Bayesian Approaches to Causal Inference in Large-Scale Surveys

David Kaplan

Department of Educational Psychology

Presented to the Economics and Applied Statistics Unit, European Commission - Joint Research Center
This talk will focus on recent developments in Bayesian approaches to causal inference applicable to large-scale surveys.

- Part 1 of my talk will focus on recent work using Bayesian propensity score analysis.

- Part 2 of my talk will focus on new ideas in Bayesian posterior predictive causal inference.

The discussion will be situated in the context of large-scale educational assessments.

Implications for the design of background questionnaires will also be discussed.
Part 1: Bayesian Propensity Score Analysis
Introduction to Part 1


- Overview of two-step Bayesian PSA

- Overview of BMA and Occam’s window

- Approximate and fully Bayesian BMA

- Design and Results of Simulation Study

- Summary Conclusions
● It is well established that in randomized experiments, individuals are assigned to treatment conditions with a known probability.

● In the limit, randomization controls for all pre-treatment differences among participants in the experiment.

● In observational studies, individuals self-select into treatment conditions on the basis of an unknown mechanism.

● Thus, the selection process is often highly nonrandom, introducing selection bias that may result in highly unbalanced covariates and thus severely weakening causal inferences.
The best that can be hoped for is that the investigator has obtained measurable and reliable covariates that relate to the selection mechanism, recognizing that unobservable covariates might still be operating to bias treatment effect estimates.

Establishing balance between the treatment and control groups on observable covariates is thus essential for obtaining unbiased treatment effect estimates.
In a classic paper, Rosenbaum and Rubin (1983) proposed propensity score analysis as a practical tool for reducing selection bias through balancing on measured covariates.

The propensity score is a scalar function of covariates so that subjects who match on their propensity scores can be treated as having similar covariate background.

This scalar function is most often obtained as the predicted probability of treatment assignment given the covariates obtained from a logistic regression.

A variety of propensity score techniques have been developed for both the estimation and the application of the propensity score.
A review of the extant literature reveals very few studies examining Bayesian approaches to propensity score analysis.

An earlier paper by Rubin (1985) argued that because propensity scores are, in fact, randomization probabilities, these should be of great interest to the applied Bayesian analyst.

Rubin (1985) argued that under the condition of strong ignorability and assuming that the estimated propensity score \( \hat{\epsilon}(z) \) is an adequate summary of the observed covariates \( z \), then the applied Bayesian will be well-calibrated (Dawid, 1982), in the sense that posterior predictions should match up with what happens in reality.

Although Rubin (1985) provides a justification for why an applied Bayesian should be interested in propensity scores, his analysis does not address the actual estimation of the propensity score equation or the subsequent outcomes equation from a Bayesian perspective.
Bayesian Propensity Score Analysis

- A paper by McCandless et al. (2009) provided an approach to Bayesian propensity score analysis for observational data.

- Their approach involves treating the propensity score as a latent variable and modeling the joint distribution of the data and the parameters for the propensity score and outcomes equations simultaneously via an MCMC algorithm.

- From there, the marginal posterior probability of the treatment effect that directly incorporates uncertainty in the propensity score can be obtained.

- A recent paper by An (2010) also put forth a joint modeling approach to Bayesian propensity score analysis.

- Gelman, et al. (2003) have argued that the propensity score should provide information only regarding study design and not regarding the treatment effect, as is the case with the Bayesian procedure advocated by McCandless et al. (2009) and An (2010).
Kaplan and Chen (2012, Psychometrika) proposed a two-step Bayesian propensity score analysis that:

1. Separates the PS equation from the outcome equation.

2. Allows priors to be incorporated into the PS and outcome equations.

3. Shows excellent covariate balance (Chen and Kaplan, in press).

4. Shows excellent frequentist properties (calibrated Bayes; Little, 2006; 2011; 2012).
Following Kaplan and Chen (2012, *Psychometrika*) consider a posterior sampling procedure of a chosen Bayesian logit model with 1000 iterations and a thinning interval of 1.

For each observation, there will be $m = 1000$ propensity score estimates $\hat{e}(x)$ calculated using propensity score model parameters $\alpha$ and $\beta$ as follows,

$$
\hat{e}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta} x)}{1 + \exp(\hat{\alpha} + \hat{\beta} x)}.
$$ (1)
Let $J = 1000$ treatment estimates be generated from posterior distribution of $\gamma$, where $\gamma$ is the treatment effect.

Assuming that $y$ is the outcome and $T$ is the treatment indicator, then Kaplan and Chen (2012) provide the following treatment effect estimator,

$$E(\gamma \mid x, y, T) = m^{-1}J^{-1} \sum_{i=1}^{m} \sum_{j=1}^{J} \gamma_j(\eta_i),$$

where $J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i)$ is the posterior sample mean of $\gamma$ in the Bayesian outcome model based on the $i^{th}$ set of propensity scores $\eta_i$, $i = 1, \ldots, m$ and $j = 1, \ldots, J$. The posterior sample mean is then averaged over $m$ sets of propensity scores.
The posterior variance of $\gamma$ is then based on the total variance formula,

$$Var(\gamma \mid x, y, T) = m^{-1} \sum_{i=1}^{m} \sigma^2_{\gamma(\eta_i)} + (m - 1)^{-1} \sum_{i=1}^{m} \{\mu_{\gamma(\eta_i)} - m^{-1} \sum_{i=1}^{m} \mu_{\gamma(\eta_i)}\}^2,$$

(3)

where $\sigma^2_{\gamma(\eta_i)} = (J - 1)^{-1} \sum_{j=1}^{J} \{\gamma_j(\eta_i) - J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i)\}^2$ is the posterior sample variance of $\gamma$ in the Bayesian outcome model under the $i^{th}$ set of propensity scores and

$$\mu_{\gamma(\eta_i)} = J^{-1} \sum_{j=1}^{J} \gamma_j(\eta_i),$$

(4)

is the posterior sample mean of $\gamma$ in the same Bayesian outcome model.
1. Joint posterior distribution of parameters in a Bayesian propensity score model

2. Posterior treatment effects based on posterior propensity scores using a conventional outcome model.
1. Joint posterior distribution of parameters in a Bayesian propensity score model

2. Conditional posterior distributions of the treatment effect using a Bayesian outcome model
Our results reveal that greater precision in the propensity score equation yields better recovery of the frequentist-based causal effect compared to traditional PSA and compared to no adjustment.

Our results also reveal a very small advantage to the Bayesian approach for $N = 100$ versus $N = 250$.

We show that greater precision around the wrong causal effect can lead to seriously distorted results.

Our findings reveal that greater precision around the correct causal parameter yeilds quite good results, with slight improvement seen with greater precision in the propensity score equation.

Our study also suggests that full optimal matching is preferred in terms of bias and precision.
Kaplan and Chen (2012, *Psychometrika*) studied the **treatment effect** and **variance estimates** of the two-step Bayesian propensity score approach comprehensively.

However, the performance of this approach with respect to covariate balance has not been examined.

Chen & Kaplan (in press) investigated covariate balance of the two-step Bayesian propensity score approach under **stratification**, **weighting**, and **optimal full matching** methods, and compared it with the balance performance of the frequentist counterparts.

A **case study** and a real-data based **simulation study** was conducted to evaluate the balance property as well as demonstrate a practical way of assessing covariate balance of the Bayesian propensity score approach.
The balance indices we used are the **standardized mean or proportion difference** (Cohen’s d) and **variance ratio** for each continuous covariate or each level of categorical covariates (binary coding) between the treatment group and control group.

- The standardized mean difference for a **continuous covariate** is obtained by $\frac{(\bar{x}_t - \bar{x}_c)}{\sqrt{(s_t^2 + s_c^2)/2}}$.

- The $\bar{x}_t$ and $\bar{x}_c$ are the sample mean of the covariate in the treatment group and control group, respectively, and the $s_t^2$ and $s_c^2$ are corresponding sample variances.

- The variance ratio for a continuous covariate is defined as $s_t^2/s_c^2$. 

The standardized proportion difference for a specific level of a categorical covariate is calculated by

\[
\frac{\hat{p}_t - \hat{p}_c}{\sqrt{[\hat{p}_t(1 - \hat{p}_t) + \hat{p}_c(1 - \hat{p}_c)]/2}},
\]

where \(\hat{p}_t\) and \(\hat{p}_c\) are proportions of participants in the treatment group and control group, respectively, for that level.

The variance ratio for a certain categorical level is calculated by

\[
\frac{\hat{p}_t(1 - \hat{p}_t)}{\hat{p}_c(1 - \hat{p}_c)}.
\]

The Bayesian propensity score approach naturally provides the 95% posterior probability intervals together with the point estimates of the balance indices.

For comparison purpose, we bootstrap 1000 data sets to obtain the 95% bootstrap confidence intervals of the balance indices for frequentist PSA methods.
Figure: Covariate Balance for Stratification in the Simulation Study.
Figure: Covariate Balance for Optimal Full Matching in the Simulation Study.
The Bayesian propensity score approach assumes that the propensity score model itself is fixed.

Rather, as a model for treatment selection, it is reasonable to assume that many possible models could have been chosen.

Moreover, the goal in Bayesian model selection is to choose a model that has the best predictive capacity. Here we are trying to optimally predict selection into treatment.

Kaplan & Chen (2014) argued that a full accounting of uncertainty in propensity score analysis should address model uncertainty and optimize prediction, and thus the purpose of this paper is to explore Bayesian model averaging in the propensity score context.
Let $M_k$, $k = 1, 2, \ldots, K$, be a set of propensity score competing models that are not necessarily nested. The posterior distribution of $\Delta$ given data $y$ can be written as

\[
p(\Delta | y) = \sum_{k=1}^{K} p(\Delta | M_k) p(M_k | y). \tag{5}
\]

where $p(M_k | y)$ is the posterior probability of model $M_k$ written as

\[
p(M_k | y) = \frac{p(y | M_k)p(M_k)}{\sum_{l=1}^{K} p(y | M_l)p(M_l)}, \quad l \neq k. \tag{6}
\]

$p(M_k) = 1/K$ for each $k$. 

Bayesian Model Averaging
The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k,$$

(7)

where $p(y|\theta_k, M_k)$ is the likelihood under model $M_k$ and $p(\theta_k|M_k)$ is the prior density of $\theta_k$ under model $M_k$. Equation (7) can be approximated using the BIC.

BMA is a natural way to address covariate choice from a Bayesian framework.
Occam’s Window

- To reduce the size of the model space, we use the procedure of Occam’s window (Madigan & Raftery, 1994).

- Models are eliminated if they predict the data less well than the model that provides the best predictions. Exclude models NOT belonging to the set

\[
A' = \left\{ M_k : \frac{\max_l \{p(M_l|y)\}}{p(M_k|y)} \leq C \right\}. \tag{8}
\]

- Models are eliminated from consideration if they receive less support from the data than simpler sub-models (Occams Razor).

\[
B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k, \frac{p(M_l|y)}{p(M_k|y)} > 1 \right\}. \tag{9}
\]
• Bayesian model averaging is simplified by replacing equation (5) with

\[ p(\Delta|y, A) = \sum_{M_k \in A} p(\Delta|M_k, y)p(M_k|y, A), \]  

(10)

where \( A \) is the relative complement of \( A' \) and \( B \).

• The models under consideration for Bayesian model averaging are those that are in \( A' \) but not in \( B \).

• The search algorithm is given in Madigan and Raftery (1994).
Approximate Bayesian Model Averaging

- For approximate Bayesian model averaging propensity score analysis, we first estimate the propensity score using the “bic.glm” routine in the R package “BMA” (Raftery, Hoeting, Volinsky & Yeung, 2013).

- A weighted average of the approximate posterior mean estimates, with the posterior predictive probabilities as weights, are used to create the model-averaged propensity score estimates.

- Based on the propensity score estimates, the treatment effects are estimated in the outcome model via propensity score stratification, weighting, optimal full matching and regression adjustment methods.

- However, the approximate BMA approach is limited in that each unit of analysis has only one estimated propensity score and thus the uncertainty of the propensity score itself is ignored in the treatment effect estimation.

- The posterior distribution of propensity scores can not be directly obtained and has to be normally-approximated using the posterior mean and standard deviation estimates provided by the “BMA” package.
We propose a fully MCMC Bayesian model averaging procedure within the propensity score framework.

1. Select propensity score models (covariates) using the R program “BMA”. Certain cumulative posterior probabilities (e.g., top 50%, 70% and 90%) can be used to limit the selected models to the most crucial ones.

2. Use the Bayesian logistic regression program “MCMClogit” within the “MCMCpack” to obtain the posterior distribution of the propensity score for each selected model. Priors can be specified at this stage.

3. Sample from the posterior distribution of the propensity score in each model with the posterior probabilities as weights to obtain the final posterior distribution of the propensity score.

4. Estimate the treatment effects in the outcome model via stratification, weighting, optimal full matching or regression adjustment based on the posterior estimates of the propensity score.
1. Independently generate random variables $x_1, x_2, \ldots, x_{10}$ as ten covariates under sample size $n = 200$, such as

$$x_1 \sim N(0, 1) \quad x_6 \sim Bernoulli(0.3)$$
$$x_2 \sim Poisson(2) \quad x_7 \sim N(-1, 3)$$
$$x_3 \sim Bernoulli(0.5) \quad x_8 \sim N(2, 2)$$
$$x_4 \sim N(0, 2) \quad x_9 \sim N(1, 0.8)$$
$$x_5 \sim Bernoulli(0.6) \quad x_{10} \sim N(2, 1)$$

2. Let $x = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10})'$ be the vector of covariates. Obtain the true propensity scores by the model:

$$e_i(x) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)},$$

(11)
3. Calculate the treatment assignment vector $T$ by comparing the propensity score $e_i(x)$ to a random variable $U_i$ generated from the $Uniform(0, 1)$ distribution. Assign $T_i = 1$ if $U_i \leq e_i(x)$, $T_i = 0$ otherwise.

4. Generate outcomes $y_1, \ldots, y_n$ using the model:

$$y_i = \lambda x_i + \gamma T_i + \epsilon_i,$$

where $\lambda = (0.2, 0.1, 0.2, -0.1, -0.2, 0.2, -0.2, 0.1, 0.2, 0.1)$, $\epsilon \sim N(0, 1)$ and $\gamma$ is the true treatment effect taking the value of 5.

5. Replicate the above steps 200 times.

We are providing the frequentist properties of a Bayesian procedure (Little, 2006, 2011, 2012).
Differences across all conditions are modest. However,

1. Occam’s window does not influence treatment effect estimation for the approximate Bayesian model averaging approach.

2. BMA-MCMC and BMA-Approx perform similarly when the models with top 50%, 70% or 90% cumulative posterior probabilities are used for propensity score estimation.

3. Priors on the propensity score model parameters have little impact on the treatment effect estimation, but this finding could be due to the particular choice of priors in this study.
1. BMA-Approx works best for stratification, optimal matching, and regression. BMA-MCMC works best for weighting.

2. BMA-MCMC has larger variance estimates as expected. Captures more uncertainty.

3. BMA-approx overall produce better estimates of treatment effects than the two-step BPSA which does not account for model uncertainty.
We also assessed the predictive performance of the two BMA propensity score approaches and compared them with the Bayesian propensity score approach with a single propensity score equation.

Because the propensity score outcome was binary, we adopted the Brier score (Brier, 1950) to evaluate the quality of the predicted posterior propensity scores.

We first randomly split the data set into two halves, one training data set and one testing data set. (We did this 100 times to account for randomness in the splittings).

The propensity score model parameters are estimated using the training data set and then the predicted posterior propensity scores are obtained based on the testing data set and the estimated propensity score model parameters.
The posterior mean of the predictive propensity scores for each student is denoted as $p_i$.

Let $T_i$ denote the treatment selection, half-day or full-day kindergarten program, for student $i$, where $T_i = 0$ or 1.

The Brier score is defined as $\sum_{i=1}^{n} (T_i - p_i)^2$.

The smaller the models Brier score across competing models, the better prediction the model makes.

We find the both BMA approaches provide smaller Brier score values than the two-step approach with only a single propensity score equation.
Conclusions

- PSA covariate choice is a problem of model choice and the Bayesian approach to addressing uncertainty in model choice is via Bayesian model averaging.

- We provide a fully Bayesian propensity score approach that
  1. Accounts for model and parameter uncertainty.
  2. Can incorporate informative and non-informative parameter and model priors.
  3. Yields very accurate treatment effects. Overall, excellent frequentist properties. Applying the log scoring rules remains to be done.
  4. An R package *BayesPSA* is available at http://bise.wceruw.org/publications.html

- Our approach can be adopted by those who wish to conduct propensity score-based causal inference within the Bayesian paradigm.
Part 2: Bayesian Predictive Causal Inference
The dominant framework of causal inference in the education sciences is the potential outcomes framework of Rubin (1974) based on earlier work of Neyman (1923).

Important extensions of the potential outcomes framework for causal inference with applications to education can be found in the work of Imai (2010) and his colleagues on causal mediation analysis, as well as in the ideas of principal stratification (Frangakis and Rubin, 2002).

Other frameworks for causal inference can also be found in the extant literature (e.g. Pearl, 2009 and Robins, 1986).

We discuss a new approach based on the work of Arjas on the use of Bayesian predictive causal inference.

We examine the utility of this approach for developing methods of causal inference in large-scale assessments.
How might the concept of Bayesian prediction based on the use of the predictive distribution be used as a framework for causal inference?

- Let $U$ be a set of unobservable, pretreatment variables, which we will call potential confounders;
- Let $X$ be a set of observed covariates;
- Let $A$ be the causal variable
- Let $Y$ be the outcome.
Following Lindley (2002), the issue in observational studies is that we have no direct control on $A$.

We write the probability of $Y$ under this situation as

$$p(Y|X, see(A))$$

(13)

to emphasize that we only observe $A$ and have no control over the values it can take.

For $A$ to assume a causal meaning, we need to be able to at least imagine a hypothetical manipulation of $A$.

Notice that if such a manipulation were possible, we would modify equation (13) in a manner after Pearl (2009) as

$$p(Y|X, do(A))$$

(14)
For observational studies (e.g. large-scale assessments) how can we go from \( do \) probabilities when the supporting evidence comes from \( see \) data along with the fact that \( A \) is usually endogenous in a \( see \) study?

One approach is to adopt assumptions given by Lindley (2002); namely,

“[t]he joint (prior) distribution of the variables \( U \) and \( X \) is the same in the hypothetical \( do \)-experiment as in the real \( see \)-study, and also that the response \( Y \) given \( X \), \( A \), and \( U \), behaves the same way regardless of whether the event \( A = a \) was done or merely seen. Stated more explicitly, we assume that 
\[
p(Y|U, X, see(A)) = p(Y|U, X, do(A)).
\]”
These assumptions are debatable but not unreasonable. If we consider the full vs. part-time kindergarten program example,

Assumption (i) states that observed background variables such as parent socio-economic status and unobserved variables such as parent attitudes toward home/life balance would be the same regardless whether they were part of a see-study or a do-experiment.

Assumption (ii) states that the outcome of beginning reading skills should respond the same, regardless of whether we directly manipulate assignment to kindergarten program type or whether we simply observe whether children are in full or part-day kindergarten.
Under the assumptions listed above, the problem of causal inference in an observational setting reduces to the ubiquitous problem of confounding and focuses our attention on the conditional distribution \( p(A|U, X) \).

In line with Arjas (2012), predictions in an observational study \( p(Y|X, see(A = a)) \) and \( P(Y|X, see(A = a')) \) could be different due to \( U \) and not \( A \).

Thus, the goal is to find conditions in which the predictions obtained under a \( see \)-study could be considered as having arisen from a \( do \)-study.
We say that $A$ is unconfounded relative to a potential confounder $U$ if $A$ and $U$ are conditionally independent given $X$, that is if

$$p(A|U, X) = p(A|X)$$ (15)

In other words, the posterior distribution of $U$ based on $X$ and $A$ does not depend on $A$.


If $A$ is unconfounded in the sense of Definition 1, then the two posterior distributions $p(U|X, see(A))$ and $p(U|X)$ are the same.
The consequences of Definition 1 and Proposition 1 for causal inference in observational studies are profound.

Under these conditions, we could imagine that the values of kindergarten program type (full or part-day) were chosen in advance and fixed at those values, or perhaps chosen as a function of the covariates \( X \) alone, as through self-selection.

Regardless, the posterior distributions of the unobserved confounders \( U \) would be the same because they are independent of kindergarten program type.
With a slight change of notation, let “obs” indicate an observational see-study and let “ex” indicate a hypothetical do-experiment.

Then, the joint distribution of all observed and unobserved variables can be written as

\[ p_{obs}(U, X, A, Y) = p_{obs}(U)p_{obs}(X|U)p_{obs}(A|U, X)p_{obs}(Y|U, X, A). \] (16)

Under the assumption of no confounding, we can write

\[ p_{obs}(A|U, X) = p_{obs}(A|X). \] (17)
Retaining Lindley’s Lindley (2002) assumptions, we can rewrite equation (16) maintaining the “obs” notation, except in the case where we are assigning values of $A$ as we would in a hypothetical $do$-experiment.

\begin{equation}
 p_{ex}(U, X, A, Y) = p_{obs}(U)p_{obs}(X|U)p_{ex}(A|U, X)p_{obs}(Y|U, X, A). \tag{18}
\end{equation}

This leads to Proposition 2 of Arjas (2012):

If $A$ is unconfounded in the sense of Definition 1, then the posterior distributions of $U$ based on the observed data $(X, A)$ are the same in both schemes.

that is

\begin{equation}
 p_{obs}(U| X, A) = p_{ex}(U| X, A) \tag{19}
\end{equation}
Here, neither of these posterior distributions depend on $A$.

The posterior distributions of $U$ based on the observed data $(X, A, Y)$ are the same in both schemes, that is

$$p_{obs}(U|X, A, Y) = p_{ex}(U|X, A, Y).$$  \hfill (20)

The predictive distributions of $Y$ based on observed data $(X, A)$ are the same in both schemes, that is

$$P_{obs}(Y|X, A) = p_{ex}(Y|X, A).$$  \hfill (21)
The significance equation (21) for our full vs. part-day kindergarten example is that assuming that kindergarten program type is unconfounded in the sense of Definition 1, and retaining Lindley’s assumptions, the predictive distribution of the reading outcome under the observational study can be considered the same as what would be obtained under a hypothetical experiment where children have been assigned to full or part-day kindergarten.

In other words, kindergarten program type has a causal effect on the distribution of the outcome in the see-study as it would if it were a do-experiment.
Although it is clear that there are important assumptions that are being made and that these need to be assessed when considering Bayesian predictive causal inference, these assumptions are no more stringent than those under other methods of causal inference. Thus, the Bayesian predictive approach to causal inference in observational settings should result in a powerful tool for rigorous and evidence-based education research.
The two methods of causal inference discussed in this talk do not exhaust the range of possibilities for causal inference in large-scale assessments, e.g.

1. Causal mediation analysis
2. Instrumental variable estimation

Our approach is pragmatic. There many different types of causal questions and there is no “one size fits all methodology.

However, the approach advocated here merges Pearls “do“-calculus with Rubins potential outcomes framework.
What are the design implications for LSAs if a goal is to draw causal inferences?

The first issue concerns the cross-sectional nature of most large-scale assessments.

In many LSAs, data are collected more or less simultaneously.

We will assume that the temporal dimension is defined by the nature of the question under the assumption that respondents are reliably accessing their memories of actual events.

The question might concern a “treatment” that occurred some time in the past - e.g. attending full or part-day kindergarten.

Although this assumption is debatable, it is no more questionable than other assumptions that are encountered in models of causal inference (e.g. potential outcomes, sequential ignorability, or exclusion restrictions).
With all of this potential confounding, how do we warrant causal claims?

We argue that statistical warrants for causal claims about a real or hypothetical intervention are always made within the context of a specific set of observed and unobserved explanatory variables measured before the onset of the intervention.

Mackie suggests that the problem in distinguishing between conditions and causes is addressed by considering that causes take place in a context, or what Mackie (1974, pg. 35) refers to as a *causal field*. 
What is said to be caused, then, is not just an event, but an event-in-a-certain-field, and some conditions can be set aside as not causing this-event-in-this-field simply because they are part of the chosen field, though if a different field were chosen, in other words if a different causal question were being asked, one of those conditions might well be said to cause this-event-in-that-other-field.
What advice can be given for the design of background questionnaires to support causal inference?

1. The causal variable should reflect an actual event in the life of the respondent (e.g. kindergarten attendance) that is relevant to the policy purposes of the survey.

2. The causal variable should encode a counterfactual statement – a hypothetical manipulation that could have occurred in a real-life experiment.

3. Additional variables should be obtained that represent confounders within the relevant causal field.

4. Focus should be on the predictive distribution of the outcome under treatment assignment.

5. Sensitivity to assumptions should be obtained whenever possible.
GRAZIE MILLE