whether the treatment has an effect on the outcome independent of a post-randomization mediator, then it makes sense to statistically control that mediator.

My own preference when making covariate choices is to prioritize theory and past research when deciding what covariates to include in a per protocol analysis but I also examine the data itself to help me make coherent choices. Again, the idea is to focus on covariates that (a) are non-trivially imbalanced when comparing the treatment and control conditions after eliminating treatment non-compliers, and that also (b) are causally related to the particular dependent variable in the analysis.

Per Protocol Analysis of the JOBS Intervention

For the per protocol analysis of the JOBS intervention, I used as covariates the baseline measures of depression, education, and economic hardship, all of which were associated with posttest depression. I also included non-white status to illustrate programming principles for binary covariates. The Mplus syntax for the per protocol analysis is shown in Table 27.4. Much of the syntax is already familiar to you based on prior chapters. I highlight lines that may be new or that are noteworthy.

Table 27.4: Mplus Syntax for Direct Covariate Approach

```
1. TITLE: EXAMPLE DIRECT COVARIATE PER PROTOCOL ;
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. DEFINE:
4. CENTER depbase educ econ nonwhite (GRANDMEAN) ;
5. VARIABLE: NAMES ARE depress risk Tx depbase age motivate educ assert
6. single econ nonwhite x10 c1 c2 tncomply depchang prob;
7. USEVARIABLES ARE depress Tx depbase educ econ nonwhite ;
8. USEOBSERVATIONS tncomply EQ 0 ;
9. ANALYSIS: ESTIMATOR = MLR ;
10. MODEL:
11. depress ON Tx depbase educ econ nonwhite ;
12. OUTPUT: Samp StdYX Mod(All 4) Residual Cinterval Tech4 ;
```

Line 3 invokes the `DEFINE` command to transform the covariates. Line 4 mean centers each of the covariates and stores the results back into the same variable. Line 8 tells Mplus to do the analysis for a subset of the observations, namely those who have a score of 0 on the variable called `tncomply`. This is a variable I created in the input data to identify the per protocol sample; individuals with a score of 1 were randomized to the treatment condition and failed to comply with the study protocol; individuals with a score of 0 were either in the control condition or were in the treatment condition and complied with the study protocol. Note that this variable is not listed on the `USEVARIABLES` command because technically, it is not part of the model; it is only used as a screener.

The model is just identified so model fit indices are moot. The total sample size after non-complier elimination was 350. The core Mplus output is as follows:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.    Est./S.E.  P-Value

 DEPRESS  ON
    TX              -0.193      0.080     -2.420      0.016
    DEPBASE         -0.844      0.135     -6.241      0.000
    EDUC            -0.026      0.019     -1.400      0.162
    ECON             0.115      0.047      2.458      0.014
    NONWHITE        -0.040      0.109     -0.367      0.714

 Intercepts
    DEPRESS         -0.350      0.059     -5.917      0.000
```

The per protocol effect of the treatment on depression was estimated to lower depression by -0.193 units (MOE = ± 0.16, critical ratio (CR) = 2.42, $p < 0.02$). The intercept is the

mean depression value when all predictors equal 0. Because I mean centered all predictors except the treatment condition, the intercept, - 0.350, is the mean depression for the control group when each covariate is held constant at its "typical" value. The mean depression for the treatment group was -0.350 + (-0.193) = -0.543. You can isolate the predicted means and margins of error for any predictor profile of interest using the profile analysis methods described throughout prior chapters, beginning in Chapter 6. Parenthetically, the unadjusted ITT analysis yielded an effect that equaled -.074, which was statistically non-significant (MOE = ±0.15, CR = 0.99, $p < 0.32$). If people participate in the program and adhere to its protocol, the program does indeed tend to lower depression. However, if you just offer the program to people in general, there is not good evidence that it will make a difference vis-à-vis traditional ITT analysis. As a policy maker, how would you act on this information?

A variant of the direct covariate approach that you may encounter in the randomized trial literature is known as **principal stratification**, variants of which I introduced in Chapter 4. When executed as part of study design and before data are collected, stratification requires that you block or create a factor for an anticipated source of imbalance and then randomly assign individuals to the treatment and control conditions in equal proportions for groups defined by that factor, thereby removing imbalance. For per protocol analyses, some researchers perform the blocking after data have been collected and then use covariate-like analytic methods to render any residual imbalance moot (Rosenberger & Lachin, 2015). Continuous confounds usually are divided into strata (e.g., if the prognosticator is age, three age groups might be defined), which yields a degree of analytic crudeness because we essentially turn a many-valued continuous construct into a crude few-valued one, thereby throwing away potentially useful information (see Chapter 3). As well, blocking on many variables can be challenging in some contexts. Because I believe there almost always are better approaches for per protocol analysis than such stratification, I do not delve into this strategy here (see VanderWeele, 2011, 2012). However, I recognize that as an *a priori* sampling strategy and for purposes of addressing moderator questions, stratification has positive features. It is just not very strong as a way of addressing per protocol questions in the direct covariate approach.

A weakness of the direct covariate approach is that one must anticipate potential sources of imbalance due to protocol infidelity when designing one's study and then obtain measures of those sources/variables so that one can control for them during the modeling enterprise. Good trial design requires giving considerable thought to the matter. Another weakness of the direct covariate approach is one I mentioned for principal stratification, namely that when adherence/compliance is continuous or many-valued,

researchers often reduce it to a dichotomy of "compliers" versus "non-compliers" to define the per protocol sample, thereby throwing away potentially useful information. Nevertheless, the approach is often superior to naïve per protocol analyses.

Inverse Probability Treatment Weighting

Inverse probability treatment weighting (IPTW) is a covariate based strategy that uses propensity scores. A **propensity score** is a person's probability of being in the intervention versus control condition given one's scores on a set of (usually baseline) covariates. It reflects one's propensity to be in one treatment condition or the other based on characteristics of the individual and the context. For the JOBS study, it is the predicted probability *using the per protocol subsample* for being in the intervention group as derived from a logistic (or probit) regression that regresses the binary per protocol treatment dummy variable (1 = in the intervention condition, 0 = in the control condition) onto the covariates that you desire to control imbalance on. From Table 27.2, the possible predictors in the logit model are baseline depression, age, motivation, education, assertiveness, marital status, economic hardship and ethnicity.[2] Austin and Stuart (2015) suggest not selecting just any variable that shows imbalance for IPTW analysis but instead argue that variable selection should be judicious following the same logic I described for the direct covariate approach; you select predictors that first and foremost have non-trivial imbalance *and* that non-trivially impact the outcome. One then uses the individual-based predicted probabilities that results from the logistic analysis as a covariate to statistically adjust for the imbalance in the per protocol analysis. The adjustment strategy can use either formal matching between the intervention and control groups on the propensity scores, stratification on the propensity scores, direct covariate adjustment using the propensity scores, or inverse probability of treatment weighting (Austin, 2013). The assumption for each of these methods is that individuals with comparable propensity scores have similar baseline profiles on the relevant covariates. The idea is that one can work with propensity scores directly rather than the individual covariates to reduce imbalance (Rosenbaum & Rubin, 1983). It turns out that this is indeed the case but under somewhat restrictive conditions (see King & Nielsen, 2019). The advantage of using propensity scores is the reduction of the number of covariate dimensions to deal with during the formal modeling of the effect of the treatment on an outcome/mediator because the covariates have been reduced to a one-dimensional score. From among these different methods, my focus here is on inverse propensity treatment

[2] In principle, any appropriate binary regression method can be used to derive the predicted probabilities, including methods from machine learning (McCaffrey et al., 2004, 2013). Chapter 15 presents a Bayesian Additive Regression Tree method that can handle linear and non-linear functions and is quite flexible.

weighting. It tends to perform best relative to the other propensity based methods, but there are exceptions.

IPTW adjusts for measured confounder imbalance using statistical weighting. To provide some intuition for this method, consider the traditional formula for the sample mean, which is the sum of the scores divided by the sample size. A more general formula for the mean can be written that incorporates weights. It is:

$$M_X = \Sigma w_i X_i / \Sigma w_i$$

where M_X is the estimate of the population mean for variable X, w is a weight assigned to each individual *i*, and the summation occurs across individuals. Each individual's score on X is multiplied by his or her "weight value" and the sum of these weighted scores is divided by the sum of the weights. Suppose I assign everyone a weight of 1. The sum of the weights (the denominator) will equal N (the sample size) and the sum of the weighted X scores will be the sum of the X scores. The result is the traditional sample mean. In this sense, the traditional formula for a sample mean is a special case of the above formulation, namely the case where all the individual weights are 1.0.

The IPTW approach was first proposed by Robins (1986). Conventional IPTW weights for each individual in the treatment condition are set equal to 1.0 divided by the propensity score (PS) for that individual, i.e., one divided by the probability of being in the intervention condition based on the logistic regression. For individuals in the control group, the weight equals 1 divided by (1-PS) or one divided by one minus the probability of being in the intervention condition. These weights are then applied to the sample data to yield an estimate of the average treatment effect, per protocol (see Robins, 1986). You may encounter other weighting schemes in the IPTW approach Greifer & Stuart, 2021) but the above is considered standard for traditional per protocol populations/analyses.

In some scenarios, a few individuals in the intervention group may have propensity scores near 0 or a few individuals in the control group may have propensity scores near 1. Such cases can make the IPTW analysis unstable. A stabilized inverse probability treatment weight is sometimes used to deal with this problem that further multiplies the IPTW by the probability of receiving the actual treatment received, i.e., by the overall proportion of people in the intervention group when calculating the weight for those in the intervention group and the overall proportion of people in the control group when calculating the weight for those in the control group (Austin & Stuart, 2015). An alternative strategy is to use trimmed or truncated weights in which a threshold is set at the two ends of the distribution such that weights that exceed the upper threshold or are lower than the lower threshold are set to the value of the upper or lower threshold, respectively (Cole & Hernán, 2008; Lee, Lessler & Stuart, 2011). The threshold typically

is based on weight quantiles, such as the 1st and 99th quantiles or quantiles that are empirically determined (Cole & Hernán, 2008).[3] A third strategy is to coarsen the propensity scores into a few categories and assign individuals who fall into a given category the weighted mean propensity score for that category, or some variant thereof. With this approach, if you coarsen too much, it reduces estimation efficiency and can create bias by not adequately resolving imbalance. On the other hand, moderate coarsening can reduce the impact of extreme weights and introduce weight stability. At present, we do not have good guidance on strategies for weight coarsening (Kang & Schafer, 2007; Cole & Hernán, 2008). Below, I show you how to use stabilized inverse probability treatment weights in Mplus.

The IPTW approach assumes (a) that no unmeasured confounders have been excluded from the logistic model used to generate the PS scores (often called the **exchangeability assumption** or **ignorability**) and (b) that the logistic model is correctly specified. It also makes what is known as a **positivity assumption** or **common support** which is that there are both intervention exposed and intervention unexposed individuals at every level of the confounders. In practice, it likely is impossible to include all relevant confounders when applying the IPTW approach, suggesting that the exchangeability assumption is rarely met. The methods I described in Chapter 2 on covariate choice can be used to help prioritize covariates to include in IPTW. Some researchers try to compensate for exchangeability violations by including a large laundry list of covariates in the hopes that relevant ones are included. My discussion of good controls versus bad controls vis-à-vis atheoretical partialling in Chapter 2 questions this practice.

When using the IPTW method, it is recommended that one perform checks on the weighted data to ensure reasonable covariate balance has been achieved. I describe methods for doing so below. When evaluating group differences on balance, some researchers rely on significance tests (e.g., Rosenbaum & Rubin, 1984); others argue that one should eschew significance tests in favor of effect size analysis (Imai, King & Stuart, 2008). One rationale for not relying on significance tests is that they may have low statistical power when working with small N.

Both the direct covariate approach and the IPTW approach assume their respective models are correctly specified. Some researchers recommend combining the two approaches to form what they call **doubly robust methods** for per protocol analysis (Robins et al., 2007; Funk et al., 2011; Wang, Ogburn & Rosenblum, 2019). I describe

[3] If you also need to use sampling weights in a complex design, you can multiply the sampling weight and the propensity score weight and use the product weight in the analysis; see DuGoff et al., (2014).

these methods below.

Analysis of the JOBS Intervention

I applied IPTW analyses to the JOBS data for a per protocol analysis using the same data and covariates as that for the direct covariate approach. I used stabilized PS-based weights for estimating the per protocol ATE. In the sections that follow, I first illustrate preliminary analyses that evaluate the achieved covariate balance and the weight distributions. I then estimate the ATE_{PP}.

Preliminary Analyses of the JOBS Intervention. The first set of analyses determined how well the weights minimized imbalance between the intervention and control groups. In Table 27.3, I found that the unweighted index of education (years of education) was statistically significantly larger in the treatment complier group than in the control group, 13.79 versus 13.34, a difference of 0.45. The pooled within group standard deviation for the two groups was approximately 2.0, yielding a Cohen's *d* of 0.23. Table 27.5 presents the Mplus syntax that compares the treatment complier and control group means and variances on education but using the weighted data in a multi-group model in Mplus (see Chapter 20).

Table 27.5: Mplus Syntax for Weighted Data for Covariate Balance

```
1. TITLE: COVARIATE BALANCE ;
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. VARIABLE:
4. NAMES ARE depress risk Tx depbase age motivate
5. educ assert single econ nonwhite x10 c1 c2 tncomply
6. depchang prob wght stabwght ;
7. USEVARIABLES ARE educ stabwght  ;
8. USEOBSERVATIONS tncomply EQ 0 ;
9. WEIGHT = stabwght;
10. GROUPING is Tx (0 = control, 1 = treat) ;
11. ANALYSIS:  ESTIMATOR = MLR ;
12. MODEL:
13. [educ] ; educ ;
14. MODEL CONTROL:
15. [educ] (meanc) ; educ (varc) ;
16. MODEL TREAT:
17. [educ] (meant) ; educ (vart) ;
18. MODEL CONSTRAINT:
19. NEW (MEANDIFF VARRATIO) ;
20. MEANDIFF = meant - meanc ;
21. VARRATIO = vart/varc ;
22. OUTPUT: Samp StdYX Residual Cinterval Tech4 ;
```

The variable `prob` in the `NAMES ARE` statement is the predicted probability of being in the treatment group that I calculated using a logistic regression before inputting the data into Mplus. I also calculated the IPTW weights and stored them in the variable `wght` and the stabilized version of the weights in the variable `stabwght` using the first method of stabilization described earlier. Line 8 eliminates treatment non-compliers from the analysis. Line 9 tells Mplus to do a weighted analysis using `stabwght` as the weight variable. Line 10 identifies the two comparison groups, the treatment and control conditions. Line 13 tells Mplus to estimate the weighted means and variances for `educ` for all groups and Lines 14 to 17 also make this request but add labels to each parameter using parentheses. These labels are referenced later in the `MODEL CONSTRAINT` command. Lines 18 and 19 ask Mplus to conduct two contrasts. Line 21 tests the weighted mean difference between the treatment and control groups and Line 22 calculates a variance ratio of the variances of the two groups. If the group distributions are perfectly balanced for `educ`, this ratio should equal 1.0. Here is the core output for this just identified model:

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
Group CONTROL				
Means				
EDUC	13.562	0.159	85.236	0.000
Variances				
EDUC	3.994	0.360	11.089	0.000
Group TREAT				
Means				
EDUC	13.579	0.157	86.446	0.000
Variances				
EDUC	4.359	0.415	10.516	0.000
New/Additional Parameters				
MEANDIFF	0.016	0.224	0.074	0.941
VARRATIO	1.091	0.143	7.630	0.000

The weighted mean years of education is 13.562 for the control group and for the treatment group it is 13.579, a difference of 0.016. From the `New/Additional Parameters` section of the output, the margin of error for the difference is 0.45, the critical ratio (CR) is 0.07, and the p value is < 0.95. The pooled within group standard deviation is the square root of the weighted average of the two variances (3.99 and 4.36)

which is approximately 2.05. This yields a Cohen's *d* of 0.016/2.05 = 0.01. The weighted group mean difference on this covariate seem trivial and affirms the use of the IPTW weights. McCaffrey et al. (2013) suggest that standardized mean differences less than 0.20 for IPTW propensity modeling is reasonable, but this is somewhat arbitrary.

The variance ratio of the two weighted variances, 4.36 and 3.99, was 1.09. From the confidence interval section of the output, the 95% CI was 0.81 to 1.37.[4] The variances in the two groups seem reasonably close, as are the standard deviations (2.09 and 2.00). I obtained comparable results for the other covariates when I evaluated them.

If the covariate is binary, I can evaluate balance using the syntax in Table 27.6 that uses a single group solution.

Table 27.6: Mplus Syntax for Weighted Data for Binary Covariate Balance

```
1. TITLE: BINARY COVARIATE BALANCE ;
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. VARIABLE:
4. NAMES ARE depress risk Tx depbase age motivate
5.  educ assert single econ nonwhite x10 c1 c2 tncomply
6.  depchang prob wght stabwght;
7. CATEGORICAL ARE nonwhite ;
8. USEVARIABLES ARE nonwhite Tx stabwght  ;
9. USEOBSERVATIONS tncomply EQ 0 ;
10. WEIGHT = stabwght;
11. ANALYSIS:  ESTIMATOR = MLR ;
12. MODEL:
13. nonwhite ON Tx (p) ;
14. [nonwhite$1] (a) ;
15. MODEL CONSTRAINT:
16. NEW (PTREAT PCNTRL DIFF) ;
17. PTREAT = exp(-a+p*1)/(1 + exp(-a+p*1)) ;
18. PCNTRL = exp(-a+p*0)/(1 + exp(-a+p*0)) ;
19. DIFF = PTREAT - PCNTRL ;
20. OUTPUT: Samp StdYX Residual Cinterval Tech4 ;
```

Line 7 defines the target variable as binary using the `CATEGORICAL` command. The use of `MLR` on Line 11 tells Mplus to analyze the data using logistic regression. On Lines 13 and 14, I label the logit coefficient (`p`) and the threshold (`a`), with the latter representing the intercept of the logit equation if it is multiplied by -1 (see Chapter 12). I use the `MODEL CONSTRAINT` commands to calculate the weighted proportion of cases that are non-white in the treatment group (`PTREAT`) and the control group (`PCNTRL`) and the difference between the proportions (`DIFF`). I convert the predicted log odds to odds and then convert

[4] For variance ratios, confidence intervals can be asymmetric, so bootstrapping might be preferable in this case.

these odds to probabilities (see Chapter 5). Here is the core output:

```
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.  P-Value

NONWHITE   ON
    TX                -0.002      0.292     -0.006      0.995

 Thresholds
    NONWHITE$1         1.586      0.205      7.731      0.000
New/Additional Parameters
    PTREAT             0.170      0.029      5.783      0.000
    PCNTRL             0.170      0.029      5.874      0.000
    DIFF               0.000      0.041     -0.006      0.995
```

The weighted proportion of non-whites in the intervention group is 0.170 and in the control group it is 0.170, a trivial difference (difference = -0.001, CR = 0.006, $p < 0.996$).

As a further check, I also examined the distribution of the stabilized weights. Cole and Hernán (2008) note that stabilized weights that have a mean near 1.0 and a small standard deviation tend to yield smaller standard errors in the final model. Weight variability also can help you choose between different models for generating the PS probabilities; the model that produces the lowest standard deviation of weights might be preferable, although how well the algorithm balances the covariates also should be considered. If the stabilized weights are highly variable, this might give you pause about IPTW because it suggests the underlying assumptions may be questionable. For the JOBS data, the mean stabilized weight was 0.99, SD= 0.18. The minimum/maximum values were 0.66 and 1.61. These results are reasonable.

Per Protocol Analysis of the JOBS Intervention. I next performed an IPTW analysis on the per protocol sample using the Mplus syntax in Table 27.7.

Table 27.7: Mplus Syntax for IPTW Per Protocol Analysis

```
1. TITLE: PER PROTOCOL IPTW ANALYSIS ;
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. VARIABLE:
4. NAMES ARE depress risk Tx depbase age motivate
5. educ assert single econ nonwhite x10 c1 c2 tncomply
6. depchang prob wght stabwght;
7. USEVARIABLES ARE depress Tx stabwght  ;
10. USEOBSERVATIONS TNCOMPLY eq 0 ;
11. WEIGHT = stabwght;
12. ANALYSIS:  ESTIMATOR = MLR ;
13. MODEL:
14.  depress ON Tx ;
```

```
15. OUTPUT: Samp StdYX Mod(All 4) Residual Cinterval Tech4 ;
```

Here is the core output for this just-identified model:

```
                                                    Two-Tailed
                   Estimate       S.E.    Est./S.E.  P-Value

 DEPRESS  ON
    TX               -0.190      0.087     -2.179      0.029

 Intercepts
    DEPRESS          -0.341      0.067     -5.056      0.000
```

The per protocol effect of the treatment on depression was estimated to lower depression by -0.190 units (MOE = ± 0.17, critical ratio (CR) = 2.18, $p < 0.03$). The intercept is the mean depression value when all predictors equal 0 and it estimates the mean depression for the control group, -0.341. The mean depression for the treatment group was -0.341 +(-.190) = -0.531. These results closely align with those of the direct covariate approach.

Earlier I mentioned the doubly robust strategy to per protocol analysis that combines the direct covariate approach with the IPTW approach. There are several variants of this method. One early strategy was to use IPTW but to also include the covariates from the direct covariate approach in the model analysis if the covariate showed non-trivial imbalance in imbalance diagnostics after weighting. A newer variant has been proposed by Lunceford and Davidian (2004; see also Funk et al., 2011). Here are the steps to execute after isolating the per protocol sample:

Step 1: Fit a logistic regression model predicting the treatment arm from the desired covariates to yield a propensity score for each individual, PS, per usual IPTW methods.

Step 2: Regress the outcome, Y, onto the covariates for the intervention group only. Use the resulting equation to obtain a predicted outcome value for each member of the entire sample. I call this predicted value for a given individual $\hat{Y}_{TREAT}$.

Step 3: Regress the outcome on the covariates for the control group only. Use the resulting equation to obtain predicted values for each member of the entire sample. I call this predicted value for a given individual $\hat{Y}_{CNTRL}$.

Step 4a: For each individual in the intervention, define DR1 as $Y/PS - [\hat{Y}_{TREAT}*(1-PS)]/PS$ and DR0 as $\hat{Y}_{CNTRL}$

Step 4b: For each individual in the control group, define DR1 as $\hat{Y}_{TREAT}$ and DR0 as

$Y/(1\text{-}PS) - [\hat{Y}_{CNTRL}*PS]/(1\text{-}PS)$

Step 5: Calculate the difference score DR1-DR0 and run the Mplus syntax in Table 27.8.

Table 27.8: Mplus Syntax for Doubly Robust Per Protocol Analysis

```
1. TITLE: DOUBLE ROBUST ANALYSIS ;
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. DEFINE:
4. Y0=1.593 + (-0.922)*depbase + (-0.011)*educ +
5. (0.135)*econ + (0.004)*nonwhite ;
6. Y1=1.578 + (-0.751)*depbase + (-0.040)*educ +
7. (0.081)*econ + (-0.098)*nonwhite ;
8. IF (Tx EQ 1) THEN DR0 = Y0 ;
9. IF (Tx EQ 1) THEN DR1 = depress/prob - (Y1*(1-prob))/prob ;
10. IF (Tx EQ 0) THEN DR1 = Y1 ;
11. IF (Tx EQ 0) THEN DR0 = depress/(1-prob) - (Y0*(prob))/(1-prob) ;
12. VARIABLE:
13. NAMES ARE depress risk Tx depbase age motivate
14. educ assert single econ nonwhite x10 c1 c2 tncomply
15. depchang prob wght stabwght;
16. USEVARIABLES ARE DR1 DR0  ;
17. USEOBSERVATIONS TNCOMPLY eq 0 ;
18. ANALYSIS:  ESTIMATOR = MLR ;
19. MODEL:
20.    [DR1] (mtreat)  ;
21.    [DR0] (mcntrl)  ;
22.    DR1 WITH DR0  ;
23. MODEL CONSTRAINT:
24. NEW (DIFF) ;
25. DIFF = mtreat-mcntrl ;
21. OUTPUT: Samp StdYX Residual Cinterval Tech4 ;
```

Lines 3 to 11 use the `DEFINE` command to execute steps 1 to 4 of the doubly robust method. Line 21 asks Mplus to estimate the mean of `DR0` and `DR1` and models the covariance between them (and by default their variances). The `MODEL CONSTRAINT` command calculates the mean difference between `DR1` and `DR0` using the label statements. Here is the core output for this just identified model:

```
                                                    Two-Tailed
                   Estimate       S.E.     Est./S.E.    P-Value

 Means
    DR0              -0.344      0.064     -5.365      0.000
    DR1              -0.534      0.053     -9.990      0.000
 Variances
    DR0               1.439      0.264      5.453      0.000
```

```
    DR1                1.000      0.142      7.033      0.000

New/Additional Parameters
    DIFF              -0.190      0.081     -2.355      0.019
```

The estimated per protocol mean depression at the posttest was -0.534 ±0.11 and for the control group it was -0.344 ±0.13. The difference between the means was -0.190 ±0.16, CR = 2.36, $p < 0.02$.

I now offer commentary on this latter version of the doubly robust test. One advantage of the test over other variants of the double robust strategy is that it allows for interaction effects between the covariates and the treatment condition when deriving predicted Y scores. This is because a separate equation is used to construct $\hat{Y}_{TREAT}$ and $\hat{Y}_{CNTRL}$ (see steps 2 and 3), hence the coefficients can differ for the intervention and control groups. At a more general level, the doubly robust method uses two models to correct for confounds, (a) one that models selection into the treatment condition and (b) one that models the relationship of confounds to the outcome per se. When both of these models are correctly specified, the doubly robust estimator is semiparametric efficient. When both the exposure and outcome models are non-trivially misspecified, the doubly robust estimate will be biased, making the approach of limited value; two wrong models do not make a right. In such cases, it is unclear if it's better to use a single moderately misspecified model in the context of traditional IPTW analysis or the doubly robust model. Finally, if one of the models is correctly specified but the other is not, the doubly robust estimator often yields unbiased coefficients. However, the standard errors can be misbehaved in some contexts, especially with small N (see Funk et al., 2011).

The Mplus program in Table 27.8 used a robust maximum likelihood standard error based on Li and Shen (2020, see their supplement). Several researchers have found that bootstrapped standard errors perform better than theoretically derived standard errors (Funk et al., 2011) but more research is needed on the best form of bootstrapping to use. I tend to rely on the doubly robust method primarily as a sensitivity check because it can be difficult to fully implement in models with mediation and moderation (see below). I discuss other forms of double robust estimation below in the context of G estimation and targeted maximum likelihood estimation.

As a technical aside, the traditional IPTW approach and the direct covariate approach are reasonably (but not perfectly) aligned for ATEs when the outcome in a randomized trial is continuous. For randomized trials where the outcome is binary, a count, or some other form that invokes a non-linear model, estimand correspondence between the approaches can break down because of the issue of noncollapsibility discussed in Chapter 12. For elaboration of this point and an application of Mplus to a

binary outcome using the IPTW approach, see the document *IPTW Analysis of a Binary Outcome* on my webpage.

In sum, for the JOBS data, each of the IPTW methods yielded similar per protocol conclusions and these conclusions mapped closely onto the conclusions from the direct covariate approach. This is not to say such correspondence always will occur. However, the fact that it does so increases our confidence in the result. For additional methods that use propensity scores, see the section on BART modeling in Chapter 15.

CACE and Instrumental Variable Analysis

The CACE framework evaluates per protocol effects by identifying individuals in the control group who likely would be intervention compliers had they been assigned to the intervention condition. The ATE_{CACE} is the difference between the average outcome for these individuals as compared to individuals who were actually assigned to the intervention condition and who complied with the intervention protocol. Many researchers believe this is the best estimand for making inferences about intervention efficacy because it controls for compliance confounds but in an innovative way. The challenge for CACE analysis is identifying likely intervention compliers from the control group had they been assigned to the intervention group. If we had a direct measure of this construct for control group individuals, the analysis would be straightforward; but usually we do not. CACE strategies use indirect methods to obtain the estimate of ATE_{CACE}.

Several approaches have evolved for estimating the CACE average treatment effect (Little & Rubin, 2000; Dunn, Maracy & Tomenson, 2005). One approach uses instrumental variables (Angrist & Imbens, 1995; Angrist & Pischke, 2008), others are based in maximum likelihood analysis (Dunn et al., 2005; Sobel & Muthén, 2012), and still others are tied to Bayesian modeling (Imbens & Rubin, 1997). I focus here on maximum likelihood methods that are grounded in mixture modeling (Sobel & Muthén, 2012). All of the approaches make assumptions in order for the underlying mathematics to work. The assumptions are reasonable in many intervention contexts but not all such contexts. Most of the assumptions deal with the adequacy of the random assignment to the intervention versus control condition and to different forms of contamination that can arise once the study has begun. Let me discuss the latter phenomena first.

Angrist et al. (1996) distinguish four types of people in a randomized trial. **Compliers** are people who follow whatever their treatment assignment is - if assigned to the intervention condition, they do the intervention per protocol; if assigned to the control condition, they don't do the intervention. These individuals are disposed to fully comply with whatever they are told to do in the study. **Always takers** are people who receive or seek out the intervention regardless of their treatment assignment - if they are assigned to

the intervention condition, they do the intervention; if they are assigned to the control condition, they still manage to do the intervention by one means or another. **Never takers** are people who do not receive nor seek out the intervention regardless of condition assignment - if they are assigned to the intervention condition, they don't do the intervention; if they are assigned to the control condition, they also don't do the intervention. Finally, **defiers** are people who do the opposite of whatever their treatment assignment is - if assigned to the intervention condition, they don't do the intervention; if assigned to the control condition, they do the intervention. Although there are some randomized trials where all four types of such contamination occur, randomized trials usually are structured so that always takers, never takers, and defiers are minimized.

If one assumes there are no defiers in a study (a reasonable assumption), Angrist et al. (1996) have shown that with proper random assignment, the intent to treat average treatment effect that researchers often focus on can be decomposed into the sum of three weighted average treatment effects representing the above groups:

$$ATE_{ITT} = \pi_C \, ATE_{CACE} + \pi_A \, ATE_{ALWAYS\text{-}TAKERS} + \pi_N \, ATE_{NEVER\text{-}TAKERS} \quad [27.2]$$

where π_C is the proportion of people who are compliers, π_A is the proportion of people who are always takers, and π_N = the proportion of people who are never takers, and the ATE terms are the average treatment effects for the three different subgroups. The goal is to isolate ATE_{CACE}., namely the effect of the treatment for people who are disposed to compliance. The idea is that by equating the groups on the tendency to comply, then differences between them must be due to the treatment condition they were assigned to. To make this work, researchers must make some assumptions. First, we need to assume that for always takers, the effect of being assigned to either the intervention or the control condition does not, on average, affect their outcome. This is reasonable because, in theory, always takers are exposed to the intervention regardless of the treatment condition they are assigned to. As such, the mean outcome difference between the treatment and control conditions for just these individuals should be zero, or $ATE_{ALWAYS\text{-}TAKERS} = 0$. This assumption results in the $ATE_{ALWAYS\text{-}TAKERS}$ term dropping out of Equation 27.2.

The same is true for $ATE_{NEVER\text{-}TAKERS}$. These individuals are not exposed to the intervention regardless of the treatment condition they are assigned to. As such, the mean outcome difference or value of $ATE_{NEVER\text{-}TAKERS}$ also should be zero because they essentially have the same lack of intervention exposure no matter which treatment condition they are exposed to.

The result of these two assumptions about $ATE_{ALWAYS\text{-}TAKERS}$ and $ATE_{NEVER\text{-}TAKERS}$ is that these ATEs drop out of Equation 27.1, yielding:

$ATE_{ITT} = \pi_C\ ATE_{CACE} + 0 + 0 = \pi_C\ ATE_{CACE}$

If we divide both sides of the above equation by π_C, we isolate the value for ATE_{CACE}, which is what we seek:

$$ATE_{ITT} / \pi_C = ATE_{CACE} \qquad [27.3]$$

The ATE_{CACE} is simply the classic ATE_{ITT} (which is straightforward to compute given no missing data) divided by the proportion of individuals who are disposed towards compliance in the study.

The key to CACE analysis, then, is to obtain a reasonable estimate of π_C. It turns out that there are non-trivial technicalities in doing so and I do not want to get sidetracked into them here (see Angrist et al., 1996, for the details). The main point of the above exercise is to illustrate that coupled with some working assumptions, estimation of the ATE_{CACE} is viable.

In practice, calculating significance tests and confidence intervals for ATE_{CACE} can be complicated. I show you here a mixture modeling approach. The mathematics of the mixture modeling strategy are described in Jo et al. (2008) and Sobel and Muthén (2012) and I leave the mathematical details for you to consider in those references. Here, I focus more on the conceptual foundations and the pragmatics of estimation.

The conceptual bases of mixture based CACE modeling is captured in the influence diagram in Figure 27.3. To keep things simple, I omit disturbance terms but they are part of the model, as appropriate. Exogenous variables also are assumed to be correlated.

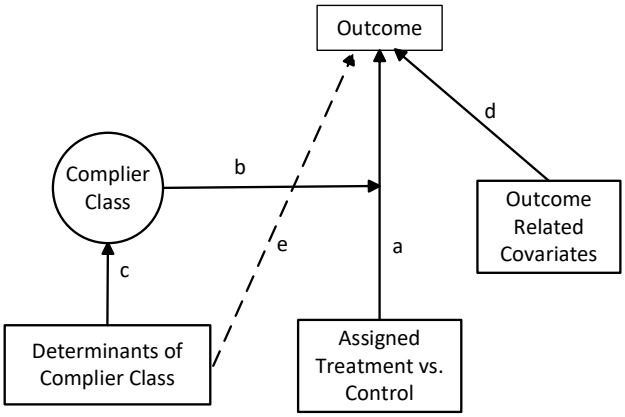

FIGURE 27.3. The CACE Model

The path for the overall average treatment effect that subtracts the control group mean from the intervention group mean is path *a*. The strength of this path is moderated by a latent variable called "complier class" (path *b*), which, has two classes or groups of people, (a) people who are disposed towards compliance or, more simply, "compliers" and (b) people who are disposed towards non-compliance or, more simply, "non-compliers." The effect of the treatment condition (intervention versus control) on outcome for people in the non-complier class/group is assumed to be zero because by grossly non-complying with the intervention, no individuals assigned to the intervention condition in this class are exposed to the intervention, hence it cannot have an effect on the outcome. For people in the complier class, by contrast, the intervention can have an effect so the ATE for this group or class is the ATE_{CACE}.

Technically, the complier class latent variable is unmeasured but it turns out we have information in the data set that can help us classify individuals into the two classes. These are called **training variables** and are used by Mplus to make informed classifications of the individuals into the two groups. For example, in the JOBS study, we have information for people in the intervention condition about whether they complied (showed up to the seminars) or failed to comply with the intervention protocol and this information can be used as "training data" when forming the complier class latent variable. The training data are omitted from Figure 27.3, and I will explain how we use it within Mplus shortly.

The causal model also includes a set of variables, usually measured at baseline, that are thought to impact the complier class that an individual is in (path *c*). By default, Mplus assumes a logistic function that regresses the binary complier class variable onto its hypothesized determinants using logistic regression. The path coefficients associated with path *c* are of interest because they give us insights into the type of people who comply or do not comply with protocols. Note that if it is appropriate, a given determinant of the complier class can be modeled to also have a direct effect on the outcome (see path *e*), but there are some cases where doing so will result in an under-identified model.

Table 27.9 presents the Mplus syntax for a CACE model as applied to the JOBS data. The syntax is inefficient and does not make use of Mplus defaults but it is written to allow me to make certain programming points. The full data set of 502 cases is used without "selecting out" per protocol cases as I did with the direct covariate and IPTW approaches.

Table 27.9: Mplus Syntax for CACE Analysis

```
1. TITLE: CACE ANALYSIS ;
```

```
2. DATA: FILE IS c:\mplus\ret\jobs.dat ;
3. VARIABLE: NAMES ARE depress risk Tx depbase age motivate
4. educ assert single econ nonwhite x10 c1 c2 tncomply
5. depchang prob;
6. USEVARIABLES ARE depress Tx depbase age motivate educ
7.    assert single econ nonwhite c1 c2;
8. CLASSES = c(2);      !specify number of classes
9. TRAINING = c1 c2;    !specify training variables
10. ANALYSIS: TYPE = MIXTURE; MITERATIONS = 30;
11. MODEL:
12. %OVERALL%
13. depress ON Tx depbase educ econ nonwhite;
14. c#1 ON age educ motivate econ assert single nonwhite depbase ;
15. %c#1%   !c#1 is the complier class
16. depress ON Tx depbase educ econ nonwhite (c1p1 p2 p3 p4 p5) ;
17. [depress] (c1int) ;
18. depress (dvar) ;
19. %c#2%  !c#2 is the noncomplier class
20. depress ON Tx@0 depbase educ econ nonwhite (c2p1 p2 p3 p4 p5) ;
21. [depress] (c2int) ;
22. depress (dvar) ;
23. OUTPUT: Samp StdYX Mod(All 4) Residual Cinterval Tech4 ;
```

Line 8 is the `CLASSES` subcommand and tells Mplus the label you will use for the classes and the number of classes. The number of classes is contained in parentheses (in this case, 2). I use the letter `c` as the label for each class, but you can use a different label if you want. The label occurs just before `(2)`. Mplus will add a `#` followed by a sequential integer to the label to give a unique name to each class. In this case, the classes are `c#1` and `c#2`. I conceptualize in the program `c#1` as the complier class and `c#2` as the non-complier class and I write the rest of the syntax accordingly.

Line 9 specifies the training variables in the input data set. The number of training variables must equal the number of classes. Because there are two classes in the current example, I have two training variables. The training variables are variables I created in the input data file. They are called `c1` and `c2` (see the `NAMES` subcommand) but I can use other labels if I want. The variable `c1` refers to the "complier" class and the variable `c2` refers to the non-complier class. Each person receives a score on `c1` and a score on `c2` in the data set, either a 1 or a zero. A zero means the person is not allowed to be in the class; a one means the person could, in principle, be in the class. Individuals in the intervention group have a known class based on their complier/adherence score. A person in the intervention group who is a "complier" receives a 1 on `c1` and a 0 on `c2`. A person in the intervention group who is a "non-complier" receives a 0 on `c1` and a 1 on `c2`. A person in the control group could be in the complier class or s/he could be in the non-complier

class. As such, all individuals in the control group are assigned a 1 on `c1` and a 1 on `c2`.

Line 10 tells Mplus to conduct a mixture model and `MITERATIONS` is the number of iterations to use in the underlying EM algorithm. In the `MODEL` section starting on Line 11, I specify an overall model that applies to each class/group (in the `%OVERALL` section) and then I specify deviations from the overall model within each class, first the `c#1` or complier class on Line 15 and then the `c#2` non-complier class on Line 19. On Line 13 in the `%OVERALL` section, I tell Mplus to regress depression onto the treatment condition and its associated covariates. On Line 14, I tell Mplus to conduct a binary logistic regression that regresses the complier class variable onto the predictors of class membership. By specifying `c#1` as the dependent variable, it will be re-scored 1 and `c#2` will be re-scored 0 for the logistic analysis.

Within the `%c#1%` model section on Line 16, I restate the within class equation from the `%OVERALL` section but I add labels to each of the path coefficients. I also provide labels to the intercept of the equation (Line 17) and the disturbance/residual variance for depression (Line 18). I do the same for the `c#2` model on Lines 19 to 22. Note that by using the same labels to refer to a given parameter in each class I impose an equality constraint for the parameter across the classes. The equality constraints I imposed are commonly used in CACE analysis but if you want, you can remove one or more of them (but be careful of under-identification). Note also that I fixed the effect of the treatment dummy variable at 0 in the non-complier class, consistent with my earlier discussion of CACE assumptions. The disturbance/residual variances are set to be equal across the classes, but this assumption also can be relaxed.

The output provides information about model fit as well as parameter estimates. I consider the model fit results and ancillary parameter information in the Appendix. Here is the core output for the ATE_{CACE}, taken from the `MODEL RESULTS` section:

```
                                                    Two-Tailed
                  Estimate       S.E.  Est./S.E.    P-Value

Latent Class 1

 DEPRESS    ON
    TX            -0.353      0.131     -2.689      0.007
    DEPBASE       -0.908      0.105     -8.655      0.000
    EDUC          -0.029      0.017     -1.765      0.078
    ECON           0.121      0.039      3.120      0.002
    NONWHITE       0.076      0.088      0.864      0.388

 Intercepts
    DEPRESS        1.996      0.385      5.183      0.000
```

```
 Residual Variances
    DEPRESS            0.500      0.036     13.906      0.000

Latent Class 2

 DEPRESS    ON
    TX                 0.000      0.000    999.000    999.000
    DEPBASE           -0.908      0.105     -8.655      0.000
    EDUC              -0.029      0.017     -1.765      0.078
    ECON               0.121      0.039      3.120      0.002
    NONWHITE           0.076      0.088      0.864      0.388

 Intercepts
    DEPRESS            1.705      0.368      4.639      0.000

 Residual Variances
    DEPRESS            0.500      0.036     13.906      0.000
```

The ATE_{CACE} was -0.353 ±0.26, CR = 2.69, $p < 0.01$. Note that this estimate differs from those of the direct covariate approach and the IPTW approach described earlier because it represents a different estimand. The direct covariate and IPTW approaches estimate the per protocol estimand. The CACE approach, by contrast, estimates the average difference between intervention compliers and control individuals who would have complied with the intervention protocol had they been assigned to the intervention group. Note also that the treatment effect for non-compliers in the section `Latent Class 2` equals zero because I fixed it to be so based on CACE assumptions.

The output does not include the estimated posttest mean depression values within the complier class for the intervention and control groups. I can obtain these by re-running the syntax but adding the following `MODEL CONSTRAINT` commands just before Line 23 (I assume you are familiar with the use of `MODEL CONSTRAINT` commands from prior Chapters):

```
22a. MODEL CONSTRAINT:
22b. NEW(MTREAT MCONTROL) ;
22c. MTREAT = c1int+c1p1*1+p2*2.45+p3*13.57+p4*3.54+p5*.1657 ;
22d. MCONTROL = c1int+c1p1*0+p2*2.45+p3*13.57+p4*3.54+p5*.1657 ;
```

Line 22c calculates the predicted mean for the complier intervention group and Line 22d does so for the complier control group. All terms refer to labels used in the complier class linear equation for depression. The predicted mean is the sum of the intercept plus the five path coefficients (p1 through p5) multiplied by a constant of my choosing. The constant represents the profile score on the variable the path label is associated with. In the current case, I multiplied each predictor/covariate by its grand mean value for the

total sample, i.e., the mean of the predictor calculated across all individuals. For example, the mean number of years of education across all 502 individuals was 13.57. Note for the model treatment condition coefficient (labeled `c1p1`), I multiplied it by 1 for the intervention group and 0 for the control group, consistent with the dummy variable that represented it. You, of course, can calculate the predicted means for any predictor profile using the methods described in previous chapters (see, for example, Chapters 6 and 11). Here is the output for the above `MODEL CONSTRAINT` commands:

```
                                                    Two-Tailed
                    Estimate       S.E.    Est./S.E.    P-Value

New/Additional Parameters
    MTREAT            -0.542      0.052    -10.472      0.000
    MCONTROL          -0.189      0.119     -1.586      0.113
```

The predicted mean depression for compliers in the intervention group was -0.542 ±0.10 and in the control group it was -0.189 ±0.24.

The output from the first run also reports the estimated proportion of people in each of the two complier classes, `c#1` and `c#2`. Here is the output:

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THE ESTIMATED MODEL

    Latent
   Classes

       1        271.06984          0.53998
       2        230.93016          0.46002
```

The estimated proportion of people in the complier class was 0.54 and in the non-complier class it was 0.46.

Finally, the output for the first run I performed contains information about the predictors of complier status (class `c#1` versus class `c#2`). Here is the logistic output:

```
Categorical Latent Variables

 C#1        ON
    AGE                0.079      0.015      5.220      0.000
    EDUC               0.308      0.068      4.499      0.000
    MOTIVATE           0.672      0.157      4.278      0.000
    ECON              -0.161      0.150     -1.072      0.284
    ASSERT            -0.378      0.145     -2.605      0.009
    SINGLE             0.514      0.280      1.838      0.066
```

```
    NONWHITE           -0.531      0.323     -1.644      0.100
    DEPBASE            -0.424      0.415     -1.025      0.305

 Intercepts
    C#1                -7.812      1.839     -4.244      0.000
```

and here are the relevant odds ratios:

```
LOGISTIC REGRESSION ODDS RATIO RESULTS

                                                 95% C.I.
                   Estimate       S.E. Lower 2.5% Upper 2.5%

 C#1      ON
    AGE               1.082      0.016      1.051      1.115
    EDUC              1.360      0.093      1.190      1.556
    MOTIVATE          1.958      0.308      1.439      2.664
    ECON              0.851      0.128      0.634      1.142
    ASSERT            0.685      0.099      0.516      0.911
    SINGLE            1.673      0.469      0.966      2.896
    NONWHITE          0.588      0.190      0.312      1.107
    DEPBASE           0.654      0.272      0.290      1.477
```

Statistically significant predictors of complier versus non-complier status included age, education, motivation, and assertiveness. Compliers tended to be older, more educated, more highly motivated to secure re-employment, and more assertive. I show in the Appendix how you can use the `MODEL CONSTRAINT` command to conduct profile analyses using probabilities in a logistic model, following guidelines outlined in Chapter 12.

The CACE based Mplus program offers much flexibility. As noted, applications typically impose equality constraints on the path coefficients across the classes vis-à-vis the use of the common labels p1 through p5 in Table 27.9 for both the complier and non-complier classes. This restriction can be lifted for any given path by using a different label for the same path across classes. I might change the label `(p3)` on Lines 16 and 20 to read `c1p3` and `c2p3` so they have different labels which would remove the equality constraint for them. Note that the disturbance variances also have an across class equality constraint. This also can be removed by using different labels for it in the two classes.

Some researchers feel that CACE estimands are the most appropriate estimands for evaluating treatment efficacy. Other researchers believe they are of limited utility because people who comply to a treatment are not necessarily the same as people who are treatment eligible. The CACE estimate, the argument goes, is thereby limited in terms of its generalizability (Marbach & Hangartner, 2020). In my view, most per protocol estimands focus on hypothetical populations (and this also is true of ITT estimands) and

issues of generalizability and applicability to different subgroups ultimately must be addressed empirically. For interesting extensions and applications of the CACE method, see Jo (2000a, b; Jo, Asparouhov & Muthén, 2008; Jo, Ginexi & Ialongo, 2010), Sobel & Muthén (2012), Stuart, Perry, Le, & Ialongo (2008), Peugh et al. (2017), Hesser et al., (2017), Hesser (2020), and Ashworth et al. (2020).

G Computation and Targeted Maximum Likelihood Estimation

Another modern approach for estimating per protocol efficacy is known as **G computation** or G estimation. The approach has been used less frequently than the approaches I have discussed, but it is gaining in popularity. After describing G computation, I consider an extension of it known as **targeted maximum likelihood estimation**.

There are many forms of G computation ranging from fairly simple parametric modeling with an outcome measured at a single time point to rather complex structural nested models (SNMs) that involve multiple observations over time with repeated administration of the treatment, such as cancer trials of responses to chemotherapy continuously delivered over periods of months with efficacy assessments made after each administration.[5] In this section, I consider the most elementary forms of parametric G estimation to give you an intuitive feel of its logic (Snowden, Rose & Mortimer, 2011; Vansteelandt & Keiding, 2011).

In simple parametric G-computation, the first step is to isolate the per protocol sample and then to fit a regression model that predicts the outcome from the treatment condition and the covariates that one believes should be controlled to deal with confounds that undermine per protocol inferences, much like the direct covariate approach. This first step regression model is called the **Q model**. It does not have to take the form of a linear model but it often does; recent applications sometimes use machine learning algorithms to assist predictor selection, but I have reservations about such strategies because of their atheoretical nature, as noted earlier.

Once the Q model is estimated using the per protocol sample, the model is then used to generate counterfactual or potential outcome values for each individual as a function of the two treatment conditions. For example, a "potential outcome" value is generated for the first individual under the assumption the individual is in the intervention group and then another "potential outcome" value is generated for the same individual under the assumption s/he is in the control group. This is accomplished using methods similar to the calculation of average marginal effects for a dummy predictor that

[5] Some investigators view inverse probability weighting as a form of G computation, for reasons I describe shortly.

I described in Chapters 5 and 12. In the JOBS data, the regression equation (Q model) predicting depression from the dummy coded treatment condition (0 = control group, 1 = intervention group) and the covariates was

Depress = 1.674 + -.193 Tx + -.844 Depbase + -.026 Educ + .115 Econ + -.040 Nonwhite

For the first individual, I calculate that person's predicted depression score by substituting in that person's scores into the above equation on each of the covariates but I assign a score of 1 to the treatment condition variable, Tx, no matter what condition the person was in. This yields the predicted score for that person as if the person had been exposed to the intervention. It is one of the "possible outcomes" the person could have. I then repeat this process for the same individual but now I assign a 0 to Tx no matter what condition the person was in. This yields the predicted score for that person as if the person had *not* been exposed to the intervention. It is the other "possible outcome" for the person. The difference between the two calculated possible outcomes is the marginal effect of the intervention on depression *for that particular person*. I repeat this process for every case/person in the sample, calculating an individualized marginal effect for each one. Finally, I compute the average of these individualized marginal effects across all individuals. The result is a sample estimate of the per protocol average treatment effect.

If I use this method I am, in essence, comparing two populations, one population in which every person has been exposed to the intervention as compared to a population in which these same people have not been exposed to the intervention. By definition, each population has the same distribution of values on the covariates or other predictors because, after all, they are the same individuals in both populations. Given that the only difference between the two populations is their "exposure" to the intervention, the intervention must be the source of their mean outcome differences. In the JOBS data, the estimated ATE_{PP} using G computation was -0.193 ±0.16. When the outcome is binary, the Q model can take the form of a logit, probit or modified linear probability model and the two potential outcomes for a given individual are operationalized using predicted probabilities for each individual based on the Q model.

Calculating the estimated standard error, p values, and confidence intervals for ATE_{PP} in G computation is not straightforward and one usually resorts to bootstrapping to do so. I provide a program on my website called *G computation* that does the calculations for either binary, continuous, or count outcomes. The video associated with the program provides examples with binary and continuous outcomes.

There is a subtle difference in the way G computation conceptualizes average treatment effects compared to how traditional regression does in the direct covariate

approach. The direct covariate approach uses a conditional regression model. In the JOBS study, the per protocol equation predicts depression from the following equation:

Depress = a + b1 Tx + b2 Depbase + b3 Educ + b4 Econ + b5 Nonwhite

The estimated average treatment effect is b1, which is the mean depression difference between the intervention and control groups *conditional on values of the other predictors being held constant at a specific set of values* (such as Educ = 13, Nonwhite = 1, and so on). In a linear main effect model with a continuous outcome such as the above, it turns out the value of b1 will be the same for any set of predictor profile values defined by the other predictors. In G estimation, by contrast, the reported average treatment effect is the mean difference between the intervention and control potential outcomes *calculated across the distributions of the other predictors* rather than at specific values of those predictors. The G computation approach is more compatible with the philosophy of average marginal effects discussed in Chapters 5 and 12. Technically, G computations are a hybrid because the Q model is a conditional model. However, there is a subtle difference at work that has analytic implications in some contexts, such as with binary outcomes.

Recently, attempts to improve G computation have been suggested in the form of a method known as **targeted maximum likelihood estimation** (TMLE). As noted earlier, confounding can be addressed by "breaking" or eliminating the association between the outcome and the confounding variables and/or by "breaking" or eliminating the association between the confounders and selection into the treatment condition. Traditional G-computation uses the former. TMLE addresses both and in this sense, it is a doubly robust method. The technique was first proposed by van der Laan and Rubin (2006) and also has been extended to include machine learning methods for variable selection, if desired. TMLE computes potential outcomes much like G computation but it introduces a "targeting step" that also addresses imbalance between the intervention and control conditions vis-à-vis propensity score methods. Using a complex iterative algorithm, the method simultaneously seeks to minimize treatment condition imbalance while also addressing the function relating the outcome to the confounders. For tutorials on TMLE, see Schuler and Rose (2017), Luque-Fernandez, Schomaker, Rachet & Schnitzer, (2018) and Oang et al. (2016). An R package *tmle* implements the method. Balzer et al. (2019) have extended the method to clustered designs. The advantages of TMLE over more traditional G computation and IPTW methods need further exploration but the method has promise given its doubly robust nature. I provide a program on my web page for conducting TMLE analyses.

Dosage Analysis

Compliance or adherence to protocols is central to efficacy analysis because protocol adherence ultimately affects the "dosage" of the intervention a person receives. In all of the prior examples, compliance or adherence was treated dichotomously in the sense that a minimum dose threshold was set that was deemed necessary for the intervention to work. Dropping out of a treatment affects intervention dosage as does remaining in treatment but failing to follow treatment protocols does. In the behavioral sciences, many interventions consist of multiple sessions and the threshold for meaningful exposure is defined as the number of sessions participants complete. The criteria are based on past research, common sense, or logical criteria, although often the choices seem arbitrary. Stuart, Perry, Le & Ialongo (2008) considered two multi-faceted interventions, one consisting of eight classes participants were to attend and another consisting of 66 take home activities that participants were to complete. For the former, researchers deemed attending at least four of the eight classes to be the minimum number a participant should attend in order to benefit from the intervention. For the second intervention, completing 45 of the 66 take-home activities was defined as "full" participation. It is not uncommon in the per protocol analytic strategies described earlier for researchers to perform sensitivity analyses under different adherence thresholds. For example, one can evaluate what the estimated ATE_{CACE} is if the threshold for defining compliers is set to attending 4 or more classes; what is it if the threshold is set to attending 5 or more classes; and so on.

Some scientists treat adherence and dose exposure as a continuous or many valued quantitative construct and seek to determine the relationship between the amount of adherence/exposure and the outcome. Sometimes this takes the form of making dose a formal design factor in which individuals are randomly assigned to different dosage groups. Other times, adherence (and consequently dose) are simply measured and treated as any other variable in an RET, perhaps as a mediator, a moderator, or a covariate.

A common strategy for exploring the relationship between adherence and treatment response is to construct a model of the link between them focusing only on individuals in the intervention group. In some contexts, such as the taking of placebos over time in a biologic trial, one can include the intervention group in such analyses (thereby advantaging the total N of the study) because one has an index of the extent to which the controls have completed their placebo-infused regimen. Despite this, intervention protocols can present different challenges to participants than control protocols (e.g., side effects might be more prevalent or severe in the intervention group) so that compliant subgroups in the two conditions are not comparable. CACE analyses are intended to address this matter. Nevertheless, researchers often deal with the intervention-control complier non-equivalence by studying compliance-outcome links for just those

individuals in the intervention group.

Such a practice removes randomization from the picture and turns the enterprise into a purely correlational/observational analysis with all the limitations that go with it. Covariate and confound controls become crucial. Some researchers correlate indices of compliance with baseline to posttreatment change scores, but this practice is ill advised for reasons I discussed in Chapter XX. Where possible, using the baseline outcome as a control covariate when relating compliance to the outcome at posttest is likely a better strategy. A host of propensity score methods have been proposed for working with such observational data but consideration of these methods is beyond the scope of this book. Many of the methods work with concepts I outlined for IPTW adjustments. See DuGoff, Schuler, & Stuart (2014) and Guo, Fraser & Chen (2020) for tutorials.

I like to think of the adherence construct not so much in terms of participant protocol compliance but rather in terms of intervention dosage. Doses can differ in frequency, duration, and amount, each of which represents a distinct facet of dosing that can contribute independently to the outcome. Non-adherence might affect some of these facets but not others. Interventions often can be decomposed into different subparts; a part that addresses mechanism A, another part that addresses mechanism B, and a third part that addresses mechanism C. Often the components are addressed sequentially in the intervention, so the timing of dropping out of treatment or of skipping a session impacts the particular components one is exposed to. These qualitative differences in dosing also can differentially affect the outcome. Factors that impact the frequency, amount, duration and nature of intervention doses can include patient-centered factors, therapy-related factors, social and economic factors, healthcare system factors, and disease factors. The bottom line is that mapping adherence, compliance or dosing onto treatment outcomes is far more complex than simply relating the number of sessions attended or the number of homework assignments completed to outcomes. In some ways, popular per protocol analytic strategies are limited because they treat the construct of “per protocol” so crudely.

Efficacy Analyses with Missing Data

My discussion of analytic methods for dealing with per protocol non-adherence has assumed there is no missing data. In many program evaluations, missing data can occur and this is especially true for treatment dropouts who are unable to be contacted for purposes of outcome assessment at the scheduled time of the posttest. For efficacy analyses, treatment dropouts are eliminated from the per protocol sample so no additional steps are needed to deal with them. In the control group, you may have individuals who provide baseline data but not posttest data and the question becomes the best ways to deal

with them.

One way to approach such missing data is to use listwise deletion but this assumes the missing data are missing completely at random (MCAR) and that violations of MCAR are not sufficient to create meaningful bias (see Chapter 26). If one assumes data are missing at random (MAR) rather than MCAR, then standard full information maximum likelihood (FIML) analyses can be applied for both the direct covariate approach and for the CACE approach.[6] However, in CACE analyses, if compliance is assessed only in the intervention condition, then addressing the MAR assumption is messy because missingness is assumed to depend on compliance status plus the measured covariates in the intervention group but only on the measured covariates in the control group (see Frangakis & Rubin, 1999, 2002). For example, in a family intervention, a low level of completion of intervention activities may be due to family instability, such as the tendency to move to a new residence or having to deal with financial stress. Such non-compliers also may be harder to reach for posttest assessments, meaning that missingness will be higher among well-complying individuals, including those in the control group who would have complied with the treatment protocol had they been assigned to the intervention condition. Missingness is thus impacted by unobserved compliance status in the control condition which violates MAR. The violation can introduce bias if the degree of violation is non-trivial.

For CACE modeling, Jo, Ginexi & Ialongo (2010) suggest ways of dealing with this dilemma by using a variant of CACE that makes weaker assumptions than MAR. They used the framework of Frangakis and Rubin (1999) who describe a missing data mechanism called **latent ignorability**. Latent ignorability assumes that potential outcomes and indicators of potential outcome missing data are independent within each level of the latent compliance variable. Jo et al. (2010) describe ways of addressing this mechanism in CACE modeling that require only minor modifications to Mplus syntax. I describe the modifications as applied to Table 27.9 in a document on my web page titled *CACE and Missing Data*.

For IPTW and G estimation, some methodologists recommend the use of multiple imputation to deal with missing data (Seaman, White & Copas, 2012; Seaman & White, 2014). These methods also assume data are MAR or that violations of MAR are not sufficient to meaningfully bias estimates.

In the final analysis, and as I stated in Chapter 26, the best method for dealing with missing data is not to have any or to have so little that it is not of consequence under listwise deletion. Do your best to minimize missing data.

[6] FIML is the default used in Mplus for both of these methods.

Concluding Comments on Efficacy Focused Analyses

Efficacy-based randomized explanatory trials are important. They provide us with insights into the relationships between outcomes and potential determinants of those outcomes while at the same time giving us a sense of our ability to change those determinants, all without the clutter of non-adherence and poorly implemented protocols. This is particularly important in the early stages of intervention development and refinement. Issues of non-adherence and poor intervention implementation can and should be addressed at the design stage of an efficacy trial with the idea of minimizing implementation infidelity. However, as hard as we try, there often will be slippage in implementation fidelity and adherence to protocols. In these cases, there are an array of modern analytic tools that can be brought to bear to keep the focus on treatment efficacy. These include the direct covariate approach, IPTW weighting, CACE analysis, G computations, and TLME. Each method has strengths and weaknesses and it probably is best to approach one's data from multiple analytic perspectives in the spirit of a sensitivity framework. A poor strategy for dealing with unavoidable non-adherence in an efficacy trial is to shift to ITT analyses. Such a shift changes the research questions. As the eminent statistician John Tukey notes, it is "far better to have an approximate answer to the right question than an exact answer to the wrong question." Dallal (2012) takes the matter further by referring to ITT analyses in such contexts as "fraudulent."

Critics are not wrong when they say that traditional per protocol analyses can compromise randomization. They can. The question is whether in a given study randomization has, in fact, been compromised by focusing on a per protocol sample (it may or may not be) and, if so, does the nature of the imbalance that results between the treatment and control groups lead to inaccurate inferences? We can test for imbalance between the treatment and control conditions on variables *we have measured* but we have no idea if imbalance has been created on variables we have not measured. And, imbalance that might occur on some unmeasured or measured variables (e.g., shoe size, to give a tongue in cheek example) don't matter if those variables are unrelated to or do not impact the outcome. Imbalance is only relevant to variables that matter.

All this means that a researcher who is planning a study needs to think long and hard about identifying nuisance variables that (a) impact the outcome in non-trivial ways, and (b) that might be subject to imbalance between the treatment and control conditions when a per protocol sample is defined for purposes of analyzing efficacy. After making a list of such variables, the researcher should be sure to measure them (or the most important ones) because the newer, more modern methods of per protocol analysis that adjust for consequential imbalance typically require that we have measures of the biasing variables. If variables that create significant bias are unmeasured, then this is potentially

problematic. Of course, we can't measure everything, so instead we must measure those variables that we a priori think will matter the most. With those measures in hand, we then can (a) evaluate the extent of imbalance that occurs for them and (b) if imbalance exists, we can apply a modern method (direct covariates, IPTW, CACE, g computation, targeted maximum likelihood) to negate their impact and make more accurate efficacy statements.

Having said all that, program administrators typically want to know if their programs are effective and how they can make their programs more effective. They are not so much interested in advancing science as they are with helping their clients vis-a-vis their particular program. As I already noted, the best advice for increasing the effectiveness of a program is garnered by pursing both efficacy and effectiveness analyses in one's evaluation effort because depending on their results, you might make different recommendations for program improvement. A program that is ineffective because it lacks efficacy requires different remedial actions than a program that is efficacious but lacks effectiveness. To be sure, the design of hybrid studies that provide perspectives on both efficacy and effectiveness is challenging because efficacy analysis, in principle, necessitates designs that create high levels of adherence even at some cost to what is realistically possible in applied contexts. Good hybrid design requires a careful balancing of what might be versus what is. Hybrid designs benefit from a focus on both efficacy based and adherence based mediators and moderators and they require researchers have in their analytic toolbox a strong set of tools for addressing both efficacy and effectiveness questions. The present chapter describes such tools.

Proper per protocol analyses have been extended to more complicated designs than the ones I have considered here, including survival modeling and treatments that are administered and can vary across time (e.g., Toh & Hernán, 2008; Toh, Hernández-Díaz, Logan, Robins & Hernán, 2010; Lodi et al., 2016). See my website for additional extensions of analytic strategies. I discuss later in this chapter extensions to trials with mediators and moderators.

EFFECTIVENESS (INTENT TO TREAT) FOCUSED ANALYSES

ITT analyses are relevant to questions of treatment effectiveness. They compare outcome means or proportions for the intervention versus control groups for all individuals who were randomized to condition irrespective of any contamination or non-compliance that may have occurred after randomization. The idea is that contamination and non-compliance operate in real world settings so one should not adjust for them if one wants to determine the effectiveness of the intervention in applied settings. Of course, it is

possible that contamination and non-adherence in an artificial randomized trial may not be reflective of the nature and amount of contamination and non-adherence in real world settings. To the extent that such generalizability is lacking, without covariates or design adjustments to make contamination/non-compliance representative of real-world contamination/compliance, ITT results can be misleading.

As examples, suppose, as a lay person, I enroll myself in a randomized trial and I am compensated for participation. Could the fact that I am compensated for participation impact my motivation to adhere to the intervention protocol? (The answer is "yes."). I might know that I am in a trial that can improve a program/intervention for others in my community and might infer that my failing to adhere to the intervention protocol can affect the validity of the trial. Could such knowledge affect my adherence to protocols? (The answer is "yes."). Is the type of person who fails to adhere to the intervention protocol under RCT/RET participation the same as the type of person who would not adhere to intervention protocols outside the context of a randomized trial? We do not know. Suppose in a family intervention I learn a truly valuable parenting tip that I think would benefit my closest friend who is the parent of a troubled adolescent. Would I be just as likely to share that information with my best friend if I knew I was formally enrolled in a randomized trial and had been told not to share information as compared to if I was not in a randomized trial and gained this information in a parenting program I attended at my daughter's school? ITT analyses of an RCT or RET permit "noise" to enter conclusions about treatment effectiveness but what if that noise is unrepresentative and unrealistic?

Despite the preference for ITT analyses by many researchers, such analyses often are misunderstood or poorly implemented. In this section, I consider three matters about treatment dropouts and ITT analyses, (1) the timing of dropping out of treatment in a pre-post control group design, (2) the use of full information maximum likelihood analysis to deal with missing data due to treatment dropouts, and (3) imputation strategies to deal with missing data due to treatment dropouts. In my discussion, I assume you are familiar with the material on missing data in Chapter 26.

An underappreciated fact about ITT analysis is that proper ITT analysis *requires having posttreatment data for treatment dropouts*. In many randomized trials, when a person drops out of a study, such data are not available. This includes dropouts in the intervention condition as well as the control condition if the control condition is an active control. Missing posttest data for intervention dropouts must be dealt with in light of the points I made at the outset of this chapter. If treatment dropouts are relatively few or if treatment dropout is completely random for both treatment and control groups, then listwise deletion of cases will yield unbiased estimates of the ITT average treatment

effect but with reduced statistical power relative to the no missing data case. If the sample size is large enough, power reductions will not be problematic. If these conditions do not hold, then the missing data must be dealt with by other means, which I describe below.

The Timing of Treatment Drop Out Relative to Baseline Assessment

Some researchers use trial designs where baseline assessments are taken prior to randomization; others randomize participants first and then take baseline assessments prior to the commencement of intervention activities. I have encountered cases where randomization is pursued initially only to have a subset of people drop out of the study before completing the baseline assessments. In such cases, there are no data for the dropouts whatsoever, so not much can be done about it analytically. I try to incentivize such dropouts for exit interviews to help me understand their drop out dynamics, I ask them to complete a brief assessment to allow selected baseline comparisons between them and those who continue with the study. Some researchers in such scenarios replace pre-baseline dropouts by randomly selecting a new individual from the population to take the person's place. That replacement then completes the baseline and the rest of the study protocol to which the dropout had been assigned. Technically, randomization is compromised by the use of such replacement participants but it may be that the degree of bias, if it occurs, is not consequential.

For designs where individuals complete baseline assessments but who fail to commence treatment, randomization can be preserved for ITT purposes by retaining them in the study and treating their posttest data as missing if such data have not been collected. Such dropouts are conceptualized as non-compliers.

Full Information Maximum Likelihood Analysis and ITT Analyses

Many researchers apply traditional full information maximum likelihood (FIML) algorithms for ITT analyses that have missing data due to treatment dropouts who could not be followed up at posttest. FIML may or may not be misleading in such cases. If the missing outcome posttest data is MCAR or MAR, then estimates of ATE_{ITT} using FIML will indeed be unbiased. If, for example, some people miss a single session intervention due to bad weather and do not complete the post-intervention survey at the posttest assessment session because of that bad weather, the posttest missing data are likely MCAR and can be handled by FIML. If, on the other hand, missing data due to dropout are not MAR, estimation bias can result. Stated another way, certain conditions must be present for the FIML strategy to adequately handle treatment dropouts. Let me illustrate this dynamic by revisiting the weight loss example by Dallal (2012) that I provided at the outset of this chapter.

Recall that Dallal described a study comparing two diets, one that, unbeknownst to the investigator, was effective and the other ineffective. Dallal characterized a dropout mechanism in which people who were not losing weight as the diet progresses stop the diet and then drop out of the study, thereby providing no post-diet weight data. For those on the ineffective diet, some participants lose weight regardless of the ineffectiveness of the diet and they stay in the study. Others gain weight and drop out. The result is that the effective diet (where people lose weight) appears less effective than the ineffective diet because the only people who remain in the study on the ineffective diet and who provide posttest data are those losing weight.

I simulated a large data set to minimize sampling error for my demonstration of the bias dynamic at work. I sampled 500,000 women who were then randomly assigned to one of the two programs, D1 (the effective diet) or D2 (the ineffective diet). The baseline weight of the women was normally distributed with a mean of 165 pounds and a standard deviation of 7 pounds. Assuming everyone provided posttest data in the population, the true average difference in weight loss between those randomly assigned to D1 versus D2 was -10 pounds, i.e., D1 was, on average, 10 pounds more effective than D2. This is the true ATE_{ITT} in my simulation.

Now, suppose I introduce a version of Dallal's treatment dropout rule: For both D1 and D2, if a woman failed to lose weight by serendipitously *gaining* half a pound or more halfway through the intervention, she stopped the diet, dropped out of the study, and did not provide posttest data. When I created my simulation, I created a mid-treatment assessment of weight that reflected, on average, a -5 pound change in weight in the D1 condition i.e., half the true total effect of diet type, -10, plus some random noise. For the D2 condition, there was no average change in weight at midtreatment, but some women changed upward and others downward because of the random noise that was operating in D2. I then applied the above drop-out algorithm to the data based on the midtreatment data: If a woman in either condition gained 0.5 or more pounds at midtreatment, I changed her observed score at the posttest to missing data. The result was that 43.3% of the women in D2 dropped out of the study following the midtreatment assessment but only 3.4% of women in D1 did so. The differential dropout rate was due to the fact that D1 was indeed effective for the vast majority of women in reducing weight midway through treatment (by about -5 pounds) whereas D2 was not. Note that in theory I do not know what the treatment dropouts in either condition did with respect to their eating or lifestyle habits after dropping out of the study; all I know is that when it was all said and done, if I measured the posttest weight for all the women who initially enrolled in the study, women randomly assigned to D1 were, on average, 10 pounds lighter than those randomly assigned to D2 at posttest, i.e., $ATT_{ITT} = -10$.

In this scenario, the missing data are NMAR. Dropping out of treatment, and hence having no posttest data, is impacted by a women's weight halfway through the program given their baseline weight. I listwise deleted the cases with missing data at the posttest and then regressed the posttest data onto the treatment condition (D1=1, D2=0) and baseline weight using the following Mplus syntax:

```
1. TITLE: Diet program simulation
2. DATA: FILE IS weightM.txt;
3. LISTWISE = ON ;
4. VARIABLE: NAMES ARE id treat basew mid post ;
5. USEVARIABLES ARE treat post basew ;
6. MISSING ALL (-9999) ;
7. ANALYSIS: ESTIMATOR = MLR;
8. MODEL:
9. post ON treat basew ;
10. OUTPUT: Samp StdYX Residual Cinterval Tech4 ;
```

All of the syntax should be familiar to you. The path coefficient for the treatment dummy variable `treat` in this analysis was -8.15 which, as Dallal suggested, underestimates the true average ITT treatment effect of -10.00 by almost 2 pounds. Using listwise deletion in the face of missing data that are MNAR is not a good idea in this case.

I next applied FIML to the data by eliminating Line 3 from the syntax and making use of the Mplus default to apply missing data FIML to the endogenous variables in the model. The path coefficient for `treat` in this new analysis was again -8.15. Applying FIML does not help. The reason FIML fails is because the data are not MAR, which violates FIML assumptions.

As I discussed in Chapter 26, one way of dealing with missing data that are NMAR is to identify the systematic source of missingness and then to bring into the modeling effort a measured variable that reflects that source in order to control for it, i.e., turn the NMAR case into a MAR case. In the simulated data, the variable `mid` is the assessed midtreatment weight of women in D1 and D2 and reflects, more or less, the source of the missingness in the statistical model conditional on one's baseline weight. I can formally bring this variable into my model as a covariate or, if it is not of substantive interest, I can use a saturated correlates approach in conjunction with the `AUXILIARY` command in Mplus to address the NMAR (see Chapter 26 for discussion of this method). I used the latter approach by adding the following command just after Line 6 in the above syntax:

```
6a. AUXILIARY = (m)  mid ;
```

The path coefficient for `treat` in this new analysis was -10.00, which reflects the true ATE_{ITT}.

In sum, FIML can be used to address treatment dropout missing data in ITT analyses in cases where some of the treatment dropouts do not provide posttest data as long as the statistical model formally takes into account sources of NMAR so as to make the missing data MAR or MCAR. The FIML approach yields reduced statistical power relative to the case where one is able to obtain posttest assessments on all or most treatment dropouts, but it often remains viable under such circumstances. This means when planning an RET one should give careful thought to the sources of treatment dropout so that they can be measured and brought under control analytically. For cases where one is concerned about people dropping out because the intervention is ineffective or because of adverse side effects of a medication, a strategically placed "during treatment" assessment might provide a means for addressing the resulting NMAR through the use of the saturated correlates auxiliary command in Mplus, as I showed for the case of the variable `mid` in the above example. If you have measures of mechanisms producing problematic missing data for loss to follow-up as discussed in Chapter 26 and measures of other mechanisms producing treatment dropouts as discussed in the present Chapter, then you can include both sets of variables in the saturated correlates auxiliary command to control for all of them.

In the literatures I follow, many RCTs apply FIML in the context of ITT analysis but they mention nothing about possible bias due to MNAR at the posttest for purposes of controlling covariates associated with treatment dropout nor do they include any covariates in their modeling efforts. Or, they engage in practices in which they claim they are conducting ITT analyses when, in fact, they are unwittingly using per protocol analyses. Here is an example. In one study, no posttest data were obtained on treatment dropouts. The study conducted "ITT" analyses by using traditional FIML for missing data per Chapter 26 as applied to the data at hand. FIML seeks to construct an estimate of the population means and covariance matrix of all variables in the analysis in light of missing data but in this case, the population (and sample) matrix includes *no information* about the posttest scores of treatment dropouts. This is analogous to doing the study with no treatment dropouts in it, i.e., it functionally is a per protocol analysis. We need to be more rigorous about such analyses and this, at a minimum, requires that we make a good faith effort to get substantial amounts of data at posttest for treatment dropouts.

Imputation Strategies and ITT Analyses

A second approach to dealing with missing data from treatment dropouts in ITT analyses is to use some form of imputation. Two commonly used single imputation strategies in RCTs are the **last observation carried forward** (LOCF) and the **worst case** (WC) imputation methods. LOCF imputes a score for the missing data that equals the value of

the same variable measured at the closest time point prior to the occurrence of the missing data. In a pre-post control group design with no mid-treatment assessments, this means imputing the baseline score for an individual into the posttest score, essentially producing no change in the outcome for that individual. If a mid-treatment assessment is taken, then the mid-treatment score on the outcome is imputed to the posttest missing data. If a variable is relatively stable over time and one expects little change in it from one time period to the next in the real world, then variants of LOFCs may be a reasonable imputation strategy. For example, chronic pain patients often feel the same level of pain across extended periods of time and it is not unreasonable to assume that their pain levels remain constant from baseline to posttest if they have dropped out of treatment. However, in a pre-post control group design, use of LOCF generally will inflate the correlation between the pretest and the posttest because one uses the same score to represent the outcome values at the two time points. This presumed stability, if false, can bias coefficients when the baseline is used as a covariate and it also can bias standard errors. Lachin (2016) has shown that LOCF requires data that are MCAR and that for pretest-posttest designs where the values at time 1 are used to impute the missing values at time 2 by LOCF, the result is a mixture of the time 1 and time 2 distributions that almost always yields biased average treatment effects unless the two time periods have identical distributions. Despite the Food and Drug Administration embracing LOCF in some circumstances (Lachin, 2016), statisticians generally recommend against its use because of these types of bias (see Kenward & Molenberghs, 2009; Li & Stuart, 2019). Indeed, some leading biomedical journals have a formal policy of not accepting articles that use LOCF (Little et al., 2012; Newgard & Lewis, 2015). The Panel on Handling Missing Data in Clinical Trials of the National Academy of Sciences recommends that methods like LOCF should not be used as a primary approach to treat missing data unless the underlying assumptions are scientifically justified (National Academy of Sciences, 2010).

In contrast to LOCF, the WC imputation method imputes to treatment dropouts with missing posttest data the worse possible score on the outcome at posttest. A variant of it sometimes used for sensitivity analyses is **best case** (BC) imputation in which the best possible score on the outcome is imputed to treatment dropouts with missing posttest data. The selection of best or worst values to impute can be based on different criteria, the specification of which can be controversial; researchers define them in ways they think are most appropriate to the study context and goals. The WC and BC methods typically are invoked in sensitivity analyses to identify worse-case and best-case scenarios for estimating the ATE_{ITT}. Another variant of the WC imputation method is the **jump to reference** imputation approach (Cro, Morris, Kenward & Carpenter, 2016). This method imputes the outcome mean of the control group into the scores of intervention dropouts

with missing data. All three of these methods are known to potentially produce biased estimates of ATE_{ITT} and biased standard errors so they are of questionable utility as a strategy for dealing with missing data.

Some researchers use variants of the above methods by introducing a small degree of random noise into the imputation so that multiple imputation can be used to obtain better estimates of the relevant standard errors. However, the bottom line is that other forms of imputation typically yield results that are less biased for estimating the true ATE_{ITT} than LOCF, BC or WC strategies.

A promising multiple imputation strategy for estimating ATE_{ITT} when there are treatment dropouts with NMAR missing data is the **retrieved dropout imputation method** (James, 2012; Wang & Hu, 2022). In this approach, you isolate all cases of individuals who dropped out of treatment, including (a) those dropouts who failed to provide data at the posttest and (b) those who you made special efforts to track down and were able to obtain posttest assessments despite the fact they dropped out of treatment, i.e., **retrieved dropouts**. For this subpopulation of retrieved and unretrieved dropouts combined, you use either the chained equation multiple imputation approach or the Bayesian H1 multiple imputation approach described in Chapter 26 to generate multiple imputations of posttest data for the unretrieved dropouts using the retrieved dropouts as the donor pool. The imputation model should include a full range of thoughtfully identified covariates including the baseline outcome measure. For treatment completers with missing data, you do a separate but parallel multiple imputation process for them and then merge their imputed data with the imputed data for the treatment dropouts into a single imputation sample. You repeat this process to generate, say, 100 imputation samples and then apply the analytic model to each of these imputed data sets. You then combine the results into a single ATE_{ITT} estimate per standard multiple imputation methods, per Chapter 26. Wang and Hu (2022) found support for the approach given the number of retrieved dropouts was not small and the amount of missing data in the dropout population was not large (less than about 30%).

When using the retrieved dropout method one typically assumes that retrievable individuals are representative of treatment dropouts more generally, an assumption that may be questionable in some contexts. Interestingly, in research that has studied retrieved dropouts, improvements in outcomes relative to the control group are sometimes observed (Farlow, Potkin, Koumaras, Veach & Mirski, 2003). This result might be due to partial exposure to the intervention, to off-protocol compensatory activities on the part of the dropouts, or to confounds of the retrieval process.

Mixed Effects ITT Analyses

As noted in Chapters 16 and 26, an approach to analyzing treatment effects in longitudinal randomized trials is mixed effects modeling. In a two group pretest-posttest control group design, the primary question of interest in a mixed effect analysis is whether the groups have different average improvements in the outcome from baseline to posttest. This is typically evaluated in terms of the treatment by time interaction effect. A touted advantage of mixed effect modeling is that it can estimate treatment effects in the presence of treatment dropouts who did not complete the posttest without recourse to imputation, relying instead on variants of the FIML method described in Chapter 26. Like FIML, these methods require that the missing data be MAR or MCAR. The same challenges I described above for FIML when applied to ITT analyses with treatment dropouts apply to mixed effects modeling; if the missing data due to treatment dropouts are NMAR, then the ATE_{ITT} can be biased. Strategies are needed to bring covariates into the model that turn the NMAR properties into MAR or MCAR properties. In Mplus, this can be accomplished using the saturated correlates approach with FIML. The retrieved dropout multiple imputation strategy also can be used in mixed effects modeling.

Concluding Comments on ITT and Effectiveness Analysis

ITT analyses of randomized trials can provide useful insights into program effectiveness. The method is best applied when data have been collected on all study participants who were randomized to condition. Unfortunately, complete data are not always available, especially for treatment dropouts who researchers lose contact with. If dropping out of treatment is completely random (i.e., the missing data are MCAR), then analyses are straightforward and viable strategies for dealing with the missing data due to dropouts include listwise deletion, FIML, or traditional multiple imputation. If the missing data are NMAR, such as in the study described by Dallal (2012), then ITT analyses must be more thoughtful. These cases requires careful analysis of the causes of treatment dropout associated with missing data and then introducing statistical controls to turn NMAR properties into MAR properties. To be sure, ITT analyses likely can tolerate some degree of assumption violations of MAR without meaningful effects on conclusions, but care is required in this regard.

In the final analysis, you should make heroic efforts to obtain posttest data on treatment dropouts rather than making guesses about how dropouts would fare had you followed them up. ITT analyses are much more straightforward in cases of the former.

EXTENSIONS TO RANDOMIZED EXPLANATORY TRIALS

To make things manageable, my discussion of modern methods of efficacy and effectiveness analysis has focused only on links between the treatment condition and the outcome. In practice, RETs include the analysis of mediation, moderation and sometimes multiple outcomes. All of the methods I discussed can be applied to RETs in straightforward ways if one uses limited information SEM (LISEM, see Chapter 8), analyzing the equations implied by an influence diagram one equation at a time. For full information SEM (FISEM), use of FIML and multiple imputation is straightforward using Mplus (see Chapter 26) although one must now think about one's entire model when making choices about auxiliary variables and which covariates to use for each equation. G estimation and targeted maximum likelihood are not amenable to FISEM. IPTW can be used in FISEM but, again, the choice of variables to use in the weighting process must be done with care and with the full model in mind. One would define the list of variables to use in the weighting process using all relevant covariates across all the equations.

CONCLUDING COMMENTS

Treatment dropouts and non-adherence to treatment protocols are common in everyday life just as they are common in randomized trials. This does not mean that the frequency and nature of dropping out of treatment and non-adherence are equivalent in the two settings. Participating in a randomized trial and all that goes with it is not necessarily the same as being exposed to treatments or being offered treatment in everyday life. In a randomized trial, people often complete questionnaires that they would not normally complete, they know they are participating in an experiment that ultimately can impact both them and larger communities, and they often are compensated for their efforts. They have contact with and can form relationships with trial staff. They sometimes are given incentives to remain in the study and are given explicit instructions to participate in ways that respect study and treatment protocols. The fact that dropping out of treatment and non-adherence occur in both real life settings and randomized trials has led some researchers to naively think that knowledge gains require that analyses of randomized trials embrace the presence of dropping out of treatment and non-adherence by using ITT analyses. I have had reviewers stubbornly demand ITT analyses and reject per protocol analyses even when my trial is efficacy focused, when I am trying to understand the determinants of efficacy (not effectiveness) and when I seek to explore the generalizability of the mechanisms of efficacy (not effectiveness) across subgroups and contexts. Inevitably, reviewer insistence comes down to statements that ITT preserves randomization, hence the need to analyze data using it. This conclusion lets methodology dictate the questions we ask rather than letting the questions we ask dictate the methods

we use to answer those questions.

Clinical trials typically ask questions about mediation (mechanisms), moderation (the generalizability of the intervention across populations and contexts), neither mediation nor moderation (an outcome only trial), or both mediation and moderation in the same trial. The outcome in any of the above four types of trials can be either (1) patient adherence, (2) clinic/provider implementation faithfulness, (3) a clinical or behavioral state (such as depression, anxiety or whether a vaccine was obtained), or (4) some combination of the above. Researchers can approach the above two facets from either (1) an efficacy orientation (where an intervention, no matter what its purpose and focus, is implemented as it was intended to be implemented, i.e., per protocol) or (2) an effectiveness orientation (where a treatment is implemented in real world settings where it may not be implemented faithfully and where patients may or may not do what they are supposed to do). In the big scheme of things, if you have adopted an efficacy focus, it does not matter what the other facets of your study are (be it mediation, moderation, an outcome focused on patient adherence, an outcome focused on anxiety, or whatever), you need to pursue per protocol analyses. By contrast, if you have an effectiveness focus, it does not matter what the other facets are, you need to pursue ITT analyses. If you are interested in both efficacy and effectiveness, you need to do both.

A problem in intervention science is that many people seem to think ITT analysis is the appropriate tool for all of the different research combinations, which is not the case. I argue that ITT analyses are not always appropriate; that per protocol analyses have been unfairly denigrated and this is unfortunate because there are indeed viable (but not perfect) modern methods of per protocol analysis that can be used.

I have argued in this chapter that both efficacy trials and effectiveness trials have their rightful place in program development, program revision, and program evaluation. Hybrid designs that explore both efficacy and effectiveness are desirable but they also are challenging given that (a) there often is misalignment between the dynamics of stopping treatment/non-adherence in real life settings as opposed to a randomized trial setting and (b) efficacy analysis requires high levels of adherence and treatment completion.

Good program evaluation means providing feedback and suggested revisions to program administrators about both the efficacy of their program per se as well adherence and dropping out dynamics relative to their program. Fortunately, trialists have evolved a host of modern analytic methods that allow us to gain perspectives on these matters. Coupled with the incorporation of mediation and moderation dynamics into program evaluations and thoughtful RET design, these methods allow us to raise the bar considerably on the quality of evaluation efforts.

To me, the dynamics of analyzing missing data are somewhat different if the source

of that missingness is dropping out of a treatment as opposed to failing to show up for a scheduled post-treatment assessment session. The variables that impact these two types of events can be different, meaning that the covariates we address to deal with NMAR missingness can differ depending on the context. As I stated at the outset of Chapter 26, the best strategy for dealing with missing data is not to have any and that holds as much for dealing with missing data due to treatment dropouts as it does to other missing data scenarios.

APPENDIX: DETAILED CACE OUTPUT

In this appendix I review in more detail output for CACE modeling when programmed using Mplus as well as additional programming strategy. I use the JOBS numerical example from the main text. The JOBS output provides several indices to evaluate model fit, although many of the traditional fit indices are not supported by Mplus. The output reports the model loglikelihood, the model AIC, and the model BIC:

```
MODEL FIT INFORMATION

Number of Free Parameters                       17

Loglikelihood

          H0 Value                          -727.379
          H0 Scaling Correction Factor        1.0432
            for MLR

Information Criteria

          Akaike (AIC)                      1488.759
          Bayesian (BIC)                    1560.475
          Sample-Size Adjusted BIC          1506.515
            (n* = (n + 2) / 24)
```

The BIC can be used to compare nested models using the methods discussed in Chapter 7. An example might be if you want to compare a model that imposes equality constraints for path coefficients across classes versus one that does not.

In the output section called `RESIDUAL OUTPUT`, Mplus reports the differences between the predicted and observed covariances of the input variables. These differences also can give you a sense of model fit, although some people find covariances difficult to interpret. Mplus also reports modification indices for each class, which can be diagnostic of ill fit at a localized level. See Chapter 7 for details.

When determining the class that a person belongs to (in this case, complier vs. non-complier) Mplus calculates a probability that the person is in each of the two classes. For example, the probability individual 1 is a complier might be estimated to be 0.92 and the probability that individual 1 is a non-complier might be estimated to be 0.08. Thus, there is a degree of uncertainty about whether a person is in or not in a given class. This uncertainty is taken into account in the overall statistical treatment of the data by the underlying algorithms. Mplus reports three different ways of estimating the proportion of people in each class, as follows:

```
FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THE ESTIMATED MODEL

    Latent
   Classes

       1        271.06984          0.53998
       2        230.93016          0.46002

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON ESTIMATED POSTERIOR PROBABILITIES

    Latent
   Classes

       1        270.83819          0.53952
       2        231.16181          0.46048

FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES
BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP

Class Counts and Proportions

    Latent
   Classes

       1              268          0.53386
       2              234          0.46614
```

The first set of estimates are based on the final fitted model and are reported in the section FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED ON THE ESTIMATED MODEL. The second set of estimates are based on Bayesian methods and are reported in the section FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED ON ESTIMATED POSTERIOR PROBABILITIES. The third set of estimates assigns people to the class that the person has the highest estimated probability of being in. These estimates are reported in the section FINAL CLASS COUNTS AND PROPORTIONS FOR THE LATENT CLASSES BASED ON THEIR MOST LIKELY LATENT CLASS MEMBERSHIP. In the JOBS data, all three estimates were quite close so it is moot which one is reported. I tend to prefer the estimated model estimates, but convincing arguments can be made for each type.

Mplus also provides information about the quality of the classification enterprise, which is used by some as an indirect index of model fit. If the indices reflecting classification quality are ill-behaved, it raises doubts about the model. One diagnostic forms a two-way table that has as rows the class a person is most likely to be in based on

the one with the highest probability for that person and as columns the mean probability result for class 1 (compliers) and the mean probability for class 2 (non-compliers):

```
Average Latent Class Probabilities for Most Likely Latent Class Membership
(Row) by Latent Class (Column)

           1        2

    1   0.916    0.084
    2   0.108    0.892
```

For example, of those individuals whose most likely class was class 1 (complier), their mean probability of being in class 1 was 0.916 and their mean probability of being in class 2 was 0.084. There is good separation between the two classes for these individuals. Of those individuals whose most likely class was class 2 (non-complier), their mean probability of being in class 1 was 0.108 and their mean probability of being in class 2 was 0.892. There also was good separation between the two classes for these individuals. This increases our confidence in the viability of the classification dynamics. This type of matrix is sometimes referred to as a **confusion matrix**.

The output also provides an index of classification quality using an entropy index:

```
CLASSIFICATION QUALITY

     Entropy                         0.734
```

The statistic is a summary of how well differentiated the confusion matrix is. It ranges from 0 to 1, with higher scores indicating more unambiguous classifications (i.e., the closer the value to 1, the better). Values greater than 0.80 are deemed good, but there is controversy about this (e.g., Ramaswamy et al., 1993); lower values are sometimes fine. For more details, see Chapter X.

Mplus output also reports within each class the squared multiple correlation for predicting the outcome (depression) from the treatment condition dummy variable and the covariates (baseline depression, education, economic hardship, and ethnicity). These are located in the section `STANDARDIZED MODEL RESULTS` and the subsection `STDYX Standardization`:

```
R-SQUARE

Class 1

    Observed                                        Two-Tailed
```

```
    Variable        Estimate       S.E.     Est./S.E.  P-Value
    DEPRESS            0.174      0.036      4.798      0.000
Class 2

    Observed                                           Two-Tailed
    Variable        Estimate       S.E.     Est./S.E.  P-Value

    DEPRESS            0.172      0.030      5.662      0.000
```

The squared correlation in Class 1 was 0.174 and in Class 2 it was 0.172.

In the main text, I reported the results of a binary logistic regression where the outcome was complier class 1 versus complier class 2 and the predictors were age, education motivation, economic hardship, assertiveness, marital status, ethnicity and baseline depression. I repeat the odds ratios here:

```
LOGISTIC REGRESSION ODDS RATIO RESULTS

                                                 95% C.I.
                    Estimate       S.E. Lower 2.5% Upper 2.5%

Categorical Latent Variables

 C#1      ON
    AGE                1.082      0.016      1.051      1.115
    EDUC               1.360      0.093      1.190      1.556
    MOTIVATE           1.958      0.308      1.439      2.664
    ECON               0.851      0.128      0.634      1.142
    ASSERT             0.685      0.099      0.516      0.911
    SINGLE             1.673      0.469      0.966      2.896
    NONWHITE           0.588      0.190      0.312      1.107
    DEPBASE            0.654      0.272      0.290      1.477
```

As discussed in Chapter 12, I prefer to work with probabilities and profile analyses instead of odds ratios and I use the `MODEL CONSTRAINT` command in Mplus to do so. Consider the motivation predictor, namely the motivation to secure employment as measured at baseline. This scale ranged from 1 to 7 with the bulk of people scoring values of 4, 5 or 6. I used `MODEL CONSTRAINT` commands to explore this predictor while holding the other variables constant at values close to their modal values. For the syntax in Table 27.9, I first added labels to the intercept and coefficients in the logit model on Line 14, like this:

```
c#1 ON age educ motivate econ assert single nonwhite depbase (lc1-lc8) ;
[c#1] (ic1) ;
```

Then I added the following syntax just before the `OUTPUT` command:

```
22a. MODEL CONSTRAINT:
22b. NEW(LODDS1 LODDS2 LODDS3 PROB1 PROB2 PROB3 DIFF23 DIFF24 ) ;
22c. LODDS1 = ic1+lc1*37+lc2*13+lc3*4+lc4*3+lc5*3+lc6*1+lc7*1+lc8*2 ;
22d. LODDS2 = ic1+lc1*37+lc2*13+lc3*5+lc4*3+lc5*3+lc6*1+lc7*1+lc8*2 ;
22e. LODDS3 = ic1+lc1*37+lc2*13+lc3*6+lc4*3+lc5*3+lc6*1+lc7*1+lc8*2 ;
22f. PROB1 = EXP(LODDS1)/(1+EXP(LODDS1)) ;
22g. PROB2 = EXP(LODDS2)/(1+EXP(LODDS2)) ;
22h. PROB3 = EXP(LODDS3)/(1+EXP(LODDS3)) ;
22i. DIFF23 = PROB2 - PROB1 ;
22j. DIFF24 = PROB3 - PROB1 ;
```

Note in lines 22c, 22d, and 22e, I calculate the predicted log odds for three different profiles that vary the value of the motivation predictor to be 4, 5, and 6 while holding the values of the other predictors constant at their modal values (I can use any profile values I want and that are of substantive interest). Lines 22f, 22g, and 22h convert these predicted log odds to probabilities and Lines 22i and 22j calculate substantively interesting differences between the predicted probabilities. (For the broader logic of this type of profile analysis, see Chapter 12). Here is the resulting output:

```
MODEL RESULTS

                                                    Two-Tailed
                  Estimate       S.E.    Est./S.E.  P-Value

New/Additional Parameters
    LODDS1          -0.676      0.425     -1.588      0.112
    LODDS2          -0.004      0.365     -0.010      0.992
    LODDS3           0.668      0.368      1.817      0.069
    PROB1            0.337      0.095      3.546      0.000
    PROB2            0.499      0.091      5.464      0.000
    PROB3            0.661      0.082      8.023      0.000
    DIFF23           0.162      0.035      4.690      0.000
    DIFF24           0.324      0.070      4.609      0.000
```

The predicted probability of being in the complier class was 0.337 ±0.20 when motivation = 4, it was 0.499 ±0.18 when motivation = 5, and it was 0.661 ±0.16 when motivation = 6. If I increase baseline motivation from 4 to 5, the proportion of people who become compliers is estimated to increase by 0.162 (± 0.07, CR = 4.69, $p < 0.01$). If I increase baseline motivation from 4 to 6, the proportion of people who become compliers is estimated to increase by 0.324 (± 0.14, CR = 4.61, $p < 0.01$). This suggests that one way of increasing compliance with the protocol (e.g., attending the four

seminars) is to increase motivation to be re-employed at baseline. I find this type of profile analysis to be more meaningful than simply documenting odds ratios.