Rejoinder

Donald B. Rubin

It is a pleasure to have the opportunity to respond to these three complementary discussions written by Edward L. Korn (ELK), Paul R. Rosenbaum (PRR) and Stephen E. Fienberg (SEF), each of whom has made substantial contributions to problems of causal inference. Thanks to all three for the generosity expressed in their comments. I am in agreement with many points raised by them, but there are a few places where we may not fully agree, possibly due to misunderstandings.

ELK

ELK's ordering of problems by their importance, displayed in his Figure 1, is difficult to dispute. I like to think of there being a continuum in causal inference from perfect randomized experiments to sloppy observational studies, and I like to "chip away" at all points in the continuum because I often feel that there is potentially a tremendous transfer of insights from work on one problem to work on another. For a specific example, the work on noncompliance presented in Angrist, Imbens and Rubin (1996) led to the general principal stratification framework in Frangakis and Rubin (2002), which provided a formal structure for embedding the resolution of the "censoring/truncation due to death" problem used a decade earlier in the context of an actual FDA submission, as noted in Rubin (1998, 2000) and the target article.

ELK's examples are highly appropriate and offer strong support for the importance of working in the context of real problems. I particularly liked his orthodontic example from Korn, Teeter and Baumrind (2001), and think that generalizations of the idea used there could be quite fruitful, especially because that idea implies specific suggestions for improvements to the design of particular types of observational studies. ELK is correct that when the principal strata are effectively observed, as in that example, distributional assumptions, such as normality, can be avoided.

As pointed out in Angrist, Imbens and Rubin (1996) and its rejoinder, the focus of our analysis of noncompliance is on the units in this experiment who would comply when assigned either treatment. This group is,

by definition, the only collection of units in this experiment who can be observed receiving and not receiving the treatment, and thus is the only collection of units providing any data about the causal effect of receiving versus not receiving the treatment. The always-takers are always observed receiving the treatment, and the never-takers are never observed receiving the treatment, and thus data from neither of these latter two groups provides any evidence about the effect of taking versus not taking the treatment. In other words, in this experiment "efficacy" can only be estimated for the compliers. "Effectiveness" is a joint property of (a) efficacy for the compliers, (b) how the treatment is "marketed" (i.e., how compliance is enforced) and (c) "placebo" effects of assignment on the noncompliers (i.e., the always-takers and the never-takers). This point is discussed at length in Sheiner and Rubin (1995).

Consequently, the emphasis within the principal stratification framework is on separately estimating efficacy and the other components of effectiveness, in order to help the generalization to other situations with possibly different placebo effects or different marketing effects. Thus I am puzzled by ELK's statement that the principal stratification approach "…negates one of the usual reasons for being interested in efficacy and not effectiveness…"—quite the opposite in my mind, but perhaps I misunderstood his meaning.

PRR

I do understand ELK's desire to avoid the entire problem of censoring due to death in the quality-of-life example by assigning the lowest possible QOL score to those who are dead, but as I have argued, this approach, to me, mixes up issues of (a) estimating the scientific effect of a treatment intervention in groups of units where it can estimated and (b) individual value judgments about the value of death versus various qualities of life. PRR seems to agree, even providing a Seneca quote in support! If we accept ELK's suggestion to use, for example, a rank-sum test, it seems an approach such as the one advocated by PRR is quite attractive because it avoids having to make a particular choice of a single value for the QOL of someone who is dead, common