UN humanitarian interventions
–
a critical assessment of impact evaluation

by

Theresa Weippert

Submitted to

Central European University

School of Public Policy

In partial fulfillment of the requirements for the degree of Master of Public Policy

Supervisor: Achim Kemmerling

Budapest, Hungary

June 2017

Abstract

The UN has been intervening in 70 conflicts since its foundation in 1945. It is a highly disputed question whether these interventions had been successful or not. This is reflected in the diverse outcomes of impact evaluations on UN interventions, which lay the foundation for evidence-based policymaking. By conducting a comparative review, which analyses the findings of research on UN interventions, this thesis will fill a gap in research on UN interventions. The thesis investigates the manner of how UN interventions have been evaluated and whether the findings can be regarded as validly assessing the impact of UN interventions. By means of contextual, data coding and methodological criteria, the scopes, research questions, designs and validity claims of 16 studies will be assessed. The comparative review shows that differences in the underlying definition of success as well as different data coding manners results in a diversity of findings. While a majoritarian consensus about the effective impact on reduction of hostility in the short run existis, the long-term impact of UN intervention stays highly disputed.

Table of Contents

Introduction 1

Chapter 1: Contextual considerations on UN humanitarian interventions 5

1.1 Focus on historical context of UN interventions 5

1.2 Focus on short- and long-term goals of UN interventions...7

Chapter 2: Data coding and methodological considerations on impact evaluation 10

2.1 Limitations and key challenges of data coding...10

2.2 Limitations and key challenges in identification methods for causal effects 12

2.2.1 Three challenges in identifying causal effects 12

2.2.2 Matching as method for the identification of causal effects...........................15

Chapter 3: Comparative review of research papers on impact evaluation of UN interventions....17

3.1 Contextual considerations: scope and causal theories of the studies.............................17

3.1.1 Short-term intervention goals: causal theory and empirical findings 18

3.1.2 Long-term intervention goals: causal theory and empirical findings.............22

3.2 Data coding considerations: the influence of data input on impact output 26

3.3 Methodological considerations: theory first in the design-based approach...................28

3.3.1 Ignorable treatment assignment: accounting for selection bias 29

3.3.2 Common support region for counterfactuals: the reality check 30

Conclusion...35

List of references 37

Appendix 41

List of figures and tables

Figure 1: Illustration of omitted variable bias …………………………………………..15

Figure 2: Correlation between threshold coding and number of units………………………..26

Figure 3: Time varying employment within and across UN intervention cases……………….27

Figure 4: Model dependence beyond data support…………………………………………...31

Table 1: Criteria for comparative review of impact evaluations of UN interventions…………17

Table 2: Limited meta-analysis of point estimates for the effect of UN intervention…………25

List of abbreviations

ACD	-	Armed Conflict Dataset
COW	-	Correlates of War Dataset
SUTVA	-	stable unit treatment value assumptions
R2P	-	"Responsibility to Protect" UN doctrine
UN	-	United Nations

Introduction

Nowadays, more personnel than ever before is employed in UN humanitarian interventions around the world. Founded with the purpose to "maintain peace and security amongst the nations", the UN has been involved in over 70 operations of conflict prevention, peacemaking, peacekeeping and peacebuilding. A variety of measures such as diplomatic mediation, humanitarian assistance, economic sanctions or the interference with military power were employed in these UN operations, leading to diverse outcomes which range from resolved wars to complicated, prolonged or recurring conflicts. The very nature of conflicts regarding the actors, purpose and warfare has changed over time and with it the approach of UN interventions to address the contemporary challenges. In the wave of a generally raising demand for evidence-based poliymaking, the UN is under pressure to justify their action by proving whether and which impact their interventions, especially the unconsented use of armed force, have. The task of impact evaluation thereby is to identify and evaluate the appropriateness and success of measures and to give advice on how to best reallocate scare resources. However, the crucial aspect with all impact evaluation is whether the evaluations yield valid results and are consequently capable of giving valuable recommendations.

A broad body of literature is centred around the question what the mechanisms look like that are underlying UN peace operations and what impact these interventions bear. The literature can be divided into three interlinked streams. While the first stream consists of guidelines on evaluations of humanitarian aid interventions in general and qualitative evaluations of UN interventions from practitioners in the field; the second stream analyses and debates about the appropriate use of the statistical methods in observational studies. Both, the qualitative and quantitative streams join together in the third stream of applied research on the impact of UN humanitarian interventions.

The first, mainly qualitative stream consists of reports, guidelines, handbooks and program evaluations written by practitioners from the field. Although literature such as the World Bank´s handbook on "Impact evaluation in practice" (Gertler et al. 2016) focuses on humanitarian aid in general, the outlined framework conditions, methods, operationalization constraints and challenges of data collection and measurement issues; all of which can be transferred from humanitarian aid to UN humanitarian interventions as well. In a series of self-evaluations, (United Nations General Assembly 2015, Stamnes and Osland 2016) the UN and associated bodies have produced a series of papers for a general qualitative performance assessment of UN interventions.

The second, quantitative stream deals purely with the claim for validity and appropriate application of statistical methods. It addresses the foundational question whether, how and to what extent x causes y. Gates and Strand (2004) investigate on the prerequisite of data availability and measurement in civil war studies, which is a general issue in observational studies. Keele (2015) and Imbens and Wooldridge (2009) sketch out insight and controversies about the potential outcome approach, which is based on the question "what would have happened if…".

The third stream consists of academic research papers which reunify the theoretical insight and the methodological appropriateness to gain knowledge about the causal impact mechanisms of UN humanitarian interventions. Research papers must set their focus on a particular aspect of intervention in order to establish meaningful impact statements. The extensive research on determinants and decision-making mechanisms for selecting UN intervention assignments is not considered in this thesis. Once the UN got involved, studies deal with the impact of UN intervention in ongoing conflicts (Hultman et.al. (2013, 2014) while others investigate the duration of peace after a conflict (Beardsley 2012 and 2013, Gilligan and Sergenti 2008, Hultman et.al. 2016).

As this thesis will show, discrepancies appear, not only amongst the results of different research papers, but also comparing the recommendation of the qualitative self-evaluations, especially regarding the employment of military forces. There is a clear need for a comparative review that systematically analyses the validity, coherence and divergences of the research findings in order to facilitate an efficient evidence-based decision-making process for resource allocation to

UN policymakers. Such a comparative systematic review is lacking in the field of UN interventions´ impact evaluation. This thesis will fill this gap and contribute to the existing literature with an analysis that compares the existing research literature regarding their underlying causal explanatory theories and findings. It further investigates on the methodological process of achieving valid causal effect estimates. Thereby, this thesis seeks to answer how UN interventions have been evaluated and whether the research established a valid causal interpretation of the impact.

Methodologically, I will draw on the three strands of literature. The first strand delivers contextual considerations for impact evaluation on UN interventions. The second strand establishes data coding as well as methodological criteria for statistically valid causality claims. The comparative review in the third strand of research literature follows the common protocol of reviews as it is laid down in the Cochrane Handbook for Systematic Reviews on Interventions (2017):

Firstly, the selection of research papers is done by restricting the publishing date to the most recent evidence from 2000 onwards and selecting on 16 studies that focus on ongoing and post-conflict impact evaluations.

Secondly, the contextual considerations for the comparison of causal explanatory theories as well as data coding and methodological criteria for the assessment of validity claims, which were elaborated in Chapter 1 and 2 respectively, will be guiding the analysis.

Thirdly, on basis of these criteria, the comparative review analyses differences and similarities in the research question, explanatory theory, data and research design of the studies at hand. A quantitative meta-analysis of the scope and confidence of results is not possible because the research questions and framework conditions differ too widely, however, the results will be synthesized in a more general way.

The thesis is structured in three chapters. The first chapter provides contextual criteria that are based on the changing nature of conflicts since the foundation of the UN and the subsequent adaption of the UN interventions approach. The second chapter investigates the challenges of data coding and further discusses methodological considerations of the design-based approach and the potential-outcome model. In the third chapter, existing research is systematically analysed and compared according to the established criteria, before concluding the thesis.

Chapter 1: Contextual considerations on UN humanitarian interventions

The purpose and justification for the UN´s intervening actions the task to maintain peace and security amongst the nations. This justification can only be upheld by proving that the humanitarian interventions are indeed successful. The following Chapter first describes the historical context of the changing nature of interventions due to changing characteristics of crisis. Second, it elaborates different definitions of success for UN interventions depending on whether the target is approached from a short- or long-term perspective.

1.1 Focus on historical context of UN interventions

The end of the Cold War marked a clear cut-off in the way how conflicts look like as well as how the UN intervenes. The global security environment shifted throughout the last decades and brought forth three generations of UN interventions which responded to the respective prevailing challenges (Schulenburg 2014). The first generation of so-called traditional peace operations found its main task in maintaining peace agreements and keeping the belligerent parties separated in a time when nations confronted each other in interstate conflicts during the Cold War. These first interventions were deeply attached to the three intervention principles of impartiality, consent of the belligerent parties and non-use of arms except self-defence.

The second generation started with the end of the Cold War when circumstances and the atmosphere changed. On the one hand, international intervention and cooperation became a more common tool of dealing with conflicts around the world, so that since 1988 a total of 56 peace operations were launched compared to only 15 in the four prior decades (UN 2017b). On the other hand, the characteristics of conflicts changed significantly over the decades. Compared to the prevalence of interstate wars, since 1989 around 70% of the conflicts are defined as interstate conflicts, while the 33 ongoing conflicts from 2013 on are unexceptionally categorized as such

(Schulenburg 2014, 3f). Instead of dealing with conventional armies and governments, the UN found itself in complex circumstances with confusing conjunctions between armed forces, civilians and political actors. Within five years after the end of the Cold war, the UN launched 20 new missions and increased its personnel by seven times (UN 2017b). However, the traditional way of peacekeeping did not apply to the changed circumstances and the new demands of dealing with the social, political and economic roots lead thereto that the second generation was perceived as a generation of failures. Although Howard (2008, 5) draws attention to ten success stories like Namibia, El Salvador, Croatia or Timor-Leste amongst the 35 interventions between 1988-2008, they are left in the shadow of the paralyzing happenings in Rwanda, Somalia, Angola and Srebrenica in Bosnia where UN peacekeepers helplessly watched the war crimes and even genocides in front of their eyes. Another drawback was the relaunching of peace missions in Haiti, Liberia and Timor-Leste because the first mission left too early after a too poorly equipped operation (DPO/DFS 2009, 3). Deeply marked by these experiences, the definition and implementation strategy of UN shifted towards tackling the root causes of conflict. The so-called multidimensional way embraces a comprehensive array of tasks beyond the traditional protection and disarmament measures. Restoring the rule of law, promoting human rights and running important governance functions reaches deep into the political sovereignty of the host country. Peace operations in this reformed sense include conflict prevention through diplomat measures, peacemaking, peace enforcement with coercive measure and peacebuilding to ultimately re-enhance the state to gain legitimacy and perform the government´s core functions to its citizens (UN 2017b).

The third generation, starting with new missions in mid-2000, is characterized by a consolidated understanding of multidimensional peace missions, often settled in the midst of ongoing conflicts without a prospect of resolution, dedicated to restoring the collapsed states. The fact that the personnel as well as the timeframe for UN interventions in this generation tripped since 2000 (UN General Assembly 2015, 20) reflects the high complexity of the intrastate conflicts. With a total of more than 113.000 staff, more personnel than ever before are employed in the 16 current UN missions. Amongst those, blue helmet soldiers constitute the mayor part of with a total number of 82.5000 troops (UN 2017a). More than 17.000 soldiers are employed in the largest current operation in the Democratic Republic of Congo, followed by five more such large operations in Dafur, South Sudan, Mali, Lebanon and Central African Republic, all but one established in the last seven years.

The use of military forces compared to measures of diplomacy, police, observers and humanitarian aid assistance is highly debated, even more so since the introduction of the "Responsibility to Protect" (R2P) doctrine in 2000. The R2P principle sought to give a guideline on "when, if ever, it is appropriate for states to take coercive – and in particular military – action, against another state for the purpose of protecting people at risk in that other state" (ICISS 2001, VII). The experience of genocide, ethnic cleansings within state boundaries and war crimes of governments against their own citizens compelled the UN to reshape the principle that host countries must consent to their interventions and the sovereignty clause of Chapter I, Article 2 in the UN Charter which rules out any interference of the UN in domestic concerns. The R2P principle defines that a government has the obligation to protect its citizens and might be intervened in the name of a humanitarian imperative if it fails to do so. The R2P approach has been heavily criticised from states that fear a whitewashed legitimization of the UN to intervene (Seybold 2007, 2) and interfere in the domestic issues with their Westernized approach of governance. The third generation is therefore characterized one the one hand by a greater awareness of the limited capabilities, but on the other hand by simultaneously risen expectations on what the UN should deliver to the nations and civilians.

1.2 Focus on short- and long-term goals of UN interventions

Judging over success and failure is done by comparing the outcome to the initial target. As it will be shown, firstly, the interference between short- and long-term perspective of UN interventions makes this a difficult endeavour. Secondly, the mechanisms of different intervention mandates and their interactions with each other in the different short- and long-term phases are complex and disputed.

The foundation for evaluating the impact of an intervention is how success and failure are defined. Whereas the traditional peacekeeping missions were satisfied with keeping belligerent parties apart and ensuring the enforcement of peace agreements, the objectives of the new generations of multidimensional missions is tackling the root causes of conflict, reconciling hostile parties are securing a self-sustaining stable environment under the rule of law. Although these aims lay at heart of every sustainable peace process, the vague, ambitious a long-term oriented formulation of objectives raises the consideration of proper matching between expectation and capacity to deliver. Such success definitions can easily underestimate the effect of simply preventing greater harm in the midst of a battlefield where there is no peace in sight. However, the habit of implementing template missions (UN General Assembly 2015, 9) that become increasingly technical, bureaucratic and large and the manner of building institutions and infrastructure while overlooking the informal and social networks underneath (UN General Assembly 2015, 48) does not yield long-term satisfaction and justification of UN interventions. Acknowledging this mismatch, a series of self-evaluation and reform analysis of UN interventions was launched (DPO/DFO 2009, UN General Assembly 2015). "Politics first" is the consensus of these reports which conclude that larger military missions can help in certain aspects but since solutions to conflicts are political by nature (UN General Assembly 2015, 42), the design and implementation of interventions must be driven by political considerations. This means walking a thin line on the principle of impartiality. Critiques however say that impartiality has never been an applicable principle by the very nature of UN interventions which dominate the local arena with their masses of troops and dollars and thereby import a liberal concept of governance with ideas of representative democracy and market economy (Schulenburg 2014, 8) which, if applied to the local situation, might actually have the contrary effect of benefiting an elite under the whitewashed label of democratization and development. The unintended consequences caused by politicization of aid, distortion of local economies through creating dependence on aid (Seybold 2007, 277) can potentially outrun the positive short-term impact of a R2P mission and counteract to the justification purpose of maintaining peace and security. In order to prevent massive failures that might even cause more harm and waste UN resources which would have been better employed in a different way, the new "politics first" strategic perspective on peace operation re-emphasizes the option to even refraining from direct and military involvement (Fortna and Howard 2008, 289).

Furthermore, if success is defined in terms of genuine local ownership, another criticism comes in, namely the fact that the evolution of a stable society and ultimately democracy historically took place in time spans that are far more extended than a UN intervention which very rarely lasted longer than 15 years. If the UN is to be judged according to a target that is by definition not achievable in the given timeframe, then its justification might be short-living. Acknowledging the need of UN intervention in conflicts for the sake of the humanitarian imperative but at the same time considering the mismatch of ambitions to capability and of means to challenges, the common baseline of the self-evaluation reports is an urgent call for change of the design and implementation of UN interventions.

The outlined changing nature of circumstances and the consequently changing nature of the UN´s approach has implications for evaluating the impact of interventions. Since different subsamples display diverse characteristic and since different definitions of key concepts might result in diverging results for the same research question, any meaningful evaluation must take four main contextual criteria into account. Firstly, the difference between peacekeeping before and after the Cold War; secondly, associated with the Cold war cut-off, the difference between inter- and intrastate conflicts; thirdly the different aim and scope of the various types of mandates and fourthly, the different impact mechanism in ongoing conflicts compared to past-conflict employment.

Chapter 2: Data coding and methodological considerations on impact evaluation

Before conducting the actual statistical analysis, a prerequisite is to determine the limitations of measuring the various variables. Therefore, this chapter will first highlight limitations and key challenges of coding criteria that are essential for assessing the results of studies on UN interventions. In a second step, it will establish methodological criteria for a comparative review by laying out limitations and key challenges in identifying a causal interpretation of statistical inference through the potential-outcome model.

2.1 Limitations and key challenges of data coding

The following section will uncover the challenges that the four data coding criteria pose on UN impact evaluations. The criteria are firstly the deaths threshold for defining conflicts, secondly the coding of start and end of a conflict, thirdly the interval precision and lastly the coding of recurring or simultaneous conflicts. In the research area of peace and conflicts, two main datasets provide valid information, namely the Correlates of War (COW) and the Armed Conflict Dataset (ACD) by the Uppsala Conflict Data Program in cooperation with the International Peace Research Institute Oslo. By showing how the choice of a particular dataset and its underlying coding influences the results, Gates and Strand (2004, 21ff) confirm the necessity of paying attention to the coding criteria.

Firstly, the threshold of battlefield deaths defines wars and conflicts with an objective criterion. The "Correlates of War" defines a war by 1000 deaths in total independent of the duration of the war. This threshold was criticized for discriminating against small countries which simply do not have the basis for such numbers irrespective of the level of violence. It, for example, excludes internationally recognized conflicts like the one between Burma and Myanmar (Gates and Strand 2004, 6). The ACD in contrast makes a distinction between minor conflicts with more than

25 deaths per year and wars with more than 1000 deaths per year. However, the COW accounts for the fact that counting battlefield deaths alone underestimates the level of violence of any conflict and includes civilian casualties (Gates and Strand 2004, 4). Doing so prevents the absurd result of coding the Rwandan massacre in 1994, which had more than half a million civilian deaths, as a solely intermediate conflict and in the subsequent year as no conflict at all (Gleditsch et.al 2002, 626).

Secondly, coding the start and ending of conflicts is tightly connected with the threshold, since crossing the threshold is the major sign of onset and termination of a conflict. However, casualties might only accumulate gradually to the officially recognized threshold, therefore the ACD also presents initial events whenever applicable (Gates and Strand 2004, 9). The termination date is sometimes clearly marked with the signing of a treaty, an established ceasefire or the victory of one warring party. However, a second look is always needed to assure that such an event really stopped hostilities. This uncertainty about the definite termination of a conflict also occurs if a conflict simply fades under the official threshold. Since this phase of peace easily relapses to a conflict, the 231 conflicts in the ACD are coded as overall 403 violent episodes (Kreutz 2010, 245).

Thirdly, the precision of measurement intervals deals with the problem of counting yearly data from the 1st January onwards. This does not recognize conflicts that amount to the necessary threshold in total, but that are cut in their counting by the arbitrary line of 1st January. For example a conflict with 800 deaths in both fall and spring will not count as conflict in the COW although it exceeds the 1000 deaths in total threshold. The other way around, a conflict that only dwelled short but very violent around new year´s eve will be coded as lasting for two full years (Gleditsch et.al 2002, 625). No matter when exactly a conflict accumulated casualties beyond the threshold, in a yearly interval, the beginning will be coded at the 1st January, even if it only occurred on the 31st December in the same year. Therefore, the more precise the data, the more valid results impact evaluation can possibly yield. Gates and Strand (2004, 13) measured that 40% of conflicts last shorter than a year, with 16.5% even shorter than a week. Useful daily data on conflicts might still be rare, but monthly data becomes more and more the standard to account for time-varying factors.

Fourthly, coding of recurrence and simultaneous events creates different amount of units. Under the COW, for example, the case of Burma and Myanmar was coded as three conflict episodes compared to only two episodes in a dataset of Doyle and Sambanis (2000). Also, depending on whether the researcher takes country or conflict as the basis for the dyad, only the latter one can account for various simultaneous conflicts in the same country. This is the reason why the ACD comes up with 231 conflicts in only 151 countries (Kreutz 2010, 244).

Next to the fact that data generation is done more rigid in conflicts with UN presence and that this can result in a systematic underreporting of deaths numbers for non-intervened conflicts (Gates and Strand 2004, 18), the combination of these four coding criteria can result in a varying number of units for the same time period.

2.2 Limitations and key challenges in identification methods for causal effects

2.2.1 Three challenges in identifying causal effects

In observational studies such as on UN interventions, the potential outcome model provides a basis for interpreting statistical inference as causal impact. "Counterfactual", which is another word for potential outcome, is a term that appears in all major impact evaluation handbooks and guidelines of the UN (Gertler et.al. 2016). In order to interpret the statistical inference that is drawn in a regression model as causal inference, or in other words, in order to interpret the calculated correlation as causality, three main assumptions have to be fulfilled. These assumptions are captured in the potential outcome model, which asks the question "what would have happened if…". In the impact evaluation of various UN agency´s programs, it is possible to conduct randomized control trials, where, under the precondition of compliance, the control group serves as valid proxy for the unobserved outcome of a treated unit that had not been treated. The naïve difference of the average outcome of treated group and the averages outcome of the untreated group constitutes the causal impact there. However, such a randomized assigned control group does not exist in the case of UN interventions.

Instead of comparing two groups, the potential outcome model compares two different statuses of the same unit (Imbens and Wooldridge 2008, 2), namely the real status with an intervention by the UN versus the hypothetical outcome if it had not experienced the intervention. The model is derived by expanding the naïve difference equation

$$Y = E(Y^1|D = 1) - E(Y^1|D = 0)$$

$$Y = E(Y^1|D = 1) - E(Y^1|D = 0) + E(Y^0|D = 1) - E(Y^0|D = 1)$$

(with $D = 1$ as assignment to an intervention, $D = 0$ as assignment to no intervention, Y^0 as outcome before and Y^1as outcome after an intervention). Rephrasing yields the following equation:

$$Y = E(Y^1 - Y^0|D = 1) - [E(Y^0|D = 0) - E(Y^0|D = 1)]\,.$$

This displays the average treatment effect which consists of the average treatment effect of the treated (the intervened countries after intervention minus the same countries if the intervention had not occurred) minus the selection bias. The selection bias thereby shows how much better the intervened country group would have been anyway even without interventions. The fundamental problem of causal inference is the fact that the direct counterfactual of what would have been the outcome of the intervened countries if they had not been intervened, $E(Y^0|D = 1)$, is unobservable.

The so-called design-based approach puts this fundamental problem of causal inference in the first place and gives priority to the theoretical identification of counterfactuals. Making the exact nature of the manipulation, in our case the UN intervention, as explicit and clear as possible reveals the right scope of interpretation of the available data (Morgan and Winship, 2007, 280). This creates an a priori credibility that cannot be made up by simple sensitivity analysis because it ultimately cannot show if the scope and direction of the results are theoretically sound. The strength of the design-based approach thereby lies in the independence of a particular parametric model when settling the assumptions (Imbens and Wooldridge 2008, 5).

After having defined the causal effect, any strategy to identify this causal effect must deal with challenges to three underlying assumptions: firstly, the ignorable treatment assignment, secondly, the so-called "stable unit treatment value assumptions" (SUTVA) which include the absence of omitted variables and thirdly, the common support region for counterfactuals. The first assumption deals with the selection bias which in our case is the difference between not experiencing an intervention for the counties where the UN did not intervene and for the country where the UN did intervene. The core concept in assuming absence of this bias is independence of the two potential outcomes from the treatment to ensure that all variation, such as intervention vs. no intervention or troops vs. police vs. observers, is random (Keele 2015, 316). Since one outcome will always be unobserved, the assumption cannot be verified by data and must be well grounded in theoretical considerations before being shaped through particular statistical methods.

The second assumption, SUTVA, displays the two features consistency and non-interference of treatment. Consistency rules out omitted variables by stating that the treatment is the same for all treated units; an assumption which is hardly justifiable in the case of UN intervention, because the intensity and effectiveness varies hugely across the intervened countries. A partial solution is shrinking down the scope of research and asking for the causal effect of particularly strong military intervention. Also, non-interference between the treated units and across time (Keele 2015, 317) is hard to justify since the socio-economic effects of civil wars hardly stay within a nation´s borders and an UN intervention interferes into this mutual influence between neighbouring countries through various channels. But, although SUTVA is hard to comply with in impact evaluations of UN interventions, Morgan and Winship (2007, 38) argue that this does not equal to a limitation of the counterfactual approach itself but rather reveals the limitations of honest, credible and valid causality claims. In line with the research-design approach, setting the priority in grounding a priori credibility on a solid theoretical foundation, the SUTVA challenge should trigger deeper considerations about the right interpretation and the feasible scope of the causal inference.

The third assumption for the definition of valid counterfactuals is a region of common support. Since the potential outcome that would occur to intervened countries if they had not received the intervention must be an outcome that indeed could have occurred in reality, the distribution of potential values must overlap with the range of actual observed values (Imbens and Wooldridge 2008, 21). However, King and Zeng (2007, 200) point out that overlapping distributions on a two-dimensional scale can reveal a different non-overlapping picture as soon as a third scale of variables controlling for omitted variables is introduced. This is displayed in Graph 1, where Y is the dependent variable, Z the control variable and the solid line refers to the treatment group.

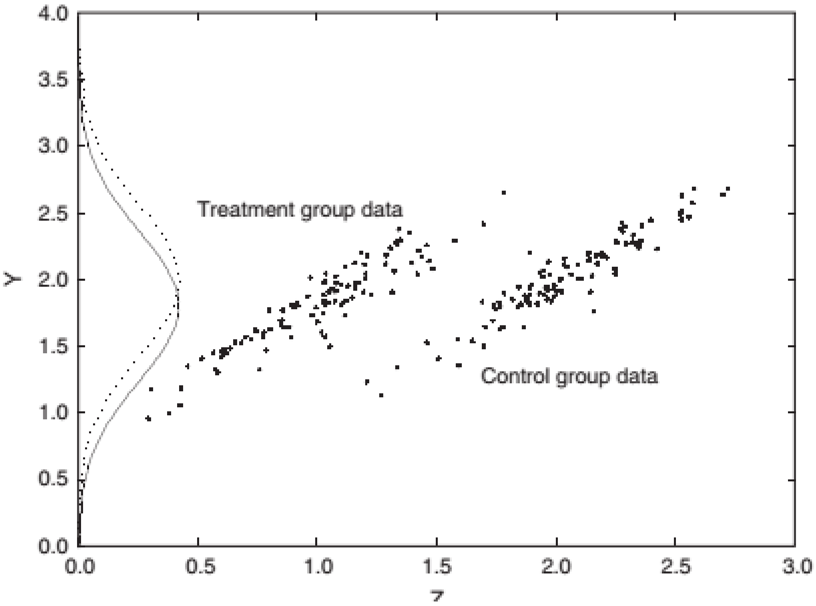

Figure 1: Illustration of omitted variable bias Source: King and Zeng 2007, 200

2.2.2 Matching as method for the identification of causal effects

Matching is one particular strategy that helps to realize the theoretical identification of counterfactuals and must therefore meet the three outlined assumptions. Matching inserts elements of randomization into observational studies and creates individual comparison pairs that are as similar as possible regarding all relevant covariates, so that the control unit serves as a manifestation of the "what would have happened if there was no UN intervention" status. In the search for a realistic counterpart to the intervened countries, matching relies completely on the assertion of a common support region. If no suitable counterpart is found for some matches, it is legitimate to simply drop this observation and thereby create a new subset of the treatment, however, the new scope and the consequently changed interpretation of the average treatment effect of the treated should be explicitly stated. Although this method creates a compelling set up of control units, Morgan and Winship (2007, 50 and 122) point out that neither perfect matching balance nor combining matching and propensity scores warrant valid causality themselves, but that solely the a priori theoretical credibility establishes a solid foundation for valid causal inference.

To sum up, before even conducting the statistical analysis, data input and the theoretical framing of statistical methods predetermine the outcomes. This chapter introduced the four data coding criteria of firstly, death threshold, secondly, coding of start and end dates, thirdly, interval precision and fourthly, coding of recurring or simultaneous conflicts. A all of them must be considered in a comparative review of impact evaluations since the difference in coding can yield systematically different results. The methodological criteria of firstly ignorable assignment, secondly SUTVA and thirdly common support region rephrase the assumptions that must be fulfilled in order to identify a causal effect through counterfactuals. Together with the contextual criteria, they will guide the comparative review of research papers in the next chapter.

Chapter 3: Comparative review of research papers on impact evaluation of UN interventions

One research study alone generally delivers a compelling argumentation for its findings, however, seen in comparison with other studies places the results in a broader framework. Such a comparative review of research on the impact of UN interventions is lacking so far. This chapter will analyse the similarities and differences from 16 studies in order to evaluate the success or failure and appropriateness of UN interventions. The focus and scope of the studies, however, differs too widely as to enable a comprehensive meta-analysis of the statistically calculated point estimates. It only allows a narrow quantitative comparison amongst four studies. The three sections of the analysis will be guided by the contextual, data coding and methodological criteria that were elaborated previously and that are summarized in Table 1. A systematic overview over the comparative review of the 16 studies can be found in the Appendix.

contextual criteria	measurement criteria	methodological criteria
1. before and after the Cold War 2. inter- and intrastate conflicts 3. intervention in ongoing conflicts vs. past-conflict employment (short-term and long-term target definition) 4. types of mandates	1. deaths threshold 2. interval precision 3. coding of duration: initiation and termination of conflict 4. coding of recurrence and simultaneous conflicts	1. selection bias: ignorable treatment assignment 2. SUTVA conditions (not considered here) 3. common support for counterfactuals

Table 1: Criteria for comparative review of impact evaluations of UN interventions, Source: own illustration

3.1 Contextual considerations: scope and causal theories of the studies

In line with the contextual considerations of Chapter 1, this section will apply the elaborated criteria. Regarding on the historical context, the cut-off of the Cold War end marked a change in the characteristics of conflicts as well as the UN´s intervention approach from traditional to multidimensional peace operations. One particular characteristics shift is the changed prevalence of intrastates civil wars since the Cold War compared to previously prevailing interstate conflicts. Another, hidden shift took place through advanced methods for data generation of the common databases (Gilligan and Sergenti 2008, 91). In summary, only 30% of the 16 studies account for these changes in a way that they separate the time periods and evaluate the post Cold War time separately. Around 40% chose the whole time period since 1945, whereas 25% take a mixed version and consider the data from the 60s onwards, since it was by then that data for economic analysis was recorded properly.

Regarding on the short- and long term perspective of UN interventions, impact evaluations must be properly grounded in a theoretical framing of causal mechanisms. The value of a comparative review lies in exposing single explanatory theories to one another and thereby finding potential complementation. Linking explanatory theory and empirical findings together, this section will compare the diverging and overlapping theories which underlie the 16 studies and analyse to which extend the empirical results support or contrast the diverse causal mechanism theories. The first part focuses on defining intervention goals in short-term as the reduction of hostility and the ending of ongoing conflicts. The second part analyses the long-term goal of sustainable peace, democratization and economic prosperity in light of two contrasting theories. Since there is no clear-cut distinction between the short- and long-term perspective of UN interventions, the two parts will refer to each other.

3.1.1 Short-term intervention goals: causal theory and empirical findings

The immediate target of UN interventions in ongoing conflicts is firstly to reduce violence and deaths, which is especially relevant in light of the unconsented military interventions that are justified with the R2P principle of protecting civilians from devastating harm. Secondly, the goal is to achieve the agreement on a ceasefire or peace agreement to end the conflict. The three studies that explicitly focus on ongoing conflict intervention consent in two causal mechanism of UN interventions, namely the separation of belligerents and secondly overcoming the prisoner´s commitment dilemma. However, as will be shown, the efficiency of different types of mandates is contested.

In all three studies, the interaction of warrying actors is framed within a bargaining theory. Reasons for conflicts are various, ranging from power pursuit and separationist ambitions to the altruistic motivation of escaping repressions. The general rationale thereby is the actor´s calculation that the violence pays off compared to other forms of shaping the circumstances. The opposing parties gain more knowledge about each other´s strategy and bargaining power the longer the conflict endures and the harder the violence is (Hultman et.al. 2016, 240). Seeing the duration of war under such perspective, the clearer bargaining positioning might lead to longer lasting peace after the conflict. It can also be the case that the information sharing of a war simply reveals that there is no overlapping reservation point. This would consequently lead to instable peace outcomes, especially if the conflict simply fades out under the threshold of battlefield deaths, but can raise up again at any time. The prisoner´s dilemma within this story is that even if a peace agreement might yield better benefits for all warrying parties, the fact of being the first to let down the arms and trust the opponent party´s commitment is an insuperable obstacle.

The two main channels allow the UN as third actor to raise the costs of violence in the rational choice of belligerents and overcome the security and commitment dilemma: On the one hand, troops create a physical separation and disarm, demobilize and reintegrate the warrying parties (Hultman et.al. 2014, 5). On the other hand, monitoring, detection of infringements and mediation in case of accidental incidences facilitates information sharing and creates more security regarding hidden intentions (Fortna 2004b, 486). As the next paragraphs will show, the functioning of the two mechanisms depend on the capacity as well as the type of mandate.

Firstly, the capacity of UN employment is recognized as decisive feature across all studies, at least in theory. For the combatants to give up their weapon-based power, both channels only work through a credible sign of the UN of being a relevant and powerful intermediator. This capability is reflected in the capacity and strength of mandate, which also display the level of commitment to the conflict case, since a large number of personnel, especially troops, is costly to engage and cannot be withdrawn easily again (Hultman et.al. 2016, 236). Since multidimensional peacekeeping is interfering also in the socio-economic and political institutions, UN interventions can be seen to raise the quality of conflict settlement and therefore expanding the bargaining frontier to mutually improved reservations points (Dorussen and Gizelis 2013). However, the big challenge of multidimensional intervention is to align this attempt to "create value" (Diehl et.al. 1996, 691) with the combatant's self-interest to "claim value" (Diehl et.al. 1996, 691). Since characteristics, such as high fractionalization, which might incentivize a peace agreement, can prove to be detrimental for upholding the agreement in the long run, a key role of the UN is to prevent such a shift of incentives in the transition phase (Beardsley 2013, 374). Although the explanatory strength of the different mechanism channels is contested, a common conclusion across the studies is that an UN intervention without the credible sign of commitment through proper mandates has no if not a negative effect (Hultman et.al. 2014, 6). The crucial feature is the absolute and not the population related number of soldiers in creating a deterrence effect, because also rebel power is unrelated to the population size (Collier et.al. 2008, 472).

Secondly, ten studies disentangle their evaluation efforts and consider the different kinds of mandates separately instead of using a binary coding. The two mechanisms of creating a buffer zone and facilitating negotiations broadly reflect the roles of armed and unarmed personnel, which are highly contested across the studies. Analysing the different kinds of mandates in more depths, Beardsley (2013) finds that diplomacy and sanctions effectively achieve termination of violence and conflict through facilitation of negotiations and pressure respectively. Troops prove to be ineffective in that sense. Also, Fortna et.al. (2004a, 283) attest that the risk of conflict relapse is decreased by 86% through the employment of observers. This stands in contrast to Hultman et.al. (2013 and 2014), who attest a significant effect on violence reduction only to troops but neither observers nor police. Observers are even found to increase the number of civilian deaths significantly (Hultman et.al. 2013, 886). However, Sambanis (2008, 16) counters the claim for a massive increase in military force by stating that the number of troops itself is not predictive for success, but rather the scope of the military mandate, whereas the transformative mandate has more effect irrespective of the size of troops.

Thirdly, several studies explicitly address potentially unintended impact. On the one hand, this is under-resourced personnel, on the other hand wrongly matching mandates to circumstances. The finding that observers are correlated with an increased number of deaths can be a manifestation of the unintended phenomenon of raising incentives for exploitative rebel´s bargaining, especially in the time period when observers already arrived on site while the deployment of troops is still on the way (Hultman et. al. 2014, 11). This unintended negative result of being used for self-interested bargaining is reflected in these studies which accuse the UN intervention of freezing a status quo, that means a sub-optimal conflict resolution, depriving the incentives for thorough negotiations and thereby increasing the probability of relapse. The simulation by Hegre et.al. (2011) as well as the evaluation by Sambanis (2008) estimate that the UN is good in settling conflicts, but cannot contribute significantly in sustaining the peace settlement. Contrary to this, Doyle and Sambanis (2006) found the UN to be failing in settling conflicts but effective in peace implementation. However, the inability of achieving a peace agreement should not be judged as failure right away if the goal of reducing deaths was achieved (Gilligan and Sergenti 2008, 113). Hegre et.al. (2011, 8) however point out the importance of war termination by calculating that in the long-term differently equipped military interventions do not show substantially different effects after a decade. The decisive fact is solely whether a peace agreement was established in the beginning or not (Sambanis 2008, 30). Diehl et.al. (1996, 684) finds that the number of prior relapses multiplies the probability of further conflict recurrences. This means that if the UN fails to establish a sustainable peace settlement, it inversely even worsens the prospects of peace.

3.1.2 Long-term intervention goals: causal theory and empirical findings

The long-term target of interventions is ultimately to resolve the cause of intervention by tackling the root causes of conflicts and fostering self-sustainable peace in democratized and economically prospering states. The definition of success ranges from the narrow view of fewer battlefield deaths than the threshold (Fortna 2004a) towards participatory peace with actual political involvement of stakeholders (Sambanis 2008) or strict peace with a minimum of democratization (Doyle and Sambanis 2000). However, due to unpredictable time-varying levels of political openness and un-codable sovereignty measures, Sambanis (2008, 22) admits that his ambitious study has to step back to the narrow definition of mere absence of violence while assessing long-term impact.

Two theoretical approaches prevail in the long-term perspective, the "politics first" and the "economy and security first" approaches. In the following, the rationale of both approaches will be elaborated together with respective empirical support and counter-evidence. "Politics first" is the clearly announced slogan of the UN (UN General Assembly 2015, 26), which means tackling the root causes of conflict onset and relapse instead of focusing solely on hostility reduction in ongoing conflicts through military involvement. The argumentation behind this approach is that a failure of governance is causing the conflict, no matter if through rent-seeking behaviour or separationist attempts (Dorussen and Gizelis 2013, 692). The broad set-up of multidimensional operations tries to account for this with interventions that do not only separate and disarm the combatants, but that go much further in reforms regarding security and justice issues, democratization and economic development (Diehl et.al. 1996, 691). According to the framework of Doyle and Sambanis (2000, 782), where the probability of successful peace is a function of international capacities substituting failing local capacities and compensating the suffered hostility, the UN either improves existing governance capacities or replaces failing and lacking state institutions.

Dorussen and Gizelis (2013) align to the "politics first" approach and define multidimensional peacekeeping solely as the involvement in policy making and goods provision. A particularly remarkably result of their research is the fact that the fostering of human rights is consistently opposed by government and rebels. This links back to the role of the UN in enhancing the bargaining scope for the combatants through increasing the quality of socio-economic and political infrastructure. Human rights enforcement in regions of bad governance pose constraints to any powerful actor and is therefore opposed. This could be seen as a confirmation that the UN should get involved in the political arena and actively reshape grievances of bad governance. It could, however, also be interpreted to the opposite that the UN should refrain from putting priority in the political arena where mechanisms are complex, opposition is strong and development is slow.

The "economy and security" approach is therefore opposing the UN´s proclaimed approach of "politics first". Contrary to wanting only few military employment, Hultman et.al. (2013) base their recommendation on the finding that "several thousand troops and several hundred police dramatically reduces civilian killings" (Hultman et.al. 2013, 875). They argue further that the risk of conflict relapse, since it is only through force that rebels are deprived of the critical ability to obtain control over territory and local population´s loyalty (Hultman et.al. 2013, 875). Collier et.al (2008) support this opposition to the "politics first" approach. They point out that rather than political motivations, such as claims that result from political exclusion, socio-economic motivations and the mere feasibility of organizing a rebellion trigger conflicts (Collier et.al. 2008, 464). Several factors favour the organization of armed rebel groups, prolong war and increase the probability of conflict recurrence. For example, low income and economic growth levels as well as high socio-economic inequality can be taken as proxy for low recruitment costs. Furthermore, a country´s dependence on natural resources, or more specific the price for exports of primary commodities, can be interpreted as a proxy for the rebel´s income and bargaining power (Collier et.al. 2004, 253). Collier et.al. (2008, 464) argue that these factors are only effectively addressed with economic and military measures and not by political means.

Collier et.al. (2008) provide empirical support for the "economy and security first" approach and find that income per capita, high level of inequality, low level of economic growth and natural resource dependence have significant effect on long-term stability. Compared to the baseline level of 40% risk of conflict recurrence, a growth rate of 10% would decrease relapse risks by 27% and a twofold increase in income level by 31% all else factors being equal (Collier et.al. 2008, 469). These results are supported with coherent findings by Sambanis (2008), who approves the significance of income level and resource dependence, as well as by Doyle and Sambanis (2000, 797) who confirm that a decline in primary commodity prices is significantly associated with a shortening of war duration. Hegre and Sambanis (2006, 531) also find significant effect of income per capita as well as primary commodity prices. Although their finding is related to the onset of wars, it can be argued that the mechanisms for the recurrence of war work similarly.

Due to their strong appeal of massively increasing the number of troops, one could suppose Hultman et.al. (2016) to be great advocates for the "economy and security first" approach. Interestingly, their study finds no significant impact of the income level, but also not the polity score on the relapse of conflict. However, it might simply depend on the fact that instead of considering these variables as pre-conflict measures, they argue that the destruction of wealth and political institutions in wars eliminated variations across countries and therefore rendered any statistical significance impossible (Hultman et.al. 2015, 245).

Collier et.al. (2008) do not only provide support for economic and military priority but also strong evidence against the UN´s politics first approach. Contrary to Hegre and Sambanis (2006, 527) who attest a country´s political transition towards autocracy, expressed in the Polity score, a significant effect in the onset of conflicts, Collier et.al. (2008, 470) show that compared to the baseline risk of relapse, autocracies tend to decrease the risk of conflict recurrence by 25%. This can be seen in accordance with the widely acknowledged finding that one-sided victory result in longer lasting peace. On the other side, countries that are not highly autocratic, but tend to democratization, substantially increase the risk of recurring conflicts by 62%. In line with this delusionary finding, post-conflict democratic elections show a misleading signal of reduced risk in the year of election which disappears right after (Collier et.al. 2008, 471). These findings consequently suggest, that the UN´s "politics first" approach and the promotion of democracy might have a perverted detrimental effect in post-conflict states, whereas a focus on sustainable implementation of peace agreement through military support and economic development is the right way to go.

A major critique to success prospects in the long-run is that the UN firstly lacks the necessary insight into local power relations and societal structures, secondly leaves its peacekeeping missions underequipped and thirdly does not even possess the necessary resources for such multidimensional long-term employment (Diehl et.al. 1996). Common to all studies is the acknowledgement that the factors driving short-term success are insufficient or even detrimental in tackling the root causes of the hostilities and the long-term implementation of peace settlements.

Irrespective of the underlying theoretical approach, there are four roughly comparable point estimates amongst the 16 studies. All of them are reported as hazard rate of conflict recurrence. Comparing to the counterfactual of no intervention, the impact of UN interventions ranges from 46% less likelihood of conflict recurrence (Sambanis 2008) to a drop of 86% (Gilligan and Sergenti 2008). This confirms an overall consensus on the positive impact of UN interventions in the post-conflict setting, however, the meanings of intervention and the size of the effect differ widely.

Characteristics	Author	Point estimate	Confidence interval	Significance level
after Cold War, UN intervention	Fortna (2004a)	0.51	(0.19)	0.1
	Gilligan and Sergenti (2008)	0.14	(-2.18)	Not specified
Total time span, UN intervention	Sambanis (2008)	0.54	(0.16)	0.05
Total time span, general intervention	Fortna (2004b)	0.13	(0.11)	0.05

Table 2: limited meta-analysis of point estimates for the effect of UN intervention, Source: own illustration

To sum up, this section showed how different mechanisms explain the role of UN in ongoing conflicts and post-conflict settings and subsequently lead to different definitions of intervention goals. There is common support that the UN is effective in decreasing hostilities in the short-run, but not whether it achieves to end conflicts. The success for differently defined peace and stability in the long run is disputed just as the mechanisms and effectiveness of different mandates such as troops, diplomacy and observers or sanctions.

3.2 Data coding considerations: the influence of data input on impact output

As different ways of coding eventuate in diverging amounts of units of analysis with distinct characteristics, it is likely that coding affects the research findings. This section will use the previously elaborated four data coding criteria to assess whether the different ways of coding result in a systematic divergence of outcomes. In general, conflicts with UN intervention have a better monitoring of events and casualties (Gates and Strand 2004, 18). That means that UN interventions are associated with higher levels of violence simply because the level of violence is observed properly compared to other conflicts. In summary, the number of the units of analysis differs widely between 36 units in Hultman et.al. (2013) and 145 units in Hultman et.al. (2014). Both studies consider intrastate conflicts after the Cold War with a 25 deaths per year threshold. In Fortna (2004b) 48 units are being analysed and 262 units in Diehl et.al. (1996), of which both use 1000 deaths in total and interstate wars, however, a longer time span and a focus on general intervention is used in the case of Fortna (2004b).

Firstly, regarding the deaths threshold of the conflict definition, two thirds of the 16 studies provided information, whereby approximately half of the studies work with the 25 deaths per year threshold. A basic scatterplot analysis in Figure 2 shows that there is no systematic correlation between the lower threshold and a

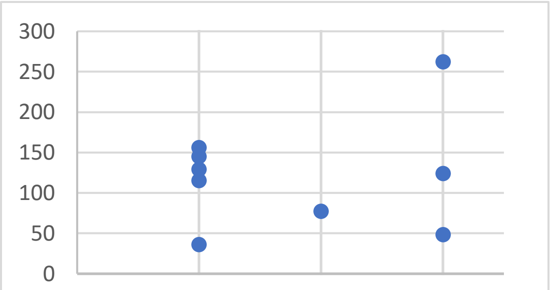

Figure 2: Correlation between threshold coding and number of units, with 1=25 per year; 2= 1000 per year, 3= 1000 in total, Source: own illustration

higher number of units of analysis. Furthermore, although the violence in conflicts has far more manifestations than solely battlefield deaths, only Hultman et. al. (2013) consider civilian deaths. None of the studies addresses further forms of violence through any kind of proxy variables or at least includes a comment on the likely bias towards inefficiency of UN interventions as theoretical consideration in their research.

Secondly, regarding the interval precision of measurements, Gates and Strand (2004), who exceptionally used daily data showed that 15.5% of the conflicts last actually less than a week (Gates and Strand 2004, 13). Such conflicts are coded as month-long conflict in two thirds of the studies which use monthly data and therefore bias the effect of UN intervention. UN interventions are treated as time constant in most of the studies, however, the capacity of different UN mandates can vary immensely within a short time-period. On the one side, the employment and withdrawal of an intervention does take place over a certain time span and on the other side, the reaction to a conflict situation can cause huge differences within a short time as the example of UNOSOM in Somalia illustrates, where the number of UN personnel increased from 700 in January to 29.000 in November 1993 and completely disappeared by February 1995 (Hultman et.al. 2013, 877). Figure 3 illustrates this point of time varying employment within a country as well as the difference of employment across countries with the example of Angola and Liberia.

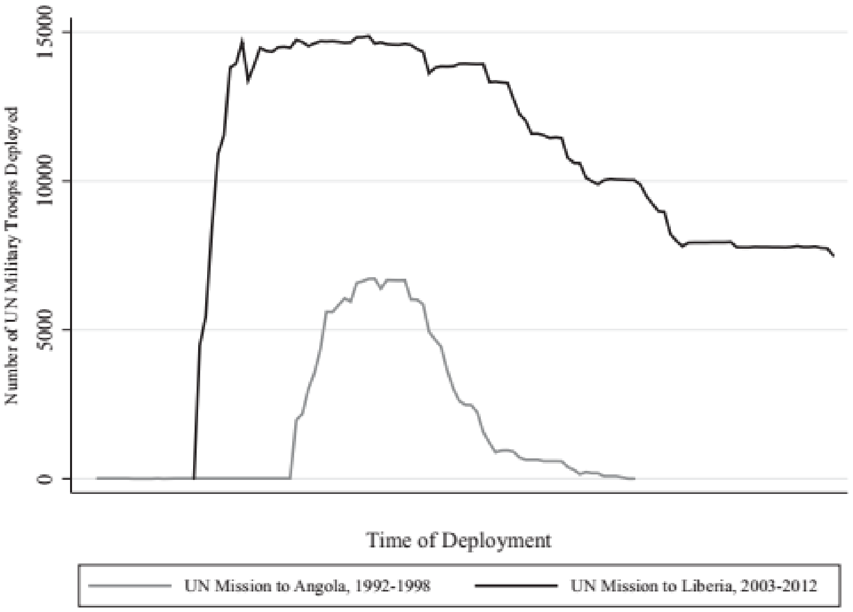

Figure 3: time varying employment within and across UN intervention cases, Source: Hultman et.al. (2014, 3)

Thirdly, regarding the coding of conflict initiation and termination, Fortna (2004a) and Sambanis (2008) reveal a discrepancy linked to conflict termination. As they define the goal of intervention as self-sustainable peace, they include the withdrawal of UN personnel into the list of termination events, because self-sustainable peace is per definitionem realized only after the UN left. However, other studies do not use this event as termination point, especially since the focus of must studies is explicitly dedicated to the impact of active UN involvement in the post-conflict period. Generally, the issue of missing or indecisive start and end point of conflicts can cause overall unpredictable bias as in the case of Diehl et.al. (1996, 695) who mentions in a footnote to one of the regression tables that 16 out of 262 conflict dyads had to be dropped.

Fourthly, regarding the coding of recurrence and simultaneous conflicts, since the different use of threshold does not obviously influence the number of units amongst the 16 studies, it might be that the coding of recurrence and simultaneous conflicts cause differences. For example, Fortna´s (2004a) estimation uses country-based dyads where recurrence of an already settled conflict is not coded as peace failure if another war in the country was still ongoing; whereas the majority uses conflict-based dyads where a government can have conflicts with several groups simultaneously.

In sum, most of the studies under consideration use the 25 deaths per year threshold, monthly data and conflict-based dyads which provides them with a solid basis for estimations. Nevertheless, based on the focus of the studies such as civilian-targeted hostility or self-sustainable peace, coding in individual studies can get non-transparent.

3.3 Methodological considerations: theory first in the design-based approach

Correlation is not causation. Therefore, rigid assumptions must be fulfilled in order to accredit validity to a causal interpretation of statistical inference. In this section, the validity of identifying causal effects of UN interventions in the 16 studies is assessed through two out of the three methodological criteria. On the one hand, the selection bias of UN intervention assignment must be taken into consideration, on the other hand, a counterfactual reading of the studies must account for common support of “what would have happened if…”- potential outcomes.

3.3.1 Ignorable treatment assignment: accounting for selection bias

The challenge in observational studies is the fact that without random assignment of treatment, it cannot be assumed that the treated group would not have been systematically different from a control group anyway. Just as medical care is more likely to be given to the sicker patients and just as firemen who fight against bigger fires are involved in more destructive happenings, in the same way UN interventions can possibly be employed in the neediest cases and therefore be associated to most deathly and devastating conflicts. However, this must not testify inefficiency of UN interventions. The success benchmark must be set in relation to the circumstances, therefore it is necessary to know whether the UN picks the easy cases with high prospect of successful interventions or indeed gets involved into the hardest conflicts where there is actually “no peace to keep” (UN General Assembly 2015, 9).

The standard way to deal with selection bias is adding control variables to the regression equation, which account for changes in both the conflict outcome and UN interventions. Sambanis (2008, 18) adds the mandate type and capacity as a manifestation of the interest of the Security Council veto powers and argues that their interest drives the Security Council decision. Other variables were used by Doyle and Sambanis (2013) and Sambanis (2008) such as a Cold war dummy to account for a time of Security Council consensus stalemate, dummies for Europe and other regions and a dummy for former colonies to account for the affiliation of Security Council members. While these variables of UN and country specific characteristics show no explanatory power, Fortna (2004a, 273) considers conflict specific characteristics and concludes that the UN is less likely to be employed in conflicts that ended with the victory of one side, simultaneously with less likelihood of going to a place with a strong government army. Since victory is associated with longer peace after war (Fortna 2004a, 286), this means that the UN is not picking the easy cases.

Interestingly, Gilligan and Stedman (2001) and Fortna (2004a) find that an employment after a treaty was signed is also less likely.

Sambanis (2008) adds some theoretical justifications for assuming randomness in the Security Council placement of interventions, which are taken up in the research design considerations of Collier et.al. (2008): The fact that bargaining within the UN is not need-driven but highly political and unsystematically complex and furthermore the fact that the actually sent interventions might look very different in the field than it was planned in the headquarter, UN interventions´ assignment to conflicts can be seen as kind of random.

Matching was presented as valid and feasible way of dealing with selection bias through making up most similar control units to establish a valid counterfactual. While one third of the 16 research papers account for the fact that UN interventions are not employed randomly to conflicts only five studies employ matching techniques prior to their statistical effect estimation. The matched samples in Gilligan and Sergenti (2008, 112) and Hultman et.al. (2014, 10) show divergences in terms of effect direction and significance, because the former study substantially decreases the size of the intervention effect, whereas in the latter it increases the size and significance level In the first study, UN intervention impact drops from 2.925 (2.01) to 0.720 (-0.44) and in the second study, the effect of troops increases from – 0.130 (0.051)* to -0.14 (0.021)** after matching.

3.3.2 Common support region for counterfactuals: the reality check

Compared to taking control units as the proxy for a potential outcome, the counterfactual reading of regressions and hazard models lies in assuming a different potential status of the treated unit itself to which the actual status is compared to. In this section, first the appropriateness and limitations of the potential outcome model is evaluated by analysing a scholarly debate about one of the studies. Then, the explicit use of counterfactuals in the other studies is investigated.

In their review of Doyle and Sambanis (2000), King and Zeng (2007) criticize the lack of a common support region for counterfactuals. In one of the model specifications, Doyle and Sambanis (2000) included a dummy variable for the seven multidimensional peace operations in their dataset and received a 23 times larger odds ratio of successful peacebuilding as outcome. Twisting the [0,1] notion of this the dummy variable values creates a counterfactual that displays the peacebuilding process of the intervened seven countries if the UN had not intervened and vice versa the peace outcome of the leftover 115 non-intervened countries if the UN had send personnel. King and Zeng (2007) take this counterfactual to show that such an extreme use of counterfactuals that have no common support within the data. This extrapolation beyond the observed data results in massive model dependence of the findings. Generally, statistical methods do a good job in finding the proper parameters that fit the model best, however, it is the analysts' job to select the right family of models first. Figure 4 illustrates that the comparison between choosing a linear and a quadratic model might not cause trouble when analysing the available data, but it will yield different results as soon as counterfactuals require the extrapolation outside the data.

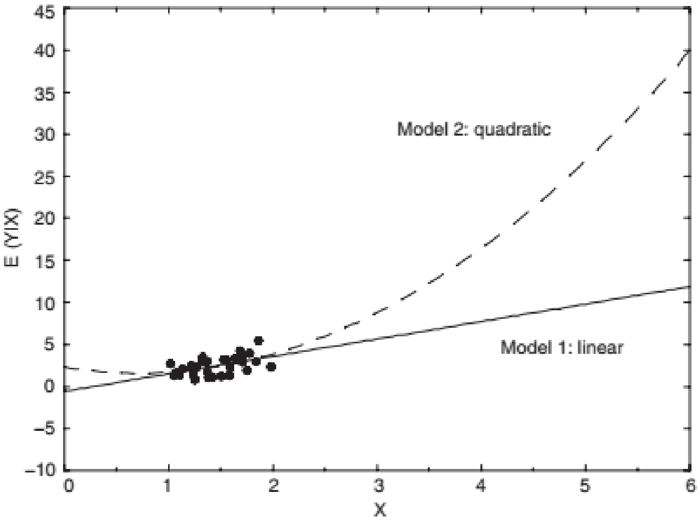

Figure 4: Model dependence beyond data support, Source: King and Zeng (2007, 184)

By means of the additional interaction between the multidimensional peacekeeping dummy and conflict duration, King and Zeng (2007, 195) show that the original model and the new interaction model conclude in different outcomes for the values of counterfactuals.

This debate on invalid causality can be seen as straw-man argumentation, because Doyle and Sambanis (2007), after being accused of using irrational and extreme counterfactuals, reject this accusation and clarify that it was never intended to read the regression analysis of conditional effects in a manner of purely hypothetical counterfactuals (Doyle and Sambanis 2007, 219). Overall, this debate reveals the question of the "potential" characteristic of potential outcomes. To filter the causal effect of a particular variation of the cause, counterfactuals rest on a ceteris paribus notion However, by completely reversing the actual state of affairs, a sphere is entered where mechanisms cannot be verified empirically any longer. In the Doyle and Sambanis (2000) case, having 115 UN instead of seven multidimensional interventions in the 122 civil wars makes up a potential world that would be too different from reality to hold the ceteris paribus assumption. However, counterfactuals must not be so extreme as they were treated in the just discussed review paper. It can be defined as any kind of clearly articulated change, such as from a lower quartile to a higher quartile, where a justifiable ceteris paribus notion does not lose the common support area.

Several of the studies under consideration explicitly used counterfactuals to estimate the impact of UN interventions. Three authors rely on a baseline to start their counterfactual reasoning. For example, Collier et.al. (2008, 473) create a clearly defined and theoretical feasible, if however not real life, scenario with four times higher military spending and an economic growth rate of 10%. Compared to the baseline hazard risk of 40%, this would decrease the hazard risk to 36.7%. Next, Beardsley (2012, 354) takes the same strategy and calculates the stepwise difference in hazard rate after six days and a month of different kinds of UN involvement. Troops, for example, cause a decline in the belligerent´s ability to compromise from 0.49% to 0.04% and a massive decline in the probability for a clear one-sided victory from 18% to 0.003% after six days. Seybold (2007, 34ff) uses the unique counterfactual of "lives that would not have been saved without UN intervention" and takes the mortality rate of a civil war population which is caused by hunger, disease and violence as counterfactual base. The after-intervention mortality rate is compared to this baseline in order to find out if military intervention successfully tackled the root causes of civilian deaths. His evaluation of the 17 UN intervention in the 90s shows that 9 indeed succeeded to save lives, but also two interventions actually worsened the situation for civilians (Seybold 2007, 270).

Three further studies simply take the counterfactual of "no intervention" to establish the impact of UN interventions. By doing so, Hegre et.al. (2011) simulate a decreasing effect of recurrence risk for multidimensional, transformational interventions as well as for higher military spending. Hultman et.al. (2013, 2014 and 2016) also take the comparison to zero UN employment and calculate that 10.000 more troops reduce battlefield deaths of 73% Hultman et.al. (2014, 9). Furthermore, Fortna (2004b) attest a decreasing impact of 87% to UN intervention in general. Based on this large sized effect, Hultman et.al. (2014) form a strong appeal to the UN decision maker to increase their troop number. However, methodologically, in accordance with the debate between Doyle and Sambanis (2000) and King and Zeng (2007), this approach of radically taking "no intervention" as comparison can be criticized of being unrealistic and out of the common support area where the ceteris paribus condition would fail. In contrast, the approach of taking a real life or at least theoretically realistic baseline is a sound way to yield valid impact evaluations. Overall and in line with the design based approach´s imperative to clearly model the theory first, the main benefit of the counterfactual reading is the focus on explicitly stating the variation in the causal variable under concern and thereby revealing the assumptions and limitations of what can actually be known as credible causality in observational studies (Morgan and Winship 2007, 13).

In conclusion, the goal of the comparative review was to gain insight on potential contextual and coding sources of differences and similarities in the findings of research on UN interventions and to shed methodological light on the validity of the causal interpretation of the statistical inference. The contextual criteria revealed that not all studies acknowledge the Cold War cut in the choice of their sample, although accordingly control variables might be employed. In the short-term perspective, the UN mechanisms of building a buffer zone and establishing a credible guarantee in the commitment dilemma of belligerent is commonly accepted, however, the results for the various mandates on short-term goals are contradictory. In the long-term perspective, there is dissent on the prevalence of the "politics first" and "economy and security first" approach, although the empirical results of the studies under investigation support the latter approach more than the first. Methodologically, regarding a design-based approach of research, some studies make the counterfactuals explicit. While some studies refer to reasonable baseline for comparing the change through UN intervention to, other studies use the scenario of zero UN intervention, which can be argued to be too unrealistic as to be counted as valid counterfactual.

Conclusion

This thesis investigated how UN interventions have been evaluated and whether the results establish valid causal impact evaluations. By filing in the gap of systematic comparison, the review of 16 studies on UN interventions, gave insight in how the choice of different definitions of success, data coding, research scope and research designs influence the respective findings. The analysis was done by establishing and applying contextual, data coding and methodological criteria.

Firstly, contextual considerations show no uncontested coherent finding regarding the direction and scope of UN intervention effect. In the short run, there is great support for the position that UN intervention can indeed reduce battlefield deaths if properly equipped; the ability of ending conflicts is however highly disputed. Regarding the long-term approach of interventions, the UN´s official approach of "politics first" does not find as much empirical support as the approach of giving priority to development and security through economic and military means.

Secondly, data coding considerations revealed that the majority of studies use fine-grained coding of 25 deaths threshold, monthly data and conflict-based dyads. No systematic link between coding and finding could be detected across the studies, however, the scope of this thesis does not employ a robustness check on coding sensitivity within and across studies.

Thirdly, on the background of a scholarly debate, methodological considerations point out that the use of "zero UN interventions" is prone to be too far from reality as to serve as a comparison. Several studies showed good examples of how to use reasonable baselines as counterfactuals.

As a result of the comparative review, implications for UN policy making cannot be articulated in a clear way based on uncontested finding. However, there is a tendency of greater empirical support for setting priority on development and security issues in post-conflict settings through economic and military interventions rather than focusing on "politics first". Nevertheless, the diverse and partly contrary findings show the need to further refine explanatory theories and align them with robust data and statistical methods.

To proceed in the endeavour of gaining comparative insight, the implications for researchers is to provide replication dataset by default and explicitly explain coding in order to make the results comparable and accountable.

A clear limitation of the presented comparative review is the limited scope of the papers under consideration. Especially including the broad literature on the selection indicators for UN interventions could contribute valuable insight. Additionally, since the terrorism of the past decade marked another acknowledged shift in the nature of conflicts, research on the UN involvement considering this type of conflicts must be considered. Further comparative research should expand the qualitative approach of this thesis with quantitative robustness checks. This can be done by systematically framing data variations with different coding patterns like 25 deaths threshold for the whole time period, for during and for after the Cold War in a first step. In a second step, statistical models of different causal theories are run with these systematic data variations to detect incoherence, similarities and remaining divergences across the research studies.

List of references

Beardsley, K. 2012. "UN Intervention and the Duration of International Crises." Journal of Peace Research 49 (2): 335–49. doi:10.1177/0022343311431599.

Beardsley, Kyle. 2013. "The UN at the Peacemaking–peacebuilding Nexus." Conflict Management and Peace Science 30 (4): 369–386.

Beardsley, Kyle, David E. Cunningham, and Peter B. White. 2015. "Resolving Civil Wars before They Start: The UN Security Council and Conflict Prevention in Self-Determination Disputes." British Journal of Political Science, January, 1–23. doi:10.1017/S0007123415000307.

"Cochrane Handbook for Systematic Reviews of Interventions." 2017. Accessed June 1. http://handbook.cochrane.org/.

Collier, Paul, Anke Hoeffler, and Måns Söderbom. 2008. "Post-Conflict Risks." Journal of Peace Research 45 (4): 461–78. doi:10.1177/0022343308091356.

DPO/DFS- Department of Peacekeeping Operations and Department of Field Support. 2009. "A New Partnerhsip Agenda- Charting a New Horizon for UN Peacekeeping." http://www.un.org/en/peacekeeping/documents/newhorizon.pdf.

Diehl, Paul F., Jennifer Reifschneider, and Paul R. Hensel. 1996. "United Nations Intervention and Recurring Conflict." International Organization 50 (4): 683–700.

Dorussen, Han, and Theodora-Ismene Gizelis. 2013. "Into the Lion's Den: Local Responses to UN Peacekeeping." Journal of Peace Research 50 (6): 691–706.

Doyle, Michael W., and Nicholas Sambanis. 2000. "International Peacebuilding: A Theoretical and Quantitative Analysis." American Political Science Review 94 (04): 779–801. doi:10.2307/2586208.

Fortna, Virgina Page. 2004a. "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace After Civil War." International Studies Quarterly 48: 269–92.

Fortna, Virginia Page. 2004b. "Interstate Peacekeeping: Causal Mechanisms and Empirical Effects." World Politics 56 (4): 481–519. doi:10.1353/wp.2005.0004.

Fortna, Virginia Page, and Lise Morjé Howard. 2008. "Pitfalls and Prospects in the Peacekeeping Literature." Annual Review of Political Science 11 (1): 283–301. doi:10.1146/annurev.polisci.9.041205.103022.

Gates, Scott, and Håvard Strand. 2004. "Modeling the Duration of Civil Wars: Measurement and Estimation Issues." Peace Pesearch Institute Oslo(PRIO).

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2016. Impact Evaluation in Practice, Second Edition. The World Bank. doi:10.1596/978-1-4648-0779-4.

Gilligan, Michael J., and Ernest J. Sergenti. 2008. "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." Quarterly Journal of Political Science 3 (2): 89–122. doi:10.1561/100.00007051.

Gledisch, Nils Petter, Peter Wallensteen, Mikael Erikson, Margaret Sollenberg, and Havard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." Journal of Peace Research 39 (5): 615–37. doi:10.1177/0022343302039005007.

Howard, Lise Morjé. 2008. UN Peacekeeping in Civil Wars. Cambridge University Press.

Hultman, Lisa, Jacob Kathman, and Megan Shannon. 2013. "United Nations Peacekeeping and Civilian Protection in Civil War." American Journal of Political Science 57 (4): 875–91.

Hultman, Lisa, Jacob Kathman, and Megan Shannon. 2014. "Beyond Keeping Peace: United Nations Effectiveness in the Midst of Fighting." American Political Science Review 108 (04): 737–53. doi:10.1017/S0003055414000446.

Hultman, Lisa, Jacob D. Kathman, and Megan Shannon. 2016. "United Nations Peacekeeping Dynamics and the Duration of Post-Civil Conflict Peace." Conflict Management and Peace Science 33 (3): 231–49. doi:10.1177/0738894215570425.

ICISS International Commission on Intervention, and State Sovereignty. 2001. “The Responsibility To Protect Report.” International Development Reserach Centre.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. “Recent Developments in the Econometrics of Program Evaluation.” Journal of Economic Literature 47 (1): 5–86. doi:10.1257/jel.47.1.5.

Keele, Luke. 2015. “The Statistics of Causal Inference: A View from Political Methodology.” Political Analysis 23 (03): 313–35. doi:10.1093/pan/mpv007.

King, Gary, and Langche Zeng. 2007. “When Can History Be Our Guide? The Pitfalls of Counterfactual inference1.” International Studies Quarterly 51 (1): 183–210.

Kreutz, Joakim. 2010. “How and When Armed Conflicts End: Introducing the UCDP Conflict Termination Dataset.” Journal of Peace Research 47 (2): 243–50. doi:10.1177/0022343309353108.

Morgan, Stephen L., and Christopher Winship. 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. 1 edition. New York: Cambridge University Press.

Sambanis, Nicholas, and Michael W. Doyle. 2007. “No Easy Choices: Estimating the Effects of United Nations Peacekeeping (Response to King and Zeng).” International Studies Quarterly 51 (1): 217–226.

Schulenburg, Michael von der. 2014. “Rethinking Peacebuilding: Transforming the UN Approach.” IPI-International Peace Institute.

Seybolt, Taylor B. 2007. Humanitarian Military Intervention: The Conditions for Success and Failure. Oxford University Press.

Stamnes, Eli, and Kari M. Osland. 2016. “Synthesis Report: Reviewing UN Peace Operations, the UN Peacebuilding Architecture and the Implementation of UNSCR 1325 | NUPI Report no.2, 2016.” Norwegian Institute of International Affairs.

United Nations. 2017a. “UN Peacekeeping Operation Fact Sheet March 2017.” http://www.un.org/en/peacekeeping/documents/bnotelatest.pdf.

———. 2017b. “Peacekeeping Operations. United Nations Peacekeeping.” Accessed May 28. http://www.un.org/en/peacekeeping/operations/.

United Nations General Assembly, and Security Council. 2015. “Report of the High -Level Independent Panel on Peace Operations on Uniting Our Strengths for Peace: Politics, Partnership and People | United Nations Official Document A/70/95–S/2015/446.”

White, Howard. 2009. Some Reflections on Current Debates in Impact Evaluation. International Initiative for Impact Evaluation New Delhi, India.

Appendix

Appendix I: Comparative Review of 16 research papers on UN interventions

Studies on interstate wars, prevention, simulation (Not included in the quantitative comparison)												
Beardsley (2012)	1946-2002	Total	Ongoing conflict	interstate	Mandates: assurance, diplomacy, military, intimidation		275	daily	matching			
Beardslye et.al. (2015)	1960-2005	mixed	prevention	intrastate	Mandates: resolution, sanctions, diplomacy, condemnation, force)				matching			
Diehl et.al. (1996)	1946-1988	Before	Post-conflict	Interstate	Mandates: passive diplomatic, active diplomatic, observer/military, other	1000/total	262		none			
Hegre et.al. (2011)	1970-2008	Mixed	Post-conflict	Intrastate	Mandates: traditional, transformational	25/year and 1000/year		yearly	Simulation of peacekeeping effect			
Studies using logit, probit and negative binomial regression models												
Beardsley (2013)	1946-2005	Total	Both		Mandates: Peacekeeping, sanctions, diplomacy, condemnation, others)	25/year		monthly		Probit regression	Duration of peace in months → measuring duration of war and duration of peace	
Doyle and Sambanis (2000)	1944-1997	total	Post-conflict	Intrastate	Mandates: monitoring/observers, traditional incl. consent based military, multidimensional, peace enformcement	1000/total	124		none	Logit regression	Dichotomous peace (lenient and strict version) → measuring probability of success/failure of peace	multidimensional 22.99** (2.87) any UN intervention 3.82* (2.13)

Fortna (2004a)	Collier et.al. (2008)	Collier et.al. (2004)	Studies using Cox proportional hazard models	Sambanis (2008)	Hultman et.al. (2014)	Hultman et.al. (2013)	Dorussen and Gizelis (2013)
1944-1997	1960-2002	1960-1999		1945-2000	1992-2011	1991-2008	1989-2006
Total	Mixed	Mixed		Total	After	After	After
Post-conflict	Post-conflict	Post-conflict		Post-conflict	Ongoing conflict	Ongoing conflict	Post-conflict
intrastate	Intrastate	Intrastate			Intrastate	Intrastate	Intrastate
Not exclusively UN, Binary and mandates: consent-based, enforced, none	UN intervention spending	General intervention. Pro government/pro rebels and economic/military		Mandates: observers, multidimensional, traditional, enforcement	Mandates: troops, observers, police	Mandates: troops, observers, police	Peacekeeping as replacement/ improvement of local authority
25/year		1000/year			25/year	25/year	-
115	68	77		128	145	36	4265 policy events
monthly	monthly	Monthly		monthly	monthly	monthly	-
logit	Dummy variable procedure, Th eoretical consideratinos of randomness			Instrumental variables, theoretical explanation of randomness	matching	matching	none
Cox proportional hazard model	Cox proportional hazard model	Cox proportional hazard model		Logit regression	Negative binomial regression	Negative binomial regression	Probit regression
Duration of peace in months → measuring hazard of conflict recurrence	Duration of peace in months → measuring hazard of conflict recurrence	Duration of war in months → measuring hazard of peace		Binary participatory peace variable, → measuring probability of participatory peace	Number of battlefield deaths	Number of civilian deaths through government and rebel violence	Probability of hostile/ cooperative reaction (not mutually exclusive) of rebels and government
UN peacekeeping 0.51 (0.19)+ observer 0.19 (0.22) js* traditional 0.14 (0.15) + multidimensional 0.47 (0.33) js* enforcement 0.57 (0.57) js*	UN peacekeeping expenditure: -0.427	Pro rebel general military intervention: 0.1994 (0.0933)*		Short-run multidimensional peacekeeping 3.1039 (1.0290)* any UN 1.9247 (0.6118)*	troops -0.130 (0.051)* observers 2.732 (1.344)	troops -0.53 (0.09)*** police -9.90 (1.55)*** observer 21.76 (3.96) ***	Peacekeeping replacing replace to cooperative reaction -0.15 (0.06)** interactions rebel*replace r cooperation -1.14 (0.14)** rebel*replace reaction to hostility

	Author	Sambanis (2008)	Hultman et.al. (2016)	Gilligan and Sergenti (2008) long-term	Gilligan and Sergenti (2008) short-term	Fortna (2004b)
Contestual criteria	Timeperiod	1945-2000	1989-2010	1988-2003	1988-2003	1946-1997
	Timeperiod with cut-off Cold War (before-after-total-mixed)	Total	After	After	After	Total
	Time horizon (ongoing conflict- post-conflict)	Post-conflict	Post-conflict	Both	Both	Post-conflict
	Type of conflict (interstate-intrastate-both)		Intrastate	intrastate	intrastate	both
	Coding of UN intervention (binary, mandates)	Mandates: observers, multidimensional, traditional, enforcement	Mandates: troops, observers, police	Binary	Binary	Not exclusively UN
Data coding criteria	Deaths threshold (25/year, 1000/year, 1000/total)		25/year	25/year	25/year	1000/total
	Number of units of analysis	128	129	156	156	48
	Interval precision (daily, monthly, yearly)	monthly	monthly	monthly	monthly	
Methodological criteria	Control for selection bias (matching, none, other)	Instrumental variables, theoretical explanation of randomness	none	matching	matching	logit
Statistical analysis p-values: + 0.1, * 0.05, ** 0.01, *** 0.001	Statistical model	Cox proportional hazard model	Cox proportional hazard model	Cox proportional hazard model	Cox proportional hazard model	Cox proportional hazard model
	Dependent variable	Duration of peace in months → measuring hazard of conflict recurrence	Duration of peace in months → measuring hazard of conflict recurrence	Duration of peace in months →conflict recurrence	Duration of war in months → hazard of peace	Duration of peace in months → measuring hazard of conflict recurrence
	Point estimate of causal impact of UN intervention	long run: UN intervention 0.54 (0.16)* traditional UN 0.48 (0.17)*	Troops: -0.0005 (0.0001)** Observers -0.0005 (0.004) Police 0.0005 (0.0005)	UN in post-conflict: 0.144 (-2.18)	UN in ongoing: 0.720 (-0.44)	General peacekeeping 0.13 (0.11)*

Appendix II: Author's Declaration

I, the undersigned Theresa Weippert hereby declare that I am the sole author of this thesis. To the best of my knowledge this thesis contains no material previously published by any other person except where due acknowledgement has been made. This thesis contains no material which has been accepted as part of the requirements of any other academic degree or non-degree program, in English or in any other language. This is a true copy of the thesis, including final revisions.

Date: 19th June 2017

Name: Theresa Weippert

Signature: