*The Annals of Applied Statistics*
2013, Vol. 7, No. 4, 2336–2360
DOI: 10.1214/13-AOAS674
© Institute of Mathematical Statistics, 2013

# EXPLOITING MULTIPLE OUTCOMES IN BAYESIAN PRINCIPAL STRATIFICATION ANALYSIS WITH APPLICATION TO THE EVALUATION OF A JOB TRAINING PROGRAM

BY ALESSANDRA MATTEI[1,*], FAN LI[2,†] AND FABRIZIA MEALLI[1,*]

*University of Florence* and Duke University†*

The causal effect of a randomized job training program, the JOBS II study, on trainees' depression is evaluated. Principal stratification is used to deal with noncompliance to the assigned treatment. Due to the latent nature of the principal strata, strong structural assumptions are often invoked to identify principal causal effects. Alternatively, distributional assumptions may be invoked using a model-based approach. These often lead to weakly identified models with substantial regions of flatness in the posterior distribution of the causal effects. Information on multiple outcomes is routinely collected in practice, but is rarely used to improve inference. This article develops a Bayesian approach to exploit multivariate outcomes to sharpen inferences in weakly identified principal stratification models. We show that inference for the causal effect on depression is significantly improved by using the re-employment status as a secondary outcome in the JOBS II study. Simulation studies are also performed to illustrate the potential gains in the estimation of principal causal effects from jointly modeling more than one outcome. This approach can also be used to assess plausibility of structural assumptions and sensitivity to deviations from these structural assumptions. Two model checking procedures via posterior predictive checks are also discussed.

**1. Introduction.** The impact of job loss and unemployment on workers' stress and mental health is a subject of much interest in psychology [see, e.g., Vinokur, Caplan and Williams (1987)]. The Job Search Intervention Study (JOBS II) [Vinokur, Price and Schul (1995)] is an influential randomized field experiment intended to study the prevention of poor mental health and the promotion of high-quality re-employment among unemployed workers. In JOBS II, participants were randomly assigned to attending job training seminars (treatment) or receiving a booklet on job-search tips (control). As in many open-label randomized intervention studies, substantial noncompliance to assigned treatment arose in JOBS II. The compliance status is a special case of intermediate variables, that is, variables, often confounded, that are potentially affected by the assignment and also affect the response. When the study goal, as in JOBS II, is to evaluate the

Received November 2012; revised May 2013.

[1]Supported in part by the Futuro in Ricerca Grant RBFR12SHVV_003 (B11J12002670001) financed by the Italian Ministero dell'Istruzione, Università e Ricerca.

[2]Supported in part by NSF Grant SES-1155697.

*Key words and phrases.* Bayesian, causal inference, intermediate variables, job training program, mixture, multivariate outcomes, noncompliance, principal stratification.

causal effect of receiving the treatment rather than the effect of assignment, the confounded intermediate variables need to be adjusted for in the analysis. Another example where intermediate variables arise is mediation analysis in observational studies: researchers are interested in knowing not only if an exposure has an effect on the response, but also to what extent this effect is mediated by some variables on the causal pathway between exposure and outcome. Other forms of intermediate variables include surrogate endpoints, unintended missing outcome data, truncation of outcome by "death" and combinations of these variables.

Our discussion will frame causal inference with intermediate variables in the context of the Rubin Causal Model (RCM) using potential outcomes [Rubin (1974, 1978)]. Under the RCM, a causal effect is defined as the comparison between the potential outcomes under different treatments on a *common set* of units. As pointed out in Rosenbaum (1984), directly applying standard pretreatment variable adjustment methods, such as regression analysis, to intermediate variables generally results in estimates lacking causal interpretation. Angrist, Imbens and Rubin (1996) and Imbens and Rubin (1997) focused on noncompliance in randomized trials and made connections with econometric instrumental variable (IV) settings: they stratify units into latent subpopulations according to their joint potential compliance statuses under both treatment and control. This is a special case of the later developed principal stratification (PS) [Frangakis and Rubin (2002)], an increasingly popular framework for handling intermediate variables. A PS with respect to an intermediate variable is a cross-classification of units into latent classes defined by the joint potential values of that intermediate variable under each of the treatments being compared. A principal stratum consists of units having the same joint intermediate potential outcomes and so is not affected by treatment assignment. Therefore, comparisons of potential outcomes under different treatment levels within a principal stratum—the principal causal effects (PCEs)—are well-defined causal effects in the sense of Rubin (1978).

However, since at most one potential outcome is observed for any unit, we cannot, in general, observe the principal stratum to which a unit belongs, so that inference on PCEs is not straightforward. There are two streams of work in the existing literature regarding this: (1) deriving large-sample nonparametric bounds for the causal effects under minimal structural assumptions [e.g., Manski (1990), Zhang and Rubin (2003)], and (2) specifying additional structural (e.g., exclusion restriction or monotonicity) or modeling assumptions to infer PCEs, and conducting sensitivity analysis to check the consequences of violations of such assumptions [e.g., Ten Have et al. (2004), Small and Cheng (2009), Sjölander et al. (2009), Elliott, Raghunathan and Li (2010), Li, Taylor and Elliott (2010, 2011), Schwartz, Li and Reiter (2012)]. In this article we introduce an alternative approach to improve estimation of PCEs, which uses multiple outcomes in a model-based analysis. For example, in the JOBS II evaluation, we will jointly model the depression score, the outcome of primary interest, and the re-employment status, a secondary outcome, to sharpen the inference for the causal effect on depression.

Multivariate analysis is beneficial for two reasons. First, models used in PS are inherently mixture models; recent results on mixture models show that with correct model specification, multivariate analysis leads to smaller variances of the parameters' estimators than those from a corresponding univariate analysis, resulting in more precise estimates [Mercatanti, Li and Mealli (2012)]. Second, some key substantive structural assumptions, such as exclusion restrictions, may be more plausible for secondary outcomes than for the primary one. For example, in JOBS II, due to the possible "placebo effect," exclusion restriction might not be plausible for depression, but it may be more plausible for re-employment status. Another example is given in Section 2. Restrictions on secondary outcomes reduce the parameter space of the joint distribution of all outcomes and, in turn, the marginal distribution of the primary one [Mealli and Pacini (2013)].

However, the additional information provided by secondary outcomes is obtained at the cost of having to specify more complex multivariate models, which may increase the possibility of misspecification. For instance, in the analysis of JOBS II data, jointly modeling depression and re-employment status involves specifying a mixture of two underlying bivariate normal distributions, increasing the number of unknown parameters compared with a univariate analysis on depression. Therefore, model diagnostics are crucial in the multivariate analysis and we develop model checking procedures via posterior predictive checks in this article.

While the use of auxiliary information from covariates to improve inference on causal effects has been discussed, the importance of exploiting multiple outcomes is less acknowledged. For example, covariates generally improve inference on causal effects by enhancing the prediction of missing intermediate and final potential outcomes [e.g., Gilbert and Hudgens (2008), Hirano et al. (2000)]. However, information on multiple outcomes is routinely collected in randomized experiments and observational studies, but is rarely used in analysis unless the goal is to study the relationships between outcomes. One exception is Jo and Muthen (2001), who demonstrated, in the context of a randomized trial with noncompliance, that a joint analysis with two outcomes outperforms the two corresponding univariate analyses. Mealli and Pacini (2013) showed that using the joint distribution of a primary outcome and an auxiliary variable (a secondary outcome or a covariate) in randomized experiments with noncompliance can tighten large-sample nonparametric bounds for PCEs.

Our work is closely related to Mealli and Pacini (2013), but it proceeds from the parametric perspective under the Bayesian paradigm instead. As causal inference problems are essentially missing data problems under the RCM, Bayesian approaches appear to be particularly useful. From a Bayesian perspective, all unknown quantities, parameters as well as unobserved potential outcomes, are random variables with a joint posterior distribution, conditional on the observed data. Therefore, inferences are based on the posterior distribution of the causal estimands defined as functions of observed and unobserved potential outcomes, or sometimes as functions of model parameters. This leads to at least two inferential

advantages. First, the Bayesian approach provides a refined map of identifiability, clarifying what can be learned when causal estimands are intrinsically not fully identified, but only weakly identified in the sense that their posterior distributions have substantial regions of flatness [Imbens and Rubin (1997)]. In particular, issues of identification are different from those in the frequentist paradigm because with proper prior distributions, posterior distributions are always proper. Weak identifiability is reflected in the flatness of the posterior distribution and can be quantitatively evaluated [Gustafson (2009)]. Second, in a Bayesian setting, the effect of relaxing or maintaining assumptions can be directly checked by examining how the posterior distributions for causal estimands change, therefore serving as a natural framework for sensitivity analysis. Moreover, the Bayesian framework allows one to quantify the impact on the causal estimates when there is a diversion from these assumptions.

The primary aim of the paper is to combine the benefits from using a multivariate analysis with the inferential advantages of the Bayesian approach for causal inference in the context of principal stratification. The rest of the article is organized as follows. Section 2 introduces the fundamentals of principal stratification and the intuition for the benefit from using a multivariate analysis. In Section 3 we propose Bayesian bivariate models for principal stratification analyses and describe the details of conducting posterior inferences for the causal effects. In Section 4 we reanalyze the JOBS II study using the proposed bivariate approach. Additional simulation studies to examine the benefits to use multivariate outcomes under various scenarios are carried out in Section 5. Two model checking procedures based on posterior predictive checks with application to the JOBS II data are discussed in Section 6. Section 7 concludes.

## 2. Fundamentals.

2.1. *Basic setup, definitions and assumptions.* Consider a large population of units, each of which can potentially be assigned a treatment indicated by $z$, with $z = 1$ for treatment and $z = 0$ for control. A random sample of $n$ units from this population comprises the participants in a study, designed to evaluate the effect of $Z$ on all or a subset of $M$ outcomes $\mathbf{Y} = (Y_1, \ldots, Y_M)'$. Without loss of generality, we will focus on the case of two outcomes ($M = 2$). For each unit $i$, let $Z_i$ be the assignment indicator with $Z_i = 1$ indicating the unit is assigned to the treatment and $Z_i = 0$ to the control. After the assignment, but before the outcome is observed, an intermediate outcome $D_i$ is also observed. In the JOBS II evaluation, both $Z$ and $D$ are binary, with $Z_i = 1$ and 0 denoting random assignment to the job training seminars and to the booklet, respectively, and $D_i = 1$ and 0 denoting actually attending the seminars or not, respectively. Also, $Y_1$ denotes the depression score and $Y_2$ denotes the re-employment status.

Assuming the standard Stable Unit Treatment Value Assumption [SUTVA, Rubin (1980)], for each outcome $Y_m$, we can define for each unit $i$ two potential outcomes, $Y_{im}(0)$ and $Y_{im}(1)$, corresponding to each of the two possible treat-

ment levels. Under the RCM, a causal effect of the treatment $Z$ on the outcome $Y_m$ is defined as a comparison of the potential outcomes $Y_m(1)$ and $Y_m(0)$ on a common set of units. However, only one potential outcome is observed for unit $i$, $Y_{im}^{\text{obs}} = Y_{im}(Z_i)$; the other potential outcome, $Y_{im}^{\text{mis}} = Y_{im}(1 - Z_i)$, is missing. Therefore, causal inference problems under the RCM are inherently missing data problems.

Since an intermediate variable, $D$, is a post-treatment variable, we can also define two potential outcomes $D_i(0)$ and $D_i(1)$ for each unit, with one being observed, $D_i^{\text{obs}} = D_i(Z_i)$, and one missing, $D_i^{\text{mis}} = D_i(1 - Z_i)$. Comparing outcomes from units with the same values of $D^{\text{obs}}$ between treatments generally leads to estimates lacking causal interpretation, because then the sets $\{i : D_i^{\text{obs}} = d, Z_i = 1\}$ and $\{i : D_i^{\text{obs}} = d, Z_i = 0\}$ are generally not the same groups of units. This concern is known as the post-treatment selection bias.

A principal stratification with respect to the post-treatment variable $D$ is a partition of units, whose sets—principal strata—are defined by the joint potential values of $D$: $S_i = (D_i(0), D_i(1))$. By definition, the principal stratum membership $S_i$ is not affected by the assignment. Therefore, comparisons of $Y_m(1)$ and $Y_m(0)$ within a principal stratum, the principal causal effects (PCEs), have a causal interpretation because they compare quantities defined on a common set of units. However, since $D_i(0)$ and $D_i(1)$ are never jointly observed, principal stratum $S_i$, which a unit $i$ belongs to, is, in general, only partially observed.

To convey the main message of utilizing multiple outcomes, we focus on the simple case of a binary intermediate variable, as is the case in JOBS II; it is nevertheless straightforward to apply the method developed here to multi-valued or continuous intermediate variables following the approaches in Jin and Rubin (2008) and Schwartz, Li and Mealli (2011). In order to highlight the role of additional outcomes, with no loss of generality, our discussion does not include covariates, although covariates can be easily included in the analysis. With a binary treatment and a binary intermediate variable, there are at most four principal strata: $S_i \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. When $D$ is the indicator of the treatment actually received, as in our JOBS II application, the four principal strata are, respectively, called never-takers ($S_i = n$), compliers ($S_i = c$), defiers ($S_i = d$) and always-takers ($S_i = a$). Though our approach applies to any binary intermediate variable settings (e.g., mediation, truncation by death), we use the familiar nomenclature of noncompliance to generically refer to $S_i$ hereafter for simplicity.

In randomized studies with noncompliance, the presence of defiers is usually ruled out assuming monotonicity of noncompliance: $D_i(1) \geq D_i(0)$ for all $i$, with inequality for at least one unit. Although often plausible in experimental studies with noncompliance, monotonicity is a substantive assumption that may not always be satisfied in other settings. An important advantage of Bayesian causal inference, in general, and our Bayesian analysis, in particular, is that the monotonicity assumption is not necessary and, consequently, violation to this assumption could be easily addressed [Imbens and Rubin (1997)].

In the JOBS II study the treatment is only accessible to the $Z_i = 1$ group, so $D_i(0) = 0$ for all $i$. Therefore, subjects who would have taken the treatment if assigned to control (defiers and always-takers) are denied to access the treatment if assigned to control and, thus, units are classified, in this experiment, only by the values of $D_i(1)$: $D_i(1) = 1$ if unit $i$ is a complier, and $D_i(1) = 0$ if unit $i$ is a never-taker. This is a typical case of one-sided noncompliance [e.g., Mattei and Mealli (2007), Sommer and Zeger (1991)].

The causal estimand of interest in this article is the population-average principal causal effect for the *first* outcome:

$$\tau_s = \mathbb{E}\big(Y_{i1}(1) - Y_{i1}(0) | S_i = s\big) \tag{1}$$

for $s = c, n$. In JOBS II, $\tau_s$ corresponds to the causal effect of being assigned to a job-search seminar on depression for compliers ($s = c$) and never-takers ($s = n$). By focusing on the population-average estimands, we can ignore the association between $Y_{i1}(0)$ and $Y_{i1}(1)$ in the analysis.[3] Depending on the models for the potential outcomes, population estimands are usually functions of more than one model parameter.

Throughout the paper, we assume that the treatment is randomly assigned, as in JOBS II. Let $p(\cdot)$ and $p(\cdot|\cdot)$ denote probability or probability density and conditional probability or conditional probability density, respectively, depending on the context.

ASSUMPTION 1 (Randomization of treatment assignment).

$$p\big(Z_i | \mathbf{Y}_i(0), \mathbf{Y}_i(1), D_i(0), D_i(1)\big) = p(Z_i).$$

Randomization implies that the joint distribution of the five quantities associated with each sampled unit, $(Z_i, \mathbf{Y}_i(0), \mathbf{Y}_i(1), D_i(0), D_i(1))$, can be decomposed into

$$p\big(\mathbf{Y}_i(0), \mathbf{Y}_i(1), D_i(0), D_i(1), Z_i\big) = p\big(\mathbf{Y}_i(0), \mathbf{Y}_i(1) | S_i\big) p(S_i) p(Z_i). \tag{2}$$

Randomization allows us to ignore $p(Z_i)$. This implies that likelihood or Bayesian model-based approaches to PS analysis usually involve two sets of models: (1) models for the distribution of potential outcomes conditional on the principal strata, and (2) models for the distribution of principal strata.

[3]Distinct from the corresponding finite-sample estimands, $\tau_s^{\mathrm{FS}} = \sum_{i:S_i=s}\{Y_{i1}(1) - Y_{i1}(0)\}/n_s$, the population causal effects (1) do not depend on the association parameters between $Y_{i1}(0)$ and $Y_{i1}(1)$, say, $\rho$. Specifically, posterior distribution of the population estimands $\tau_s$ will not be dependent of $\rho$ as long as $\rho$ is a priori independent of the remaining model parameters, while inferences for the finite sample causal estimands $\tau_s^{\mathrm{FS}}$ would generally involve $\rho$ regardless of the prior structure between parameters [for more discussion on this, see page 311 in Imbens and Rubin (1997)].

2.2. *Intuition for sharping inference from multiple outcomes*. The intuition for the benefit of jointly analyzing multiple outcomes in PS analysis is as follows.

Principal strata are inherently latent clusters. Intuitively, proper utilization of auxiliary variables provides extra dimensions to better predict the component membership and disentangle the mixtures. First, additional outcomes serve as additional predictors of principal strata membership from the outcome models. To see this, take, for example, the model for two potential outcomes under $z = 0$. By the Bayes rule, $p(Y_{i1}(0), Y_{i2}(0)|S_i) \propto p(S_i|Y_{i1}(0), Y_{i2}(0))\ p(Y_{i2}(0), Y_{i1}(0))$. Comparing to the univariate model with $Y_1$, where $p(Y_{i1}(0)|S_i) \propto p(S_i|Y_{i1}(0))\ p(Y_{i1}(0))$, it is clear to see the role of the second outcome $Y_2$ as an additional predictor of $S_i$.

As a second intuition, two (or more) distributions may be difficult to disentangle if they are similar, for example, if their means are very close; these same two means may instead be very far apart (and thus the mixture easier to disentangle) if considered in a two-dimensional space. In fact, recent theoretical results for mixture models [Mercatanti, Li and Mealli (2012)] show that, given correct model specification, the probability of correctly allocating the cluster membership of the units and the information number for the means of the primary outcome in a bivariate mixture model are generally larger than those in the corresponding marginal model. As a result, variances of the maximum likelihood estimators for the mixture means, estimated by the inverse of the observed information matrix, are generally smaller in a bivariate analysis than in a univariate one.

As a third intuition, some structural assumptions may be more plausible for the secondary outcome than the primary outcome. For example, stochastic exclusion restriction (ER) for never-takers is commonly assumed to point-identify PCEs:

ASSUMPTION 2 (Stochastic exclusion restriction for never-takers).

$$p\big(Y_{im}(0)|S_i = n\big) = p\big(Y_{im}(1)|S_i = n\big), \qquad m = 1, 2.$$

The ER implies that any effect of the assignment is mediated through the intermediate variable. Under Assumption 2 $\tau_n = 0$ and $\mathbb{E}(Y_{i2}(1) - Y_{i2}(0)|S_i = n) = 0$. But the ER is often questionable in practice. Consider a double-blinded randomized trial with the primary goal of studying the efficacy of a new drug on a health outcome, where side effects are also recorded as a secondary outcome. Due to the placebo effect, the ER may not always hold for the primary outcome. Since side effects are usually only caused by taking the drug rather than the placebo, ER appears to be more likely to hold for side effects than the primary outcome. Formally, we have the "partial exclusion restriction (PER)" assumption [Mealli and Pacini (2013)]:

ASSUMPTION 3 (Stochastic partial exclusion restriction for never-takers).

$$p\big(Y_{i2}(0)|S_i = n\big) = p\big(Y_{i2}(1)|S_i = n\big).$$

Assumption 3 implies that $\mathbb{E}(Y_{i2}(1) - Y_{i2}(0)|S_i = n) = 0$, but the average causal effect for never-takers on the primary outcome, $\tau_n$, may differ from zero. Restrictions on the secondary outcome, such as PER, will reduce the parameter space of the joint distribution of the outcomes and in turn the marginal distribution of the primary one. PER can be combined with other conditions on the association structure between outcomes to improve inference about the causal estimates [Mealli and Pacini (2013)].

**3. Bayesian bivariate principal stratification analysis.** The structure for Bayesian PS inference was first developed in Imbens and Rubin (1997) for the special case of noncompliance. As discussed before, two sets of models need to be specified, as well as the prior distribution for the parameters, $\boldsymbol{\theta}$. Denote $\pi_{i,s} = p(S_i = s|\boldsymbol{\theta})$ and $f_{i,sz} = p(\mathbf{Y}_i(z)|S_i = s, \boldsymbol{\theta})$, for $s = c, n$ and $z = 0, 1$, and assume a prior distribution $p(\boldsymbol{\theta})$ for the parameters $\boldsymbol{\theta}$. The posterior distribution of $\boldsymbol{\theta}$ can be shown to be

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{Z}, \mathbf{X}) \propto p(\boldsymbol{\theta}) &\times \prod_{i:Z_i=1, D_i^{\text{obs}}=1} \pi_{i,c} f_{i,c1} \times \prod_{i:Z_i=1, D_i^{\text{obs}}=0} \pi_{i,n} f_{i,n1} \\
&\times \prod_{i:Z_i=0, D_i^{\text{obs}}=0} [\pi_{i,n} f_{i,n0} + \pi_{i,c} f_{i,c0}],
\end{aligned} \tag{3}
$$

where the sum in the likelihood is because the units with $(Z_i = 0, D_i^{\text{obs}} = 0)$ are mixture of never-takers and compliers. Direct posterior inference of $\boldsymbol{\theta}$ from (3) is made easier using data augmentation to impute the missing $D_i^{\text{mis}}$. Specifically, we can first obtain the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{D}^{\text{mis}})$ from a Gibbs sampler by iteratively sampling from $p(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{D}^{\text{mis}}, \mathbf{Z})$ and $p(\mathbf{D}^{\text{mis}}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{Z}, \boldsymbol{\theta})$, which in turn provides the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{Z})$ and thus the posterior of the causal estimands $\tau_s$, $s = c, n$. The key to the posterior computation is the evaluation of the complete intermediate-data posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{D}^{\text{mis}}, \mathbf{Z})$, which has the following simple form:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{D}^{\text{mis}}, \mathbf{X}, \mathbf{Z}) = \pi(\boldsymbol{\theta}) &\times \prod_{i:Z_i=1, S_i=c} \pi_{i,c} f_{i,c1} \\
&\times \prod_{i:Z_i=1, S_i=n} (1 - \pi_{i,c}) f_{i,n1} \times \prod_{i:Z_i=0, S_i=c} \pi_{i,c} f_{i,c0} \\
&\times \prod_{i:Z_i=0, S_i=n} (1 - \pi_{i,c}) f_{i,n0}.
\end{aligned}
$$

Without additional assumptions, such as ER, inference on $\tau_s$, though possible and relatively straightforward from a Bayesian perspective, can be very imprecise, even in large samples. We argue that jointly modeling multiple outcomes may help to reduce uncertainty about $\tau_s$ in cases where such assumptions are questionable.

**4. Application to the JOBS II study.** In JOBS II, before randomization, participants were divided into two groups defined by values of a risk variable depending on financial strain, assertiveness and depression scores. Subjects who had a risk score greater than a prefixed threshold were classified in the high-risk category. Subsequently, the low- and the high-risk participants were randomly assigned to a control condition or an experimental condition. The intervention consisted of 5 half-day job-search skills seminars, aimed at teaching participants the most effective strategies to get a suitable position and at improving their job-search skills. The control condition consisted of a mailed booklet briefly describing job-search methods and tips.

Previous studies have found that the job search intervention program had its primary impact on the high-risk group [e.g., Jo and Muthen (2001), Little and Yau (1998), Vinokur, Price and Schul (1995)], hence, our focus is on high-risk subjects. The sample we use consists of 398 high-risk individuals with nonmissing values on the relevant variables. We focus on the outcomes measured six months after the intervention assignment. The primary outcome of interest ($Y_1$) is depression, measured with a sub-scale of 11 items based on the Hopkins Symptom Checklist. As a secondary outcome ($Y_2$), we use re-employment, a binary variable taking on value 1 if a subject works for 20 hours or more per week.

Noncompliance arises in JOBS II because a substantial proportion (46%) of individuals invited to participate in the job-search seminar did not show up to the intervention. As mentioned before, the treatment condition is only available to the individuals assigned to the intervention in JOBS II, thus, by the strong monotonicity assumption, there are neither defiers nor always-takers in the data. Some summary statistics for the sample of 398 high-risk unemployed workers classified by assignment $Z_i$ and treatment received $D_i^{\text{obs}}$ are shown in Table 1.

Comparisons of outcomes conditional on the actual treatment status do not generally lead to credible estimates of the effect of the job-search seminar attendance. However, randomization of the assignment implies that a standard intention-to-treat (ITT) analysis, which compares units by assignment and neglects noncom-

TABLE 1
*Summary statistics (means), JOBS II data*

| | | | | $Z_i = 1$ | | |
|---|---|---|---|---|---|---|
| **Sample size** | **All 398** | $Z_i = 0$ **130** | $Z_i = 1$ **268** | $D_i^{\text{obs}} = 0$ **124** | $D_i^{\text{obs}} = 1$ **144** | $D_i^{\text{obs}} = 0$ **254** |
| Assignment ($Z_i$) | 0.67 | 0 | 1 | 1 | 1 | 0.49 |
| Job-search seminar ($D_i^{\text{obs}}$) | 0.36 | 0 | 0.54 | 0 | 1 | 0 |
| Depression ($Y_{i1}^{\text{obs}}$) | 2.06 | 2.15 | 2.01 | 2.08 | 1.96 | 2.11 |
| Re-employment ($Y_{i2}^{\text{obs}}$) | 0.60 | 0.55 | 0.63 | 0.59 | 0.66 | 0.57 |

pliance, leads to valid inference on the causal effect of assignment. Under monotonicity and ER for noncompliers (never-takers), the ITT effect is proportional to the PCE effect for the subpopulation of compliers ($\tau_c$). Therefore, the ITT effect can be interpreted as indicative of the effect of the treatment, although the attribution of the PCE for compliers to the causal effect of the treatment for compliers is an assumption.

In JOBS II, assuming ER for depression may be controversial. For example, never-takers randomized to the intervention might feel demoralized by the inability to take advantage of the opportunity, whereas they would be less demoralized when randomized to the control group because the intervention was never offered. Therefore, we relax ER for depression, using information on a secondary outcome—re-employment status—to improve the estimation of weakly identified causal effects on depression.

*Models.* We assume a bivariate normal outcome model for the logarithm of depression ($Y_1$) and a latent variable $Y_{i2}^*$ underlying the binary re-employment status: $Y_{i2}(z) = \mathbf{1}(Y_{i2}^*(z) > 0)$. Specifically, for $s = c, n$ and $z = 0, 1$,

$$\begin{pmatrix} Y_{i1}(z) \\ Y_{i2}^*(z) \end{pmatrix} \Big| S_i = s \sim \mathrm{N}\left(\boldsymbol{\mu}^{s,z} = \begin{pmatrix} \mu_1^{s,z} \\ \mu_2^{s,z} \end{pmatrix}, \boldsymbol{\Sigma}^{s,z} = \begin{pmatrix} \sigma_{11}^{s,z} & \sigma_{12}^{s,z} \\ \sigma_{12}^{s,z} & \sigma_{22}^{s,z} \end{pmatrix}\right), \tag{4}$$

with $\sigma_{22}^{s,z} = 1$. This formulation is equivalent to assuming a probit model for $Y_2$: $p(Y_{i2}(z) = 1 | S_i = s) = \Phi(\mu_2^{s,z})$. Note that under PER for re-employment, $\mu_2^{n,1} = \mu_2^{n,0}$. For principal strata, we assume a Bernoulli distribution

$$p(S_i = c) = \pi_c \quad \text{and} \quad p(S_i = n) = \pi_n = 1 - \pi_c. \tag{5}$$

The parameters are $\boldsymbol{\theta} = \{\pi_c, \boldsymbol{\mu}^{s,z}, \boldsymbol{\Sigma}^{s,z}\}$.

*Prior distributions for parameters.* To simplify the notation, a priori distributions are specified omitting the superscript $s, z$. For the mean parameters, $\boldsymbol{\mu}$, we assume the independent diffused normal priors, $\boldsymbol{\mu} \sim \mathrm{N}(0, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})$, where the prior variance matrices are diagonal $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}} = v_a \mathbf{I}_p$. For the covariance matrices $\boldsymbol{\Sigma}$, due to the constraint of $\sigma_{22} = 1$, there is no conjugate prior. Letting the covariance parameters $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{12})$, we need to ensure that the distribution of $\boldsymbol{\sigma}$ is truncated to the region $\mathcal{A} \subset \mathbb{R}^2$ where $\boldsymbol{\Sigma}$ is a positive definite matrix, that is, $\mathcal{A} = \{\boldsymbol{\sigma} : \sigma_{11} > \sigma_{12}^2\}$. As in Chib and Hamilton (2000), we assume a truncated bivariate normal prior for $\boldsymbol{\sigma}$, $\boldsymbol{\sigma} \sim \mathrm{N}(\boldsymbol{\sigma}_0, \boldsymbol{\Sigma}_0)\mathbf{1}_{\mathcal{A}}(\boldsymbol{\sigma})$, where $\boldsymbol{\sigma}_0$ and $\boldsymbol{\Sigma}_0$ are hyperparameters, and $\mathbf{1}_{\mathcal{A}}$ is the indicator function taking the value one if $\boldsymbol{\sigma}$ is in $\mathcal{A}$ and the value zero otherwise.

*Prior to posterior computation.* The posterior distributions of the parameters were obtained from Markov chain Monte Carlo (MCMC) methods. The MCMC algorithm that we adopted uses a Gibbs sampler with data augmentation to impute at each step the missing compliance indicators $D_i^{\mathrm{mis}}$ and to exploit the complete

compliance data posterior distribution to update the parameter distribution. Details of the MCMC are given in the supplementary material [Mattei, Li and Mealli (2013)].

*Results*. We estimated PCEs using four models: (1) a bivariate model that does not assume ER for either depression or re-employment; (2) a bivariate model that assumes PER for re-employment; (3) an univariate model for depression that does not assume ER; and (4) an univariate model for depression that assumes ER for never-takers. We do not present results from the bivariate model that assumes ER for both depression and re-employment because under ER (and monotonicity) the improvement from secondary outcomes is only marginal, as we can uniquely disentangle the mixtures of distributions associated with principal strata without invoking any additional distributional or behavioral assumption.

The posterior distributions were simulated running three chains from different starting values [see the supplementary material Mattei, Li and Mealli (2013), for further details on chains' initial values]. Each chain was run for 10,000 iterations after a burn-in stage of 5000 iterations. The potential scale-reduction statistic [Gelman and Rubin (1992)] suggested good mixing of the chains for each estimand, providing no evidence against convergence. Inference is based on the remaining 30,000 iterations, combining the three chains.

Table 2 presents the posterior median and 95% credible interval for the estimands of interest—the PCEs on depression for compliers, $\tau_c$, and never-takers, $\tau_n$—obtained from the four models. For $\tau_n$, both the univariate model without ER and the bivariate models with and without PER for re-employment lead to a small and negligible estimated effect, suggesting that never-takers' depression status was little affected by the invitation to attend the job-search seminar. This is also evident from the posterior densities plotted in the bottom panel of Figure 1: the posterior distributions of $\tau_n$ are evenly spread around zero with a large span. These results

TABLE 2
*Summary statistics*: *Posterior distributions of PCEs on depression for compliers and never-takers*

| | **Median** | **2.5%** | **97.5%** | **Width of the 95% credible interval** |
|---|---|---|---|---|
| PCEs for compliers ($\tau_c$) | | | | |
| 1. Bivariate | −0.338 | −0.594 | −0.105 | 0.489 |
| 2. Bivariate with PER | −0.205 | −0.758 | 0.285 | 1.043 |
| 3. Univariate | −0.206 | −0.582 | 0.125 | 0.707 |
| 4. Univariate with ER | −0.260 | −0.613 | 0.049 | 0.661 |
| PCEs for never-takers ($\tau_n$) | | | | |
| 1. Bivariate | 0.043 | −0.193 | 0.263 | 0.456 |
| 2. Bivariate with PER | −0.056 | −0.684 | 0.488 | 1.171 |
| 3. Univariate | −0.084 | −0.527 | 0.287 | 0.813 |

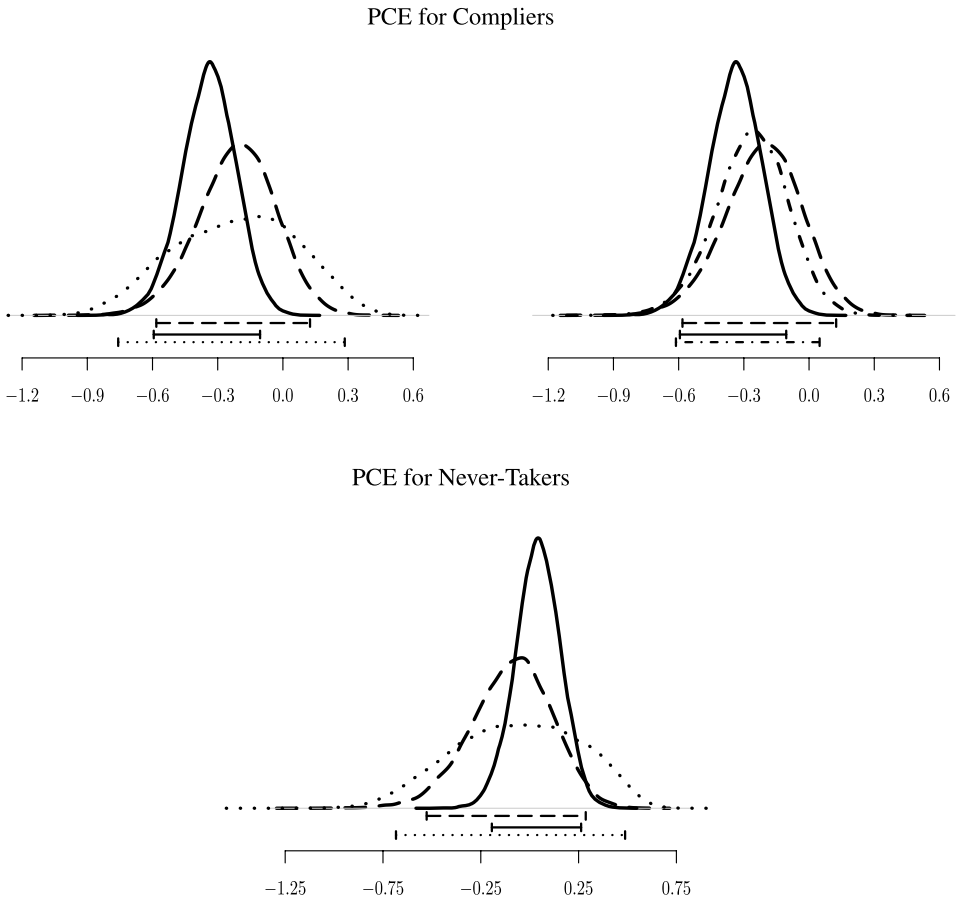


FIG. 1. *Posterior densities* (*derived using a kernel smoothing*) *and* 95% *posterior intervals of PCEs on depression for compliers* ($\tau_c$) *and never-takers* ($\tau_n$) *under the univariate approach with ER* (*dot–dashed lines*), *the univariate approach without ER* (*dashed lines*), *the bivariate approach* (*solid lines*) *and the bivariate approach with PER* (*dotted lines*).

imply that the ER assumption for depression in never-takers may be reasonable. Interestingly, the bivariate model that does not assume ER for any outcome still significantly improves inferences about PCEs, reducing the width of the credible interval for $\tau_n$ by 44% compared to that from a univariate analysis (rows 5 and 7). Conversely, the bivariate model with PER provides a large posterior credible interval for $\tau_n$ (see the discussion below).

For the PCEs for compliers, $\tau_c$, a negative point estimate is obtained from all four models: $-0.338$ in the bivariate case, $-0.205$ in the bivariate case with PER, $-0.206$ in the univariate case, and $-0.260$ in the univariate case with ER. The posterior probability of this effect being negative is greater than 75% irrespective of the approach we consider. Therefore, all the approaches show some evidence

that the invitation to attend the job-search seminars reduces depression among compliers. However, only the bivariate model leads to a 95% credible interval not covering 0, with a 99.8% posterior probability that $\tau_c$ is negative. In fact, the bivariate analysis without PER provides considerably more precise estimates for $\tau_c$ than both the bivariate analysis with PER and the univariate analyses with and without ER: the bivariate model without ER for any outcomes (row 1) reduces the width of the 95% credible interval for $\tau_c$ by 53% compared to the bivariate model with PER (row 2), and by 31% and 26% compared to the univariate model without (row 3) and with (row 4) ER, respectively. This is further illustrated by the posterior densities plotted in the upper panel in Figure 1. The bivariate approach with PER performs worse than the univariate approaches, too: the 95% posterior credible intervals for the PCEs on depression from the bivariate approach with PER are more than 30% wider than those derived from the univariate approaches. Somewhat surprisingly, the posterior distributions of $\tau_c$ and $\tau_n$ from the model with PER have large variances. This highlights an interesting phenomenon about PER that will be further investigated through our simulations: PER helps to reduce posterior uncertainty only if it does (or approximately) hold and is imposed. However, when it is imposed but does not hold, PER may force the parameters to lie in a region of the natural parameter space that is far away from the truth and thus leads to larger posterior variances. This is what may have happened in the JOBS II analysis: even if there is large posterior uncertainty about the effect of assignment on re-employment for never-takers, imposing this effect to be exactly zero leads to ill-fitted models.

It is worth noting that the bivariate approach leads to posterior distributions of $\tau_c$ and $\tau_n$ centered at slightly different medians. In light of the simulation results, which show that jointly modeling two outcomes generally leads to posterior means and medians closer to the true values, these findings suggest the bivariate estimates are more reliable, while the univariate estimates may be far from the true values.

JOBS II is a randomized experiment, and so pre-treatment covariates do not enter the assignment mechanism. Nevertheless, covariates could be still used to improve precision of the causal estimates. Our analysis can also use covariates in addition to auxiliary outcomes. Indeed, we also estimated the models previously described conditional on several relevant covariates. Similar results were obtained, but the benefits of the bivariate approach, that we want to highlight here, are particularly evident when no covariates are used. Therefore, we relegate the details for the models with covariates to the supplementary material [Mattei, Li and Mealli (2013)].

**5. Simulations.** To better understand the results of the JOBS II application and, more importantly, to further shed light on the comparison between univariate and bivariate principal stratification analyses in general settings, we conduct an extensive simulation study. We consider a wide range of simulation scenarios that often occur in practice, accounting for different correlation structures between the

outcomes for compliers and never-takers, various deviations from the PER for the secondary outcome, and different association levels between the auxiliary variable and the compliance status.

To simplify computation, we generate two continuous outcomes from a mixture of two bivariate normal distributions as model (4), and the stratum membership from a Bernoulli distribution as model (5). Although we only consider bivariate Normal distributions in our simulations, we can reasonably expect that our results are not tied to distributional assumptions: Mealli and Pacini (2013) show that secondary outcomes can also tighten large-sample nonparametric bounds for PCEs, and Mercatanti, Li and Mealli (2012) show that the use of an auxiliary variable may improve inference also in misspecified Gaussian mixture models. See also, for example, Gallop et al. (2009), Mealli and Pacini (2008), for further insights on the role of distributional assumptions in PS analysis. We assume that parameters are a priori independent and use conjugate diffuse prior distributions. The true simulation parameters are shown in Table 3. Mimicking the JOBS II data, all simu-

TABLE 3

*True values of parameters of the seven simulation scenarios. The last two columns show the ratio of the between-groups variance and the total variance of the secondary outcome under the control and the active treatment arm, where the groups are defined by the compliance status (correlation ratio)*

| Scenario | $\boldsymbol{\mu}^{n,0}$ | $\boldsymbol{\mu}^{n,1}$ | $\boldsymbol{\Sigma}^{n,0}$ | $\boldsymbol{\Sigma}^{n,1}$ | $\eta^2_{Y_2\|S,Z=0}$ | $\eta^2_{Y_2\|S,Z=1}$ |
|---|---|---|---|---|---|---|
| I | $\begin{bmatrix}2.75\\12\end{bmatrix}$ | $\begin{bmatrix}4.25\\12\end{bmatrix}$ | $\begin{bmatrix}0.16 & 0.16\\0.16 & 4\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.08\\0.08 & 4\end{bmatrix}$ | 0.639 | 0.770 |
| II | | | $\begin{bmatrix}0.16 & 0.64\\0.64 & 4\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.32\\0.32 & 4\end{bmatrix}$ | | |
| | $\begin{bmatrix}2.75\\12\end{bmatrix}$ | $\begin{bmatrix}4.25\\13\end{bmatrix}$ | | | 0.639 | 0.824 |
| III | | | $\begin{bmatrix}0.16 & 0.16\\0.16 & 4\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.08\\0.08 & 4\end{bmatrix}$ | | |
| IV | | | $\begin{bmatrix}0.16 & 0.64\\0.64 & 4\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.48\\0.48 & 9\end{bmatrix}$ | | |
| | $\begin{bmatrix}2.75\\12\end{bmatrix}$ | $\begin{bmatrix}4.25\\24\end{bmatrix}$ | | | 0.639 | 0.950 |
| V | | | $\begin{bmatrix}0.16 & 0.16\\0.16 & 4\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.12\\0.12 & 9\end{bmatrix}$ | | |
| VI | | | $\begin{bmatrix}0.16 & 0.96\\0.96 & 9\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.80\\0.8 & 25\end{bmatrix}$ | | |
| | $\begin{bmatrix}2.75\\24\end{bmatrix}$ | $\begin{bmatrix}4.25\\36\end{bmatrix}$ | | | 0.941 | 0.957 |
| VII | | | $\begin{bmatrix}0.16 & 0.24\\0.24 & 9\end{bmatrix}$ | $\begin{bmatrix}0.04 & 0.20\\0.2 & 25\end{bmatrix}$ | | |

In all the scenarios

$$\boldsymbol{\mu}^{c,0}=\begin{bmatrix}2.5\\8\end{bmatrix},\ \boldsymbol{\mu}^{c,1}=\begin{bmatrix}0.5\\6.5\end{bmatrix},\ \boldsymbol{\Sigma}^{c,0}=\begin{bmatrix}0.09 & 0.24\\0.24 & 1\end{bmatrix},\ \boldsymbol{\Sigma}^{c,1}=\begin{bmatrix}0.01 & 0.08\\0.08 & 1\end{bmatrix}$$

lated data sets have $n = 600$ units, generated using principal strata probabilities of 0.7 for compliers and 0.3 for never-takers. The simulated samples are randomly divided into two groups, half assigned to the treatment and half to the control. Three parallel MCMC chains of 15,000 iterations with different starting values were run for each of the seven simulated data sets, with the first 5000 as burn-in. Mixing of the chains was determined to be adequate and all chains led to similar posterior summary statistics.

Figure 2 shows the posterior densities and 95% posterior credible intervals of the PCEs for compliers and never-takers on the primary outcome, in both the univariate and bivariate cases. The results clearly demonstrate that simultaneous modeling of both outcomes significantly reduces posterior uncertainty for the causal estimates. In fact, the bivariate approach outperforms the univariate one in each of the scenarios considered, providing considerably more precise estimates of the PCEs for compliers and never-takers.

The benefits of the bivariate approach especially arise when compliers and never-takers are characterized by different correlation structures (scenarios III and V) and when the association between the auxiliary outcome and the compliance status is stronger (scenarios VI and VII). In addition, plots (III), (V), (VI) and (VII) in the upper and lower panels of Figure 2 suggest that the posterior distributions of the PCEs are much more informative in the bivariate case. Specifically, plots (III) and (VII) show that the posterior distributions of the PCEs for compliers and never-takers are flat in the univariate approach, but become much tighter in the bivariate case. The improvement is even more dramatic in scenarios (V) and (VI), where the plots show that posterior distributions of the PCEs for compliers and never-takers are bimodal in the univariate case, but both become unimodal in the bivariate case. Also, in the above scenarios jointly modeling the two outcomes leads to posterior means of the PCEs for compliers and never-takers much closer to the true values. The bivariate approach outperforms the univariate one also in scenarios II and IV, where compliers and never-takers are characterized by similar correlation structures. In both scenarios the bivariate approach considerably increases the precision of the estimates.

In scenario I, where PER for the secondary outcome holds, we also derived the posterior distributions of the PCEs for compliers and never-takers by specifying a bivariate model that assumes PER. The bivariate models with and without PER lead to similar results, and both clearly outperform the univariate model, leading to much less variable and more informative posterior distributions of the causal effects of interest. Several other scenarios with additional structural assumptions were also examined: magnitude of the improvement varies, but the pattern is consistent with what is described here.

Additional bivariate analyses were conducted to investigate the role of PER, by fitting the bivariate model with PER also to the six data sets generated under scenarios II through VII, where PER does not hold. Results, shown in the supplementary material [Mattei, Li and Mealli (2013)], suggest that inference for the PCE

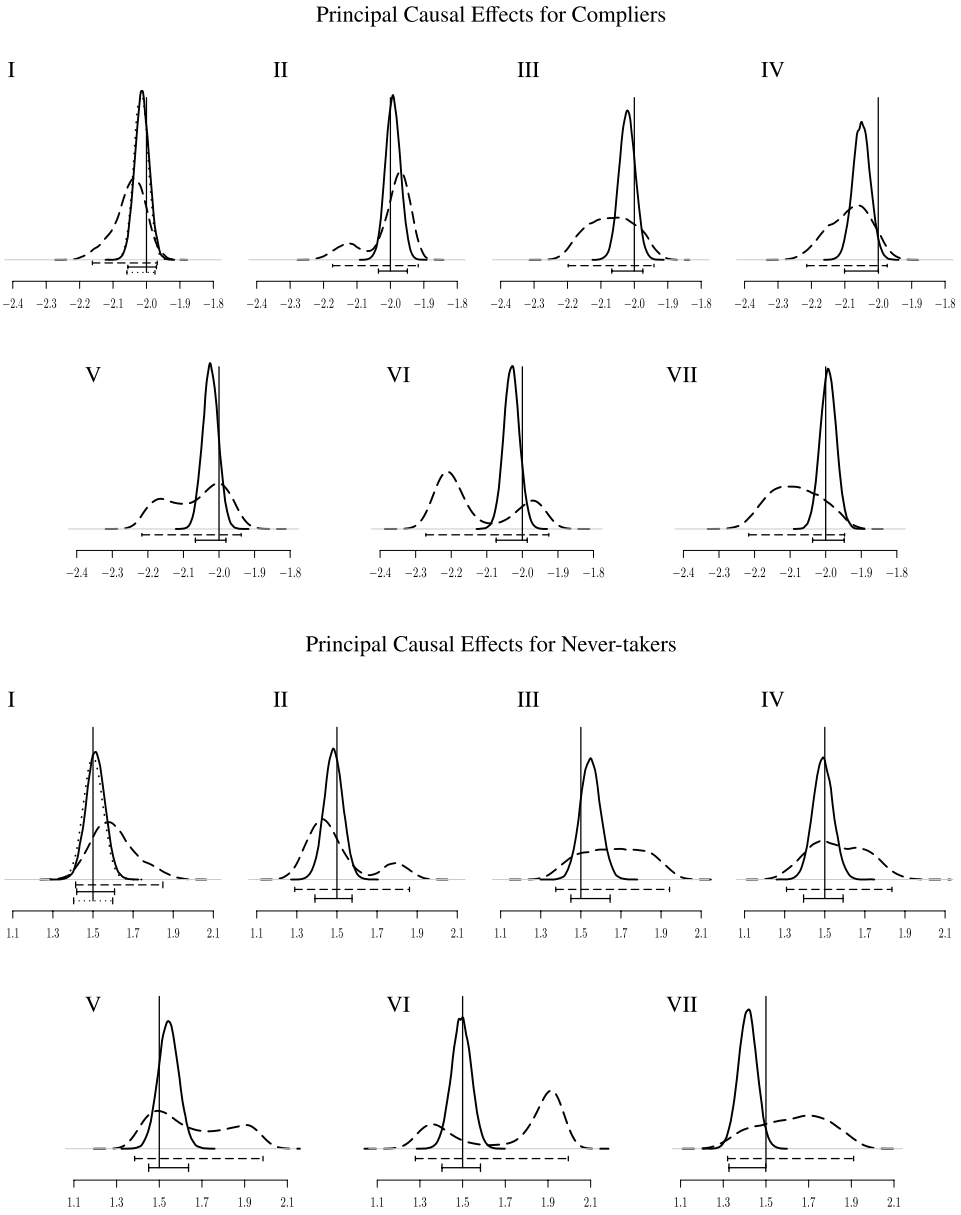


FIG. 2. *Posterior densities* (*derived using a kernel smoothing*) *and* 95% *posterior intervals of PCEs on the primary outcome for compliers* ($\tau_c$) *and never-takers* ($\tau_n$) *under the univariate approach* (*dashed lines*), *the bivariate approach* (*solid lines*) *and the bivariate approach with PER* (*dotted lines*). *The black vertical lines represent the true values*. *The Roman numbers denote the simulation scenarios described in Table* 3.

for compliers is robust with respect to violation of PER: the corrected-specified bivariate model and the misspecified bivariate model with PER perform similarly, leading to posterior distributions for the PCE for compliers characterized by similar posterior variability and similar posterior means. On the other hand, inference on the PCE for never-takers appears to be rather sensitive to the PER assumption, especially when PER is strongly violated (scenarios V, VI and VII) and when compliers and never-taker are characterized by similar correlation structures (scenarios II and IV). In these scenarios, the posterior distributions from the misspecified bivariate models with PER are characterized by larger posterior uncertainty and are centered at posterior means much farther away from the true parameters than the posterior means from the corrected-specified bivariate models. Also, the posterior distributions of the PCE for never-takers derived from the misspecified bivariate models with PER provide 95% posterior credible intervals that do not even cover the true parameter in most of the scenarios.

These results shed light on two key complementary facts about PER. First, as already anticipated, PER may help to reduce posterior uncertainty when it does hold and is imposed, although the jointly modeling of two outcomes still improves inference increasing precision, even if no exclusion restriction on the secondary outcome is imposed. It is worth noting that this is a different result from the nonparametric large-sample case, where the secondary outcome does not help sharpening inference if no exclusion restriction is imposed on it [Mealli and Pacini (2013)]. Second, PER may actually increase the posterior variability of the causal estimates and lead to misleading results, when it is imposed but does not hold. Therefore, less precise inference under PER can be viewed as evidence of violation of PER, which is the case in the JOBS II application. This highlights the importance of carefully evaluating the plausibility of ER assumptions.

In order to evaluate the accuracy and robustness of the proposed approach, we also investigated its repeated sampling properties using Monte Carlo simulations, which were summarized by calculating standard frequentist measures, including average biases, percent biases, mean square errors (MSEs) and coverage of nominal 95% confidence intervals. Results [shown in the supplementary material Mattei, Li and Mealli (2013)] confirm, and generally magnify, the findings discussed here that the simultaneous modeling of two outcomes may improve estimation by reducing posterior uncertainty for causal estimands.

**6. Posterior predictive model checking.** The use of multiple outcomes may help in improving inference, although the additional information provided by secondary outcomes is obtained at the cost of having to specify more complex multivariate models, which may increase the possibility of misspecification. Therefore, model checking procedures to ensure sensible model specification are crucial.

Bayesian goodness-of-fit methods have been proposed in the literature, including Bayes factors and marginal likelihood [e.g., Chib (1995)] and posterior predic-

tive checks [e.g., Gelman, Meng and Stern (1996), Rubin (1984)]. Here, we focus on posterior predictive checks, which are based on comparisons of the observed data to the posterior predictive distribution. A posterior predictive check generally involves the following: (a) choosing a discrepancy measure, $\Delta$; and (b) computing a Bayesian $p$-value.

The posterior predictive discrepancy measures that we use here were first proposed by Barnard et al. (2003) and can be defined as follows. Let $\mathcal{D}_{s,z}^{\text{study}} = \{i : S_i^{\text{study}} = s \text{ and } Z_i = z\}$ be the group of subjects of type $S_i^{\text{study}} = s$ assigned to treatment $Z_i = z$, $s = c, n$, $z = 0, 1$, in the *study* data, where study = obs for the observed data and study = rep for data from a replicated study, that is, outcome data and compliance status drawn from their joint posterior predictive distribution. Note that the assignment variable is fixed at its observed values. Let $N_{s,z}^{\text{study}}$ be the number of units in the *study* data belonging to the $\mathcal{D}_{s,z}^{\text{study}}$ group, and let $\overline{Y}_{m,s,z}^{\text{study}}$ and $s_{m,s,z}^{2,\text{study}}$ denote the mean and the variance of the outcome variable $Y_m^{\text{study}}$, $m = 1, 2$, for this group of units. Then, the discrepancy measures we use are

$$SI_{m,s}^{\text{study}}(\boldsymbol{\theta}) = |\overline{Y}_{m,s,1}^{\text{study}} - \overline{Y}_{m,s,0}^{\text{study}}| \quad \text{and}$$

$$NO_{m,s}^{\text{study}}(\boldsymbol{\theta}) = \sqrt{\frac{s_{m,s,0}^{2,study}}{N_{s,0}^{\text{study}}} + \frac{s_{m,s,1}^{2,study}}{N_{s,1}^{\text{study}}}}$$

and the ratio of $SI_{m,s}^{\text{study}}(\boldsymbol{\theta})$ to $NO_{m,s}^{\text{study}}(\boldsymbol{\theta})$: $SN_{m,s}^{\text{study}}(\boldsymbol{\theta}) = \frac{SI_{m,s}^{\text{study}}(\theta)}{NO_{m,s}^{\text{study}}(\theta)}$, $m = 1, 2$, $s = c, n$. These measures aim at assessing whether the model, which includes the prior distribution as well as the likelihood, can preserve broad features of signal, $SI_{m,s}^{\text{study}}(\boldsymbol{\theta})$, noise, $NO_{m,s}^{\text{study}}(\boldsymbol{\theta})$, and signal to noise, $SN_{m,s}^{\text{study}}(\boldsymbol{\theta})$, in the outcome distributions for compliers and never-takers.

In order to assess the plausibility of the posited models as a whole, we also consider the $\chi^2$ discrepancy, defined as the sum of squares of standardized residuals of the data with respect to their expectations under the posited model [e.g., Gelman, Meng and Stern (1996)]; and for the continuous outcome (depression, $Y_1$), the Kolmogorov–Smirnov discrepancy, defined as the maximum difference between the empirical distribution function and the theoretical distribution implied by the posited model.

A widely-used Bayesian $p$-value is the posterior predictive $p$-value (PPPV)—the probability over the posterior predictive distribution of the compliance status and the parameters $\boldsymbol{\theta}$ that a discrepancy measure in a replicated data drawn with the same $\boldsymbol{\theta}$ as in the observed data, $\Delta^{\text{rep}}(\mathbf{S}^{\text{rep}}, \boldsymbol{\theta})$, would be as or more extreme than the *realized* value of that discrepancy measure in the observed study, $\Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \theta)$: $p(\Delta^{\text{rep}}(\mathbf{S}^{\text{rep}}, \boldsymbol{\theta}) > \Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \boldsymbol{\theta}) | \mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{Z}^{\text{obs}}, \mathbf{X})$ [Gelman, Meng and Stern (1996), Rubin (1984)].

PPPVs are Bayesian posterior probability statements about what might be expected in future replications, conditional on the observed data and the model. Therefore, extreme $p$-values, that is, $p$-values very close either to 0 or 1, can be interpreted as evidence that the model cannot capture some aspects of the data described by the corresponding discrepancy measures, and would indicate an undesirable influence of the model in estimation of the estimands of interest.

Although the PPPVs are Bayesian posterior probabilities, even within the Bayesian framework, it is desirable that they are, at least asymptotically, uniformly distributed over hypothetical observed data sets drawn from the true model. Unfortunately, PPPVs are not generally asymptotically uniform, but they tend to be conservative in the sense that the probability of extreme values might be lower than the nominal probabilities from the uniform distribution. This conservatism property implies that PPPVs may lack of power to detect model violations. Alternative posterior predictive checks have been proposed in the literature, including partial posterior predictive $p$-values and conditional predictive $p$-values [e.g., Bayarri and Berger (2000)], calibrated posterior $p$-values [Hjort, Dahl and Steinbakk (2006)] and sampled posterior $p$-values [Gosselin (2011), Johnson (2004), Johnson (2007)]. Here we focus on sampled posterior $p$-values (SPPVs), which have been shown to have at least asymptotically a uniform probability distribution [Gosselin (2011)].

The SPPV is defined as $p(\Delta^{\text{rep}}(\mathbf{S}^{\text{rep}}, \boldsymbol{\theta}^{(j*)}) > \Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \boldsymbol{\theta}^{(j*)}) | \mathbf{Y}^{\text{obs}}, \mathbf{D}^{\text{obs}}, \mathbf{Z}^{\text{obs}}, \boldsymbol{\theta}^{(j*)})$, where $\boldsymbol{\theta}^{(j*)}$ is a *unique* value of $\boldsymbol{\theta}$, randomly sampled from its posterior distribution. Following Gosselin (2011), we calculated the SPPV associated to the JOBS II study using the following two steps: (i) draw $K$ simulated replicated data sets from the sampling distribution conditional on $\boldsymbol{\theta}^{(j*)}$; (ii) draw at random the $p$-value from a Beta distribution with parameters $a + 1$ and $b + 1$, where

$$a = \sum_{k=1}^{K} \mathbf{1}_{\{\Delta^{\text{rep}_k}(\mathbf{S}^{\text{rep}_k}, \boldsymbol{\theta}^{(j*)}) > \Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \boldsymbol{\theta}^{(j*)})\}} + \varepsilon \sum_{k=1}^{K} \mathbf{1}_{\{\Delta^{\text{rep}_k}(\mathbf{S}^{\text{rep}_k}, \boldsymbol{\theta}^{(j*)}) = \Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \boldsymbol{\theta}^{(j*)})\}},$$

$$b = \sum_{k=1}^{K} \mathbf{1}_{\{\Delta^{\text{rep}_k}(\mathbf{S}^{\text{rep}_k}, \boldsymbol{\theta}^{(j*)}) < D^{\text{obs}}(\mathbf{S}^{\text{obs}}, \theta^{(j*)})\}} + (1 - \varepsilon) \sum_{k=1}^{K} \mathbf{1}_{\{\Delta^{\text{rep}_k}(\mathbf{S}^{\text{rep}_k}, \boldsymbol{\theta}^{(j*)}) = \Delta^{\text{obs}}(\mathbf{S}^{\text{obs}}, \boldsymbol{\theta}^{(j*)})\}}$$

with $\varepsilon \sim U(0, 1)$.

A potential drawback of SPPVs is that they might provide different random results on the same data and the same model, depending on the *single* value $\boldsymbol{\theta}^{(j*)}$ of the parameter vector $\boldsymbol{\theta}$ that is sampled. To avoid this issue, we also implemented the solution proposed by Gosselin (2011), which involves drawing more than a single value of the parameter vector $\boldsymbol{\theta}$ from its posterior distribution. The steps are as follows: (a) a value $u$ from a uniform distribution on $(0, 1)$ is drawn; (b) $J >$

1 values of the parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}$, are drawn from its posterior distribution; (c) for each $j = 1, \ldots, J$, the sample posterior $p$-value associated with $\boldsymbol{\theta}^{(j)}$ is computed; (d) the SPPVs are combined using the empirical $u$-quantile of the latter distribution. We call the Bayesian $p$-value derived from this approach the *modified*-SPPV.

Table 4 shows the results from the three Bayesian $p$-values we considered. The SPPVs are based on $K = 500$ replicated data sets, and the modified-SPPVs were

TABLE 4
*Posterior predictive checks*

| **Approach** | **Signal** | | **Noise** | | **Signal-to-Noise** | | | **Kolmogorov–** |
|---|---|---|---|---|---|---|---|---|
| **Outcome** | ***c*** | ***n*** | ***c*** | ***n*** | ***c*** | ***n*** | $\chi^2$ | **Smirnov** |
| | | | | *Posterior predictive p-values* | | | | |
| Bivariate | | | | | | | | |
| Depression | 0.513 | 0.805 | 0.432 | 0.564 | 0.528 | 0.798 | 0.597 | 0.400 |
| Re-employment | 0.497 | 0.502 | 0.670 | 0.242 | 0.416 | 0.582 | 0.475 | |
| Bivariate with PER | | | | | | | | |
| Depression | 0.573 | 0.574 | 0.522 | 0.573 | 0.562 | 0.552 | 0.563 | 0.389 |
| Re-employment | 0.542 | 0.493 | 0.408 | 0.492 | 0.545 | 0.493 | 0.382 | |
| Univariate | | | | | | | | |
| Depression | 0.601 | 0.678 | 0.836 | 0.865 | 0.536 | 0.623 | 0.979 | 0.441 |
| Univariate with ER | | | | | | | | |
| Depression | 0.555 | | 0.802 | | 0.484 | | 0.939 | 0.373 |
| | | | | *Sample posterior p-values* | | | | |
| Bivariate | | | | | | | | |
| Depression | 0.545 | 0.798 | 0.697 | 0.619 | 0.473 | 0.783 | 0.866 | 0.816 |
| Re-employment | 0.379 | 0.438 | 0.830 | 0.121 | 0.262 | 0.582 | 0.663 | |
| Bivariate with PER | | | | | | | | |
| Depression | 0.693 | 0.807 | 0.512 | 0.520 | 0.663 | 0.800 | 0.592 | 0.341 |
| Re-employment | 0.856 | 0.818 | 0.527 | 0.341 | 0.863 | 0.761 | 0.416 | |
| Univariate | | | | | | | | |
| Depression | 0.170 | 0.731 | 0.747 | 0.320 | 0.154 | 0.757 | 0.699 | 0.410 |
| Univariate with ER | | | | | | | | |
| Depression | 0.190 | | 0.625 | | 0.169 | | 0.899 | 0.392 |
| | | | | *Modified sample posterior p-values* | | | | |
| Bivariate | | | | | | | | |
| Depression | 0.893 | 0.872 | 0.571 | 0.367 | 0.401 | 0.747 | 0.803 | 0.659 |
| Re-employment | 0.228 | 0.115 | 0.705 | 0.618 | 0.122 | 0.690 | 0.433 | |
| Bivariate with PER | | | | | | | | |
| Depression | 0.117 | 0.605 | 0.631 | 0.542 | 0.329 | 0.329 | 0.546 | 0.627 |
| Re-employment | 0.495 | 0.802 | 0.892 | 0.260 | 0.788 | 0.699 | 0.566 | |
| Univariate | | | | | | | | |
| Depression | 0.241 | 0.283 | 0.888 | 0.868 | 0.820 | 0.097 | 0.900 | 0.255 |
| Univariate with ER | | | | | | | | |
| Depression | 0.200 | | 0.811 | | 0.724 | | 0.692 | 0.172 |

calculated by drawing at random $J = 1000$ values of the parameter vector from its (simulated) posterior distribution and simulating $K = 500$ replicated data sets for each $j = 1, \ldots, J$.

As can be seen in Table 4, the estimated Bayesian $p$-values for the bivariate model that does not assume ER for any outcome range between 11.5% and 89.3%, suggesting that the bivariate model fits the data pretty well and successfully replicates the corresponding measure of location, dispersion and their relative magnitude. Unsurprisingly, similar results are obtained for the bivariate model with PER for re-employment. In fact, the analyses do not provide strong evidence against PER for re-employment, so it is reasonable that posterior predictive checks fail to detect the potential benefits of the bivariate model that does not assume PER over the bivariate model that does assume PER. However, the empirical results in Section 4 show that the bivariate model without PER considerably reduces posterior uncertainty for the causal estimands of interest. Therefore, also in light of the simulations, we expect that inferences drawn without assuming PER may be more reliable. On the other hand, the PPPVs and the modified-SPPVs show some evidence that the univariate models might not optimally fit the data according to the $\chi^2$ discrepancy. In addition, the modified-SPPVs suggest that the univariate model without ER might fail to replicate the signal-to-noise measure in the depression distribution for never-takers. These potential failures of the univariate models might be due to the underlying categorical nature of the depression variable. More flexible statistical models could be considered and compared, but the potential failures of the univariate models seem to be successfully fixed when the additional information provided by the secondary outcome is used, so we do not further drill down this issue in this paper, where focus is on investigating the benefits of jointly modeling multiple outcomes in causal inference with post-treatment variables.

**7. Conclusion.** Motivated by the evaluation of a job training program (JOBS II), we have demonstrated, within the framework of principal stratification, the benefits of jointly modeling more than one outcome in model-based causal analysis for studies with intermediate variables. Observed distributions in these studies are typically mixtures of distributions associated with latent subgroups (principal strata). Structural or behaviorial assumptions are often invoked to uniquely disentangle these mixtures. When such assumptions are not plausible, distributional assumptions are often invoked. But these usually lead to models that are weakly identified, weakly in the sense that the likelihood function has substantial regions of flatness. From a Bayesian perspective, even when the likelihood is rather flat, if the prior is proper, so will be the posterior. However, posterior uncertainty will still be rather large in these models, with posterior distributions of causal parameters often presenting more than a single mode, unless the prior is extremely informative.

We have shown how to sharpen inference in these weakly identified models: improvements are achieved without adding prior information or additional assumptions (such as ERs, weak monotonicity or stochastic dominance), but rather by

using the additional information provided by the joint distribution of the outcome of interest with secondary outcomes. Indeed, in the JOBS II application, ERs are not particularly plausible. Nonetheless, by jointly modeling depression, the primary outcome, and re-employment status, a secondary outcome, we have found improved evidence for a positive effect of the job-training program on trainees' depression compared to a univariate analysis on depression alone. Additional simulations further illustrate the benefits under more general scenarios.

JOBS II is a randomized study, but we stress that our framework can also serve as a template for the analysis of observational studies with intermediate variables. In observational studies, randomization (ignorability) of treatment assignment is usually assumed conditional on relevant pretreatment variables [Rosenbaum and Rubin (1983)], thereby conditioning on the covariates is not optional in observational studies but crucial for credible causal statements. However, once ignorability is assumed, the structure for Bayesian inference in observational studies with intermediate variables (e.g., mediation analysis) is the same as that in randomized experiments. The differences lie in the structural assumptions: for example, while in some experiments the design of the study can help in making the ER assumption plausible (blindness or double-blindness), the ER assumption for an *instrument* in observational studies is often questionable. As a consequence, improving inference of weakly identified models is even more relevant in observational studies.

**Acknowledgments.** The authors are grateful to the Editor, the Associate Editor and two reviewers for constructive comments, to Guido Imbens, Booil Jo and Barbara Pacini for helpful discussions, to Amiram Vinokur and University of Michigan, The Interuniversity Consortium for Political and Social Research (ICPSR) for providing the JOBS II data. Part of this paper was written when Alessandra Mattei was a senior fellow of the Uncertainty Quantification program of the U.S. Statistical and Applied Mathematical Sciences Institute (SAMSI).

## SUPPLEMENTARY MATERIAL

**Supplement to "Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program."** (DOI: 10.1214/13-AOAS674SUPP; .pdf).

**Supplement A:** Details of calculation. We describe in detail the Markov Chain Monte Carlo (MCMC) methods used to simulate the posterior distributions of the parameters of the models introduced in Section 5 in the main text.

**Supplement B:** Posterior inference conditional on pretreatment variables. We describe details of calculation and results under the alternative models conditioning on the pretreatment variables.

**Supplement C:** Additional simulation results. We present additional simulations aimed at investigating the role of the partial exclusion restriction assumption and assessing the repeated sampling properties of the proposed approach.

## REFERENCES


ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

BARNARD, J., FRANGAKIS, C. E., HILL, J. L. and RUBIN, D. B. (2003). Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *J. Amer. Statist. Assoc.* **98** 299–323. MR1995712

BAYARRI, M. J. and BERGER, J. O. (2000). $p$ values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142. MR1804239

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

CHIB, S. and HAMILTON, B. H. (2000). Bayesian analysis of cross-section and clustered data treatment models. *J. Econometrics* **97** 25–50.

ELLIOTT, M. R., RAGHUNATHAN, T. E. and LI, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics* **11** 353–372.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

GALLOP, R., SMALL, D. S., LIN, J. Y., ELLIOTT, M. R., JOFFE, M. and TEN HAVE, T. R. (2009). Mediation analysis with principal stratification. *Stat. Med.* **28** 1108–1130. MR2662200

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A. E. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GILBERT, P. B. and HUDGENS, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262

GOSSELIN, F. (2011). A new calibrated Bayesian internal goodness-of-fit method: Sampled posterior $p$-values as simple and general $p$-values that allow double use of the data. *PLoS ONE* **6** 1–10.

GUSTAFSON, P. (2009). What are the limits of posterior distributions arising from nonidentified models and why should we care? *J. Amer. Statist. Assoc.* **104** 1682–1695. MR2750585

HIRANO, K., IMBENS, G. W., RUBIN, D. B. and ZHOU, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1** 69–88.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive $p$-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. MR2463484

JO, B. and MUTHEN, B. (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. In *New developments and techniques in structrual equation modeling* (G. A. Marcoulides and R. E. Schumacker, eds.) 57–87. Erlbaum Associates, Mahwah, NJ.

JOHNSON, V. E. (2004). A Bayesian $\chi^2$ test for goodness-of-fit. *Ann. Statist.* **32** 2361–2384. MR2153988

JOHNSON, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Anal.* **2** 719–733. MR2361972

LI, Y., TAYLOR, J. M. G. and ELLIOTT, M. R. (2010). A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66** 523–531. MR2758832

LI, Y., TAYLOR, J. M. G. and ELLIOTT, M. R. (2011). Causal assessment of surrogacy in a meta-analysis of colorectal cancer trials. *Biostatistics* **12** 478–492.

LITTLE, R. J. and YAU, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods* **3** 147–159.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review* **80** 319–323.

MATTEI, A., LI, F. and MEALLI, F. (2013). Supplement to "Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program." DOI:10.1214/13-AOAS674SUPP.

MATTEI, A. and MEALLI, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63** 437–446. MR2370802

MEALLI, F. and PACINI, B. (2008). Comparing principal stratification and selection models in parametric causal inference with nonignorable missingness. *Comput. Statist. Data Anal.* **53** 507–516. MR2649105

MEALLI, F. and PACINI, B. (2013). Using secondary outcomes and covariates to sharpen inference in randomized experiments with noncompliance. *J. Amer. Statist. Assoc.* **108** 1120–1131.

MERCATANTI, A., LI, F. and MEALLI, F. (2012). Improving inference of Gaussian mixtures using auxiliary variables. Discussion Paper 12–14. Dept. Statistical Science, Duke Univ., Durham, NC.

ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

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

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593. MR0590687

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **106** 1331–1344. MR2896839

SCHWARTZ, S., LI, F. and REITER, J. P. (2012). Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Stat. Med.* **31** 949–962. MR2913871

SJÖLANDER, A., HUMPHREYS, K., VANSTEELANDT, S., BELLOCCO, R. and PALMGREN, J. (2009). Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics* **65** 514–520. MR2751475

SMALL, D. S. and CHENG, J. (2009). Comment on "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data." *Biometrics* **65** 682–686. MR2766612

SOMMER, A. and ZEGER, S. L. (1991). On estimating efficacy from clinical trials. *Stat. Med.* **10** 45–52.

TEN HAVE, T. R., ELLIOTT, M. R., JOFFE, M. and ZANUTTO, E. (2004). Causal linear models for non-compliance under randomized treatment with univariate continuous response. *J. Amer. Statist. Assoc.* **99** 16–25.

VINOKUR, A. D., CAPLAN, R. D. and WILLIAMS, C. C. (1987). Effects of recent and past stress on mental health: Coping with unemployment among Vietnam veterans and non-veterans. *Journal of Applied Social Psychology* **17** 710–730.

VINOKUR, A. D., PRICE, R. H. and SCHUL, Y. (1995). Impact of the JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology* **23** 39–74.

ZHANG, J. L. and RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death." *Journal of Educational and Behavioral Statistics* **28** 353–368.

A. MATTEI
F. MEALLI
DEPARTMENT OF STATISTICS,
INFORMATICS, APPLICATIONS
UNIVERSITY OF FLORENCE
VIALE MORGAGNI, 59
50134 FLORENCE
ITALY
E-MAIL: mattei@disia.unifi.it
mealli@disia.unifi.it

F. LI
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
122 OLD CHEMISTRY BUILDING
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: fli@stat.duke.edu