Living Up to the Name: The Causal Effect of U.N. Peacekeeping Operations on Civilian Well-Being

**Introduction**

* *Motivating Reason/Case (We Want Causation – Not Correlation)*
* *Quick Review of PKOs and Development/Growth*
* *Layout of the Document*
* *Make Sure You Explain What a Treatment and Outcome Are And What Causal Inference Is*

**Limits of Statistical Control**

The standard approach in the conflict management literature (and the social sciences generally) for addressing alternative explanations of an outcome of interest is statistical control. Researchers adjust for alternative explanations by including such control variables in a specified regression formula which includes the main explanatory variable of interest. Following the estimation of a model with this approach, researchers focus on the coefficient for the explanatory variable of interest after covariate adjustment. Researchers may then proclaim something to the effect of, “after controlling for various alternative explanations of the dependent variable, we find that the explanatory is associated with a X-unit change in the dependent variable”. Despite the prominence of this approach, it is lacking in many ways.

A common critique of statistical control is the dependence on the strict assumptions of linear regression. Researchers often insert control variables in a regression equation reflexively without assessing whether the relationship between the control variable(s) and dependent variable is linear when a non-linear specification may be more appropriate. While researchers can create polynomial transformations of variables and specify these transformations in the regression model, such specifications can only occur if researchers consciously choose to do so. A similar concern can be made for interactive effects where the validity of a model’s output is dependent on the researcher to specify such interactive effects. Additionally, compared to alternative mainstream methods, statistical control preforms poorly at balancing covariates, rendering the quality of causal inferences from such designs poor (if researchers attempt to make causal inferences at all). Alternative methods such as matching and weighting are often viewed as favorable for making causal inferences from observational data as they deliberately attempt to create a balanced data set where observations with similar values in the covariates are deliberately matched with other observations that differ only in treatment (main explanatory variable) status. Not only is the logic of such preferred methods more intuitive for the purposes of causal inference, matching and weighting are likewise not burdened by the assumption of linear relationships.

Beyond the model-based critiques of statistical control, a culture of “controlling” and adding a host of covariates into a regression model has also emerged with negative consequences for the estimation of causal and representative effects. In part, this culture has emerged due to the emphasis on omitted variable bias (OVB). While handling OVB is important, many scholars have discussed the inverse issue of including “too many” or “bad” controls (Achen 2005, Schrodt 2014, Rohrer 2018, Cinelli et al. 2022). As Cinelli et al. (2022) note, even when “bad” controls are discussed, the scope of discussion is often limited to researchers including controls that are essentially alternative operationalizations of the outcome itself. These critiques are not made lightly. Casually including a laundry list of controls can create wildly inaccurate results due to collinearity and incorrect specification of the relationship between the covariates and the outcome variable (Achen 2005, Schrodt 2014). With respect to estimating causal effects, researchers may blindly create bias in a variety of contexts, such as including controls that are correlated with the treatment, but not the outcome, including controls that are influenced by the treatment *and* outcome, or including controls that are caused by the treatment and effect the outcome (Rohrer 2018, Cinelli et al. 2022). Much work has been done on the topics of statistical control, its limitations, and the bad practices unintentionally created as a result. Such topics are not the core issue of this paper, although, they are important to review as a justification for embracing alternative methods, encouraging a culture of thinking clearly about the logic of which covariates to include and how to adjust for them, and attempting to make causal inferences from observational data.

**Causal Inference with Observational Data**

The ability to establish causation is not one that is achieved with ease. Two variables may covary with one another, but establishing the exact causal connection between two variables is much more challenging. Within a single unit, a treatment can be applied and the outcome can subsequently alter. However, other factors beyond the treatment that vary across time may also be a part of the causal story. Across spatial dimensions in which a treatment is applied to numerous units, any change in the outcome could simply be a result of unit-specific factors. If the treatment and outcome of interest are both influenced by a common confounding factor, it becomes difficult to disentangle the effect of the treatment on the outcome from the effect the confounder on the outcome. In an ideal world, researchers would have access to a time machine in which they could apply a treatment to a unit and record the outcome. Following this, the researchers could go back in time, ensure that the treatment never occurred, and record the same outcome. The difference in the outcome with treatment and without treatment would represent the causal effect of the treatment. Obviously, this is not an option, and this impossibility is known as the fundamental problem of causal inference.

In the absence of time machines, researchers have relied on randomized controlled trials (RCTs) as the “gold standard” for making causal inferences and resolving the fundamental problem of causal inference. In theory, RCTs resolve two serious problems that arise when attempting to discover a causal effect of a treatment on an outcome. First, because researchers can assign treatment in an RCT, issues of reverse causality in which the outcome effects the treatment are assuaged. Second, because access to treatment is randomized, concerns of confounding are largely ameliorated as any confounding effect cannot be correlated with treatment exposure due to the randomness in treatment assignment. With randomization, pure chance is the only factor correlated with treatment. With confounding effects resolved, the average difference in the outcome between the treated and control units represents the average treatment effect (ATE). While RCTs have their issues and shortcomings (Deaton and Cartwright 2018), their capacity to effectively eliminate the concern of unobserved confounding is impressive. Unfortunatley, for most conflict management research questions, the possibility of RCTs is either impossible or unethical. Researchers do not have the capacity to randomly assign peacekeeping operations, mediated peace agreements, or foreign aid packages. If they did, random assignment would hardly be an ethical method of applying potentially life-saving treatments. As a result, conflict management researchers must attempt to make causal inferences with treatments that are non-random. Consequentially, confounders of an impossible to know quantity - some obvious and some obscure - are introduced that complicate the exact causal relationship between a treatment and an outcome.

From these limitations, both historic and contemporary research in conflict management has tended to neglect causal research. True, while experimental designs in conflict management studies are incredibly difficult to execute, *quasi*-experimental methodology is abundant and has a rich and growing role in social science research. Further, following the emergence and popularization of *do*-calculus and directed acyclic graphs (DAGs), researchers have the capacity to isolate confounders, frame their research designs in experimental language, and isolate causal effect using observational data (Pearl 1995, Rohrer 2018). In the following paragraphs, I discuss two assumptions of causal inference, the ignorability assumption and the stable unit value treatment assumption (SUTVA).[[1]](#footnote-1) The satisfaction of the former assumption is made possible with DAGs and sensitivity analysis. However, especially for conflict and conflict management research, satisfying SUTVA remains a challenge.

Recall that the threat of unobserved confounding seriously jeopardizes the capacity of a researcher to make causal claims with observational data. The ignorability assumption (also known as the exchangeability assumption or the assumption of no unmeasured confounding) is the prime culprit for why researchers interpret regression coefficients as associations instead of causal effects. Since it is impossible to control for *all* confounding effects, a researcher cannot be sure that the unmeasured Nth confounding effect would not have seriously influenced a causal estimate had it been specified in a model. It should be noted that the possibility of unobserved confounding is *always* present with observational data. Indeed, it is this observation, in part, that led many researchers to avoid causal inference with observational data entirely. DAGs are helpful because they do the behind-the-scenes math for a user to isolate confounding effects. However, DAGs are only capable of doing so with variables that users have directly specified. To resolve this, sensitivity analysis is crucial. While there are at least dozens of different iterations of sensitivity analysis, the common goal with such methods is to examine how sensitive causal estimates are to various hypothetical unobserved confounders. Researchers specify various thresholds in magnitude of unobserved confounders and assess whether the likelihood of a confounder of such a magnitude plausibly exists and, if so, to what degree would the inclusion of said confounder mathematically alter the causal estimate. Together, DAGs and sensitivity analysis do not completely satisfy the ignorability assumption. With observational data, this is not possible. However, the combined use of these tools allows researchers to specify their models correctly and assess the extent to which unobserved confounding is a legitimate threat to observed causal estimates.

On the other hand, SUTVA is perhaps a much more threatening assumption for the prospects of causal inference in conflict management research. SUTVA implies that the outcome of one unit is impacted by their treatment status alone and not the treatment status of other units. Such “spillover effects” are naturally problematic for the study of conflict management given the large amount of research dedicated to the role of geographic contiguity and distance in conflict and peace. It is not too difficult to imagine a certain conflict management strategy, such as a UN peacekeeping operation, being implemented in Country A and this specific PKO impacting levels of violence and stability in contiguous, non-treated countries B and C. Indeed, this very scenario has some empirical support (Beardsley 2011). When SUTVA is violated, we cannot assure that non-treated units are serving as control units at all, seriously complicating any causal effect of a particular treatment. Unfortunatley, attempts at satisfying SUTVA are not yet commonplace or standardized. While satisfying SUTVA can become a problem for RCTs, the ability of a researcher to control and monitor treated units renders current efforts to satisfy SUTVA difficult for researchers working with observational data. As a result, I do not claim to satisfyingly address SUTVA in this research project. Nonetheless, considering SUTVA and strategies to satisfy SUTVA for conflict management research questions with observational data should be at the forefront of methodological concerns in this area of research.

**Attempts at Making Causal Inferences in the PKO Literature with Observational Data**

After reviewing the logic of causal inference and assumptions required for causal inference, it is worth evaluating popular quasi-experimental designs employed in the peacekeeping literature to estimate causal effects. As noted earlier, statistical control often falls short compared to alternative methods. Many researchers have acknowledged this in the peacekeeping literature (albeit, for reasons that are oftentimes not explicitly linked to the improvement of causal inferences). In the following section, I review these alternative strategies, briefly explaining the logic of such quasi-experimental design methods and their limitations as it relates to the peacekeeping literature.

Fixed effects (FE) are often employed in the peacekeeping literature (Joshi 2013, Hultman et al. 2014, Kocher 2014, Bove and Ruggeri 2016, 2018, Fjedle et al. 2018, Haas and Ansorg 2018, Beber et al. 2019, Blair 2019, Di Salvatore 2019, Phayal 2019, Phayal and Prins 2019, Bara 2020, Bove et al. 2021), oftentimes not explicitly for causal inference purposes (in many cases, FE is employed to account for omitted variable bias). The implementation of fixed effects can be helpful for making causal inferences due to its capacity to control for all unobserved *time-invariant* factors of a specified unit. By creating a dummy variable for each unit, researchers can remove confounding effects that are unit-specific. Confounding effects such as these are often hard, if not impossible, to identify individually, which lends credit to the implementation of fixed effects. However, two glaring issues with the implementation of fixed effects for making causal inferences in the PKO literature should be noted. First, for the study of post-conflict peace, the implementation of fixed effects for the study of PKOs *as an event* is impossible given that the presence of a PKO in a prior conflict spell is a *time-invariant variable*. In other words, it is *fixed*, meaning that a scholar studying PKOs would be unable to determine the effect of PKOs independent of the other unit-specific fixed factors. This problem can be avoided if one alters their measure of PKOs. If one chooses not to measure PKOs using a dummy, opting to include a fluid measure such as the number of personnel involved in the PKO instead, fixed effects can still be employed given that the PKO measure is no longer a time-invariant variable. Still, while fixed effects accounts for all time-invariant aspects of a unit, it does not eliminate the potential for a *time-variant* confounder to slip through the cracks and bias estimates. One may be tempted to use the two-way fixed effects (TWFE) to account for both unit-specific and time-specific confounders. However, as Imai and Kim (2021) have demonstrated, numerous issues arise when attempting to use TWFE for causal inference.

*Synthetic Control (Bove and Elia 2017)*

In the absence of experimental data, many scholars in the field have adopted an instrumental variables approach to making causal inferences concerning the effect of PKOs (Sambanis 2008, Vivalt 2015, Caruso et al. 2017, Ruggeri et al. 2017, Blair 2019, Bove et al. 2021). The instrumental variables approach seeks to remove aspects of the treatment that are endogenous (associated with confounders) and retain the exogenous aspects (not associated with confounders) of the treatment. The instrumental variables approach does so by identifying a variable (an instrument) that is correlated with the treatment, is not correlated with other confounding factors, and is correlated with the outcome *only through the* treatment. If these conditions are met, it can be assumed that the instrument reflects a portion of the exogenous aspects of the treatment and is untainted by confounding factors. Predicted values are generated by regressing the treatment on the instrument and these subsequent predicted values are used to estimate the causal effect of the treatment on the outcome. While this method is appealing when there are theoretical reasons to believe that confounding variables are present that current data either does not or cannot account for, this approach has not been implemented without controversy. Gilligan and Sergenti (2008) criticized the use of instrumental variables, referring to causal estimates from such an approach as invalid. These authors argued that the literature has a good grasp on the confounders that complicate the relationship between PKOs and peace, rendering the concern of unknown confounders relatively unimportant. Further, the authors were also skeptical that an instrument for this type of research *could* exist on the grounds that "Any factor that affects how long a war or its subsequent peace will last should also be taken into account by the UN Security Council when it is deciding whether or not to allocate a mission" (Gilligan and Sergenti 2007, p. 91). Essentially, the authors argued that there are no exogenous aspects of the treatment (UN PKO) given that the authorization of PKOs are heavily influenced by endogenous factors related to conflict and peace duration. Indeed, the discovery of valid instruments are particularly difficult given the challenge of satisfying the excludability assumption in which the instrument effects the outcome solely through the treatment. For example, weather is commonly used as an instrument in conflict studies employing an instrument variables approach. However, recent work has suggested that this once-reliable instrument heavily violates the excludability assumption (Mellon 2021). Such findings present a fundamental problem with the use of instrumental variables. Instruments are as valid as our ability to argue that the instrument effects the outcome solely through the treatment, rendering the validity of these instruments incredibly sensitive and subjective.

Instead of instrumental variables, Gilligan and Sergenti (2007) suggested the adoption of matching as an approach to improve causal estimates in the peacekeeping literature. The virtues of matching, as the authors claimed, can be attributed to the relative simplicity and transparency of the technique. Units are matched to each other according to their similarity with a specific number of confounding factors. They differ, however, with respect to their treatment status. Given the similarity between matched units, the difference in outcome between matched units *may* be indicative of a causal effect of the treatment. Matches can be made transparent along with the variables on which they are matched. Indeed, given the intuitive nature of this approach, matching is widely employed in the peacekeeping literature (Sambanis 2008, Kathman and Wood 2011, Hultman et al. 2013, Hultman et al. 2014, Ruggeri et al. 2017, Di Salvatore 2018, Fjelde et al. 2018, Haass and Ansorg 2018, Beber et al. 2019, Bara 2020, Mvukiyehe and Samii 2020), albeit, not always as a method to explicitly improve causal interpretation (oftentimes, matching is employed as a “robustness check” to assess the strength of results estimated with adjusted covariates). However, matching strategies have two notable drawbacks. First, because matching *matches* observations and weights these observations in a regression framework accordingly, many observations are naturally omitted from the analysis if a comparable match is not found within the data set. Especially in areas of conflict management research where the number of observations cannot afford to be significantly trimmed, this represents a major problem.

Researchers generally have two options to resolve this issue. First, researchers can relax matching criterion to allow the top *k* best matches to serve as control units rather than relying on the first best match. In addition, researchers can opt to match with replacement, meaning that a control unit can serve as a control unit for more than one treated unit if it is an optimal match for other treated units. While these measures offer a larger *N* to work with, the quality of matching breaks down. Units that are increasingly different from one another are matched (creating further issues with resolving confounding) and repeated control groups are overrepresented in the sample. Another approach is the use of weighting, particularly, inverse probability weighting. Like matching, IPW generates weights to balance the data set along specified confounders. Unlike matching, IPW does this in a manner that does not omit any observations.[[2]](#footnote-2) IPW does so by following a two-step process. First, a numeric value for the propensity of receiving treatment predicted by specified confounders is generated for each observation. Given that treatments are often binary, logistic regression is commonly employed for this purpose. Next, these observations are weighted. Observations are weighted more heavily when their propensity to receive treatment differs largely from their *actual* exposure to treatment. For example, according to the results of a logistic regression model, if a country was very unlikely to receive a UN PKO and *still received one*, this observation would be weighted heavily. Likewise, if the results of the model predicted a certain country had a very high chance of receiving a PKO and *did not*, then this observation is also weighted heavily. Observations that experience the treatment in accordance with the predicted propensity to receive treatment are less heavily weighted. Rather than relying on distance to another observation to be weighted, IPW does not require any data points to be dropped as weights are generated agnostic of other observations. While one may be tempted to match on the generated propensity scores themselves, using these scores themselves as weighting criteria creates a number of issues for making causal inferences (King and Nielsen 2019).

In theory, with a sufficiently large *N*, we could simply run models with matched and weighted data sets and leap right into results interpretation. However, as Imai et al. (2021) note, contemporary mainstream matching methods are designed to operate within a cross-sectional setting in which unit observations are not repeated over time. Given that most conflict management research utilizes panel (time-series cross-sectional) data, this is problematic. In the following section, I discuss the issues that arise when utilizing matching methods designed for cross-sectional research using panel data. I also discuss the novel approach developed by Imai et al. (2021) that allows researchers to follow a matching strategy with panel data. Lastly, I develop a research design to implement this novel method for the study of UN peacekeeping operations.

**Matching and Weighting with Panel Data**

* *Imai et al. (2021) Approach (Which Matching Estimators and Trimming for IPW?)*
* *Set of Cases (What Qualifies as Conflict)*
* *Outcome (Log-Transformation)*
* *Treatment (Treatment-Variation Map)*
* *Causal Estimate (ATT)*
* *Why No Split-Sample (DiD Takes Care of Time-Invariant Factors, Need More N)*
* *Confounders (Explain Why Not Using Precision Neutral Controls + DiD Time-Invariance)*
  + *Development*
  + *Government Capacity*
  + *War Duration*
  + *Population*
* *Sensitivity Analysis*

**Results**

* *Frequency Distribution*
* *Balance Assessment*
* *Results*
* *Placebo Test*
* *Sensitivity Analysis*

**Conclusion**

* *Recap and Summary*
* *Contributions and Suggestions*
* *Limitations (small N is a big concern, SUTVA may still be a problem, data quality concerns, only works for linear regression so peace outcomes like time-to-event, counts of violence, and ordinal scales do not work)*

1. Making causal inferences requires other assumptions as well. However, these assumptions overlap with assumptions researchers familiar with regression and statistical inference should also be familiar with, such as the requirement of a sufficiently large sample size, the reduction of measurement error, and positivity (treatment varies within each possible combination of conditioned variables). As a result, I do not go into great detail on these assumptions in this paper. [↑](#footnote-ref-1)
2. However, it is not uncommon to manually remove select observations if their generated weights are extreme. This occurs when a single non-treated observation, according to the propensity scores, was *very* likely to have been treated and vice versa (typically where the propensity score is .95 for a non-treated observation or .05 for a treated observation). [↑](#footnote-ref-2)