COMPASSS Working Paper 2010-59

www.compasss.org

How Good Are Your Counterfactuals?

Assessing Quantitative Macro-Comparative Welfare State Research with Qualitative Criteria

Patrick Emmenegger
Centre for Welfare State Research
University of Southern Denmark

Abstract: All causal statements based on historical data – both in qualitative and quantitative social research – rely on counterfactuals. In quantitative research, scholars attempt to arrive at valid counterfactuals by emulating an experimental design. However, because of treatments that are impossible to manipulate and the non-random assignment of data to treatment and control groups, causal statements are often based on invalid counterfactuals. In contrast, in qualitative research, scholars attempt to arrive at valid counterfactuals by probing the historical and logical consistency of counterfactuals and by acknowledging the interconnectedness of events. Criteria to evaluate counterfactuals have been developed, especially in the international relations literature. These criteria allow for a discussion of the quality of counterfactuals used in causal statements. In this article, we suggest using these qualitative criteria to evaluate counterfactuals in quantitative macro-comparative welfare state research. We argue that these criteria can help us identifying erroneous causal inferences in quantitative research based on historical data. The usefulness of such an approach is illustrated using the seminal contribution of Alesina and Glaeser (2004) on racial-linguistic fractionalization and social expenditure.

"Like it or not, then, counterfactual conditionals are unavoidable."
Stanley Lieberson (1985: 48)

"I now believe that counterfactuals play the crucial role in comparative thinking."
Adam Przeworski (2007: 479)

1 Introduction

According to King et al. (1994: 79), the 'fundamental problem of causal inference' is that causal statements are ultimately based on the comparison of something that did occur (the 'factual') and something that did *not* occur (the 'counterfactual'). Since we can per definition not observe the counterfactual – the outcome if the hypothesized cause would not have occurred – we never know a causal inference for certain.

For example, if we are interested in the role of incumbency on chances of re-election, we can compare the electoral success of incumbents with the electoral success of challengers. However, we cannot get evidence on the electoral success of the incumbents *if they were not incumbents*. Rather, we have to rely on the assumption that – after controlling for confounders – incumbents and challengers are reasonably comparable to allow for an evaluation of the causal effect of incumbency on re-election.

Some research designs are more apt to deal with the fundamental problem of causal inference. Experimental designs randomly allocate the units of analysis to two groups, the so-called treatment and control groups. If the sample is sufficiently large, these two groups should be equal on all relevant confounders. Subsequently, a treatment is applied to one group, but not to the control group. Due to randomization and the law of large numbers, the observed difference is then equal to the causal effect of the treatment (Holland 1986).

Unfortunately, data in macro-comparative welfare state research is nothing like that. For numerous reasons, the randomization of the units of analysis is not possible and treatments cannot be applied in a discriminatory manner. Researchers thus often employ the assumption

of conditional independence, which seems to allow for causal inference even though the treatment and control groups do not have the same hypothesized average values (Collier et al. 2004: 33). Conditional independence implies that by incorporating the appropriate control variables, the comparison of the average values of the treatment and control groups corresponds to the causal effect of the treatment. Of course, whether the observed difference really reflects the causal effect of the treatment depends on the assumption of conditional independence: Are there no omitted variables? Are there no further selection effects? Is the model specified in the correct functional form? In any case, there is *no* way to know for sure whether the assumption of conditional independence has really been met (King et al. 1994: 79; Collier et al. 2004: 33; Freedman 2010).

Holland (1986) and Rubin (1978), two of the developers of the counterfactual approach to causation described above (the Neyman-Rubin-Holland model), thus conclude that their model for causal inference is not very apt for dealing with non-randomized data. According to Brady (2004: 61), some statisticians consider conditional independence a "chimera – seldom justifiable and usually accepted by the researcher as a matter of pure faith and nothing more."

In macro-comparative welfare state research, we are facing a situation in which we want to use quantitative methods such as regression analysis to make causal inferences. However, our research designs violate some of the most fundamental assumptions. Thus, the counterfactuals we use for making causal statements might not be valid. What can we do about this?

In this article, we propose to look over the fence and to learn from qualitative methods. Qualitative researchers normally use counterfactuals to assess causality for singular events. For example, the assertion that the assassination of archduke Franz Ferdinand in Sarajevo caused the outbreak of World War I hinges on the counterfactual that if the archduke had survived, there would have been no World War I. Probing the historical and logical consistency of counterfactuals (could the assassination have failed?) and acknowledging the

interconnectedness of events (were there forces that would have led to war even in the absence of the assassination?) can subsequently help assessing the validity of this counterfactual. To this end, qualitative methodologists have developed criteria that allow for a discussion of the quality of counterfactuals used in causal statements.

However, we can also use counterfactual theorizing to evaluate regularities. According to King and Zeng (2007: 185), every quantitative analysis "is making a counterfactual prediction, and every one needs to be evaluated by the same ideas well known in qualitative research". In a similar vein, Lebow (2010: 6) argues that counterfactual worlds should be used "to probe the causes and contingency of the world we know". Thereby, the goal is not to make the case for alternative worlds and engage in reflections on 'what could have been', but to 'poke counterfactual holes in covering laws' (Tetlock and Lebow 2001).

In sum, we suggest using qualitative criteria for evaluating counterfactuals in quantitative macro-comparative welfare state research. We argue that these criteria can help us identifying erroneous causal inferences in quantitative research based on historical data. However, the ultimate goal is not to simply highlight problems in causal statements. Rather we demonstrate below that a critical discussion of counterfactuals can help us formulating more realistic hypotheses and thus improve macro-comparative welfare state research.

We proceed as follows. In the next section, we demonstrate why the theory of causation, which underlies most quantitative social science research (the Neyman-Rubin-Holland model), is not apt to deal with the kind of data researchers are confronted with in macro-comparative welfare state research. Subsequently, we develop a list of criteria for the evaluation of counterfactuals. In section 4, we apply these criteria to the literature on welfare chauvinism. More precisely, using these criteria, we evaluate the influential study by Alesina and Glaeser (2004) on the relationship between racial-linguistic fractionalization and social expenditure. A final section concludes.

2 Quantitative macro-comparative welfare state research and the counterfactual problem

Macro-comparative welfare state research often uses nation-states as its unit of analysis. Typically working with populations of 16 to 30 cases, researchers examine cross-national (and occasionally temporal) variation in social spending, labour market regulation or employment rates using a set of theoretically informed independent variables and techniques such as regression analysis.

Although rarely explicitly stated, these analyses rely on the assumption that – after controlling for some variables – the observed influence of each variable corresponds to the result found under conditions of random assignment (Lieberson 1985: 200; Shalev 2007; Freedman 2010). However, in macro-comparative welfare state research, this assumption is hardly ever valid.[1] By focusing on mature welfare states in advanced industrialized democracies, researchers accept that historical and political processes have shaped their population of cases. For example, Tilly (1975: 15) notes that "the Europe of 1500 included some five hundred more or less independent political units, [while] the Europe of 1900 has about twenty-five". Thus, today's observable nation-states are "a highly contingent set of (surviving and constructed) cases" (Ebbinghaus 2005: 138).

Recent contingent and thus non-random changes include cases such as the German reunification. In 1990, Western Germany was the second richest country in the European Community with a gross national product per capita of 40'200 DM (after Luxembourg). One year later, the reunified Germany was the sixth richest country in the European Community, just after Italy, with a gross national product of 36'000 DM per capita (Czada 1998: 26). At the same time, Germany's population grew by about 26 percent (Armingeon et al. 2008). Similarly, in May 2004 and January 2007, ten Central and Eastern European countries joined the European Union that – despite being advanced industrialized democracies – have *systematically* different historical backgrounds than the Western European member states.

There is an abundance of further examples of this kind. However, one thing is clear: In no way can we assume "that some sort of random assignment model is operating or closely approximated" (Lieberson 1985: 201).

Similarly, macro-comparative welfare state research typically relies on observational data. In contrast to experimental data, real world events and processes produce observational data, which are therefore not subject to the direct control of investigators (Brady and Collier 2004: 299). Thus, the random allocation of cases to treatment and control groups cannot fabricate independence of confounders. Consequently, researchers can never be sure that other characteristics, which might have an effect on the dependent variable in the analysis, did not influence the assignment of cases to the treatment and control groups (Lieberson 1985: 31).

For example, quantitative analyses often point to the strength of left parties to explain high levels of social expenditure in Scandinavia (e.g. Huber and Stephens 2001). However, in the absence of an experimental research design, which is not possible for this kind of research questions, we can never be *sure* that there is no unobservable 'Scandinavian variable' that causes these high levels of spending (and possibly also the electoral strength of the left). This 'Scandinavian variable' might be some cultural background variable, related to geographical peculiarities, the particular Scandinavian climate or the result of pan-Scandinavian historical experiences. This 'Scandinavian variable' might be rather unlikely, but the important point is: Statistically, we cannot rule it out (King et al. 1994: 79; Freedman 2010: 62).

As a consequence, researchers normally incorporate a set of independent variables on the right-hand side of the regression equation in order to 'control' for confounders. This way, researchers attempt to achieve conditional independence that would allow them to draw causal inferences on the basis of observational data. However, we rarely know enough about a causal relationship to be able to specify a model that incorporates all potential confounders (Lieberson 1985). And even if we knew all potential confounders, the small number of cases

available would prevent us from controlling for all of them. After all, there are only 27 EU member states and 30 OECD member states. The facts that the interpretation of regression models with more than three independent variables becomes increasingly complex (Achen 2002) and that including relevant control variables might not even reduce omitted variable bias (Clarke 2005) further aggravates the problem. Thus, the assumption of conditional independence is often very hard to justify (Brady 2004: 61; Freedman 2010: 62).[2]

Thus, the reliance on non-randomized observational data creates problems for causal inference. Although we might be able to observe regularities, we cannot infer from these regularities that there is also a causal relationship (Brady 2008: 230). In order to infer from regularities on causality, we still need a theory of causation. Quantitative social science normally relies on the counterfactual theory of causation as developed by Neyman, Rubin and Holland (King et al. 1994: 76-82; Collier et al. 2004: 25-26). This theory of causation is based on the comparison of something that did occur (the 'factual') and something that did *not* occur (the 'counterfactual'). King et al. (1994: 79) describe this reliance on counterfactuals in causal statements as the "fundamental problem of causal inference". Since we can never observe the counterfactual – the hypothetical outcome if a particular prior event had not occurred – we have to rely on assumptions about it. Ideally, we can rely on an experiment in which we introduce a treatment and evaluate its causal effect, while eliminating rival explanations by randomizing the assignment into treatment and control groups.[3] However, as noted above, data in macro-comparative welfare state research is nothing like that. It is based on real world observations and not randomized. As a result, causal inferences on the basis of observational data might be highly misleading. Put differently, we should not have much confidence in causal inferences that are based on non-randomized observational data and counterfactuals because these counterfactuals are likely to violate some of the most fundamental assumptions for causal inference. Rubin (1978) and Holland (1986) are thus

critical of applications of their counterfactual theory of causation to non-randomized data, a feeling that is widespread in the statistical community (Brady 2004: 61).[4]

We should not conclude from this discussion that we would be better off abstaining from the analysis of causal effects or that we should refrain from using quantitative methods.[5] Causality is the ultimate goal of most social science research and quantitative methods are a core part of our methodological tool-kit. Rather, we should conclude that additional guidelines are needed that help us evaluating the counterfactuals on which our causal inferences are based upon. Thus, instead of throwing out the baby with the bath water (abandoning regression analysis) or of looking the other way (ignoring the problem), we should ask ourselves whether our counterfactuals are "reasonable" (Lieberson 1985: 48).

In recent contributions, King and Zeng (2006, 2007) have introduced a new software package, which helps researchers examine whether seemingly causal inferences rely on reasonable or 'miracle' counterfactuals. In an effort to complement their efforts, we develop in the next section criteria to evaluate counterfactuals in a qualitative way. While King and Zeng's (2006, 2007) approach uses statistical criteria to evaluate whether causal inferences are a function of modelling choices (and thus relying on 'miracle' counterfactuals), we suggest using qualitative criteria to evaluate counterfactuals employed in causal inferences.[6]

3 Criteria for good counterfactuals

Counterfactuals have an odd position in social science methodology. Although their usage is often met with criticism, they are inevitable for causal statements in macro-comparative social research (Levy 2008: 629). Following Tetlock and Belkin (1996: 4), a counterfactual can be defined as a "subjective conditional in which the antecedent is known or supposed for purposes of arguments to be false". Put differently, counterfactuals are possible worlds in which the antecedent did not actually occur or occurred in a different way. In macro-comparative social research, causal statements often rely on (implicit) counterfactuals.

Unable to rerun history and manipulate the treatment, comparative research often relies on the Humean regularity approach. If, say, we observe that in Western Europe strong trade unions are associated with high levels of public social expenditure, while weak trade unions are associated with low levels of public social expenditure, we tend to conclude that trade union strength is the cause of high levels of public social expenditure.

However, as Brady (2008: 230) reminds us, association is not causation. In fact, the above statement about causality relies on counterfactuals. Suppose that other variables co-vary with trade union strength. An obvious candidate is left party power. Table 1 displays an example using fictional data: We have dichotomous data on left party, trade union strength and welfare state generosity for 18 countries. In six countries, we observe strong left parties, strong trade unions and generous welfare states. In seven countries, we observe strong left parties, weak trade unions and lean welfare states. In five countries, we observe weak trade unions, weak left parties and lean welfare states. Finally, we cannot observe the combination of weak left parties and strong trade unions anywhere.

Two observations can be made on the basis of Table 1. First, trade union strength and left party strength co-vary (Cramér's V = 0.44). Second, we can observe a perfection correlation between trade union strength and welfare state generosity (Cramér's V = 1.00). However, in the latter case, as soon as we take variation in left party strength into account, we are getting into trouble.

Table 1: Example: Counterfactual analysis and welfare state generosity

Row no.	Strong left party	Strong trade unions	Generous welfare state	Number of countries
1	Yes	Yes	Yes	6
2	Yes	No	No	7
3	No	No	No	5
4	No	Yes	?	0

Source: Ragin (2000: 105).

On the basis of Table 1, we are tempted to conclude that strong trade unions are the cause of high levels of welfare state generosity. However, this conclusion ultimately relies on the *counterfactual* assumption that in the case of weak left parties and strong trade unions – a combination that cannot be observed in the real world – we would observe generous welfare states. The problem is that we cannot *know* this, we can only *guess.* And in most cases, we do not clearly state this 'guess-work'.

In the international relations literature, the explicit usage of counterfactual theorizing is more common. Here, we can witness a burgeoning literature on the methodology of counterfactual analysis (e.g. Fearon 1991, 1996; Tetlock and Belkin 1996; Lebow 2000, 2010; Goertz and Levy 2007; Levy 2008). In the following, we use this literature to develop (best-case) criteria for good counterfactuals (see Table 2 for an overview).

Table 2: Checklist for good counterfactuals

---

1. **Clarity:** Specify the hypothesized antecedent, the consequent, the connecting principles and additional enabling conditions as clearly as possible
2. **Plausibility of the antecedent:**
    a. Make sure that the counterfactual antecedent is logically and historically possible (logical and historical consistency)
    b. Specify antecedents that require the alteration of as few historical facts as possible (minimal rewrite rule)
3. **Conditional plausibility of the consequent:**
    a. The connecting principles should be consistent with well-established theoretical laws (theoretical consistency)
    b. The connecting principles should be consistent with well-established statistical generalizations (statistical consistency)
    c. Minimize the number of causal steps between the counterfactual antecedent and the counterfactual consequent (proximity rule)
    d. Acknowledge the interconnectedness of events and consider the effects of second-order counterfactuals
4. **Projectability:** Formulate other implications that can be tested against data

---

Sources: Fearon (1991, 1996), Tetlock and Belkin (1996), Lebow (2000) and Levy (2008).

3.1 Clarity

As a general rule, researchers should specify as clearly as possible what is to be explained (the counterfactual consequent), what accounts for this outcome (the counterfactual

antecedent) and the mechanism linking the two (Tetlock and Belkin 1996: 19-21; Lebow 2000: 581-582; Levy 2008: 633-634). Furthermore, any enabling conditions and counterfactuals have to be stated as well. For example, in order to answer questions about the effect of a Richard Nixon Presidency (instead of John F. Kennedy Presidency) on the outcome of the Cuban missile crisis, we need to clearly state that the counterfactual case (a Nixon Presidency) relies on the additional counterfactual that the Republicans would have won the 1960 election. In contrast, in the case of the failed murder attempt on Archduke Franz Ferdinand in Sarajevo on June 28, 1914, no additional enabling conditions and counterfactuals are necessary as it is very plausible that Gavrilo Princip might have simply missed the Archduke (as it is for instance common to fail due to excitement).

3.2 Plausibility of the antecedent

Several factors affect the plausibility of the antecedent. First, a counterfactual antecedent should be logically consistent (Fearon 1991: 193; Tetlock and Belkin 1996: 21-23; Lebow 2000: 582; Levy 2009: 635). As argued by Elster (1978: 204-208), we cannot use antecedents in causal statements if we have a theory showing that the counterfactual antecedent could not have logically happened. To illustrate this, Elster discusses the famous thesis by Fogel (1964) on the role of railroads in the development of the American economy. Fogel (1964) argues using counterfactual analysis that 19th century America without railroads would have grown only slightly slower due to stronger incentives to develop the internal combustion engine sooner. Elster (1978) rejects this argument by noting that had the technology for the development for automobile engine been present at that time, we can also assume that the technology for the development of railroads would have been present.

Thus, logical consistency refers to connecting conditions and principles that are needed in order to make the antecedent logically possible. For example, the quantitative literature on welfare state spending has found that Christian democratic parties have a positive effect on

welfare state generosity. Christian democratic parties typically emerge in countries with proportional electoral systems. Thus, assuming the presence of a relevant Christian democratic party in a country such as Great Britain – implicitly done in quantitative analyses of welfare state generosity – forces us to further assume a different electoral system (proportional instead of majoritarian). However, a different electoral system would most likely lead to many more changes in the partisan composition of the British parliament, thereby making causal statements very difficult. Put differently, in the case of Christian Democracy in Great Britain, the enabling counterfactuals (i.e. counterfactuals needed in order to sustain the primary counterfactual) undercut the counterfactual antecedent (Lebow 2000: 582).

The problem of logical consistency is strongly related to the problem of historical consistency (Fearon 1991: 193; Tetlock and Belkin 1996: 23-25; Lebow 2000: 582-583; Levy 2008: 635). According to Max Weber (1949), good counterfactual cases should require as few changes as possible as compared to the real world ('minimal-rewrite rule'). Tetlock and Belkin (1996: 23) list three criteria: First, good counterfactual cases should start from the real world as known before asserting the counterfactual. Second, it should not require us to rewrite long stretches of history. Finally, it should not unduly disturb what we otherwise know about the original actors and their attitudes. The British example above clearly violates the minimal-rewrite rule since we are required to imagine how a counterfactual world would look like assuming proportional representation and the presence of a relevant Christian democratic party.

The minimal-rewrite rule, that is small changes to the real course of history, need not necessarily lead to small consequences. As the Archduke Franz Ferdinand example above illustrates, small changes may have tremendous consequences. Although it is an open question whether a failed murder attempt would have avoided the World War I, it suffices to

note that a failed murder attempt *may* have changed the historical course of action dramatically (Lebow 2010).

It is important to note that the problems of logical and historical consistency tend to interact. The more historical facts we have to change in order to get the necessary antecedent, the more enabling conditions and counterfactuals are needed to logically sustain this counterfactual antecedent. For example, speculating about alternative courses of history if Hitler had been a woman violates both the historical and the logical consistency criteria. First, we know that Hitler was no woman, a fact that cannot simply be changed. Second, if he had been a woman, numerous other historical *facts* would have to be changed too. It is for example doubtful that a female Hitler would have become Reich Chancellor. It is equally doubtful that a female Hitler would have made the same experiences (such as several years of trench warfare during World War I). In any case, we simply cannot know what would have been if Hitler had been a female.[7]

3.3 Conditional plausibility of the consequent

Conditional plausibility of the counterfactual consequent refers to likelihood that the counterfactual antecedent indeed leads to the counterfactual consequent. To guarantee that the counterfactual antecedent implies the counterfactual consequent, Tetlock and Belkin (1996: 25-27) suggest that the connecting principles should be consistent with well-established theoretical laws. This criterion is, however, contested. Lebow (2000: 583) argues that there are only few generally accepted theories in political science. Often researchers are confronted with several conflicting theories. As a result, competing schools of thought may simply invent counterfactuals of convenience (Weber 1996). Therefore, Tetlock and Belkin (1996: 27) suggest the usage of 'reality constraints'. They argue that counterfactuals must not only fit existing historical ('minimal-rewrite rule') and statistical data, but also provide testable

predictions, which can be empirically evaluated and possibly, together with the proposed counterfactual, rejected.

The example of the redistributive effect of welfare states illustrates the important role of theoretical consistency in the development of sound counterfactuals. As argued by Bergh (2005), currently employed indicators of the degree of welfare state redistribution (the comparison of pre tax/transfer and post tax/transfer income inequality) are flawed because they are based on the improbable counterfactual assumption that real-world welfare states do not affect labour market responses, incentive structures and the private provision of social protection. However, there is an abundance of literature in economics that shows that the welfare state has an effect on economic agents, citing effects such as moral hazard or crowding out. Thus, the counterfactual is not theoretically consistent.[8]

Statistical generalizations provide another 'reality check'. In the words of Tetlock and Belkin (1996: 29), we should use the canons of sound statistical reasoning to constrain our judgements of counterfactuals. Put differently, we should not use counterfactuals that are statistically very unlikely. For example, if we argue that globalization is the cause of financial pressure on mature welfare states, we rely on the counterfactual that in the absence of globalization, mature welfare states would not be under financial pressure. However, considering the existing statistical evidence in macro-comparative welfare state research, such a conclusion would be premature, as demographic change and slower productivity growth among others have contributed to financial pressure on mature welfare states (Pierson 2001: 82).

The conditional likelihood of the counterfactual consequent is also a function of the number of causal steps between antecedent and consequent (Fearon 1996: 66; Lebow 2000: 583). Imagine the following example: The counterfactual antecedent A is supposed to lead to the consequent B. In order to arrive at B, we need to define the connecting principles. Let's say that three causal steps, x, y and z, are needed, each of them having a high likelihood that

they will lead to the assumed outcome (0.80). If all three causal steps need to materialize in order to arrive at the consequent, there is only a 51.2 percent probability that we would indeed observe the counterfactual consequent (0.8x0.8x0.8=0.512). Thus, the more steps are needed in order to get from the antecedent to the consequent, the smaller the probability that the antecedent will indeed lead to the consequent. As a result, we should aim for counterfactual cases, "in which the hypothetical antecedent and consequent are close together in time and are separated by a small number of causal steps" (Fearon 1996: 66).

Finally, we need to acknowledge the interconnectedness of events and consider the effects of second-order counterfactuals (Lebow 2000: 584). As highlighted by Lebow, 'surgical counterfactuals' are not realistic. Changes in the past will very likely require other changes in the past in order to make the counterfactual case possible.[9] For example, a strong Christian democratic party in Great Britain would most likely require changes in the electoral system. This change would, of course, affect many other things and thereby making any causal statement questionable.

Similarly, we should consider the effects of second-order counterfactuals. Second-order counterfactuals can be the result of the long-term effects of enabling conditions or the follow-ups of the counterfactual antecedents and consequents. For example, one might argue that a failed murder attempt against Archduke Franz Ferdinand would have avoided World War I. However, it could well be that the failed attempt might have induced others to emulate Gavrilo Princip, thereby ultimately leading to a successful assassination. Similarly, the long-term effect of economic crisis might not be a lower GDP, but rather a higher GDP because the crisis might force countries to embark on long overdue economic reforms. In both cases, the second-order counterfactuals (setting an example and providing incentives for reform) undercut the first-order counterfactuals.

3.4 Projectability

Projectability is another criterion to impose some reality constraints on counterfactual cases (Tetlock and Belkin 1996: 30-31). Counterfactual cases and their enabling conditions may allow us to formulate other implications that we can test with new data. Following King et al. (1994), we should ask the question: "If my argument is correct, what else should be true?" If we can observe certain implications of the causal argument in the real world, we can be more confident in the validity of our counterfactual analysis. For example, elsewhere we argue that the Danish government would have enacted stricter job security regulations if left and far-left parties had been able to cooperate in the late 1960s and early 1970s, as they did for instance in Sweden (Emmenegger 2010). We then substantiate our counterfactual by showing that there is a strong correlation between the electoral strength of left *and* far-left parties in this period and contemporary levels of job security regulations – but not between the electoral strength of left parties alone and the level of job security regulations – in a sample of all advanced industrialized democracies.

It is important to note that these criteria reflect *best-case* situations. We are neither always able to formulate observable implications on the basis of our analysis nor are we always able to rely on well-established general theoretical laws in order to support our causal statement. However, the checklist helps us to identify the criteria best-case counterfactuals should satisfy in order to qualify as building blocks in causal statements. Deviations from this best-case scenario need not necessarily forebode invalid causal statements. Rather, we should use deviations from these criteria to discuss the validity of the employed counterfactuals and allow for a discussion of the causal statements.

4 Example: Alesina and Glaeser on racial-linguistic fractionalization and social expenditure

Qualitative researchers typically use counterfactuals to evaluate explanations for singular events. In this article, we argue that we can use the very same qualitative criteria to evaluate regularities and proclaimed causal relationships in quantitative macro-comparative welfare state research. We argue that this is a fruitful exercise because these analyses – if they attempt to *explain* relationships – rely on assumptions about counterfactuals that are systematically violated. As a result, additional evaluation is needed in order to increase our confidence in the validity of the results. We hasten to add that we do not suggest to use these counterfactuals to develop 'what if' scenarios or alternative worlds. Rather, we use these criteria for good counterfactuals to "probe causes and contingency of the world we know" (Lebow 2010: 6).

In the following, we demonstrate how researchers can use the qualitative criteria for the evaluation of good counterfactuals to improve our understanding of causal relationships in quantitative macro-comparative welfare state research. We use the well-known study of Alesina and Glaeser (2004, hereinafter AG) on the relationship between racial-linguistic fractionalization and social expenditure to assess their claim that the generous Western European welfare states might not survive in heterogeneous societies.

AG argue that high levels of racial-linguistic fractionalization negatively affected welfare state development in the United States. In contrast, European countries are mostly homogeneous societies. As a result, public social expenditure increased dramatically in the second half of the twentieth century. However, now that Europe has become more diverse as a result of large-scale immigration, they claim that it is unlikely that the generous European welfare states can survive in heterogeneous societies.

From a counterfactual analysis point of view, AG's claim suffers from at least three problems. First, their empirical analysis is likely to violate the assumption of conditional independence. AG are employing a very heterogeneous sample of 52 to 55 countries in their

cross-national analysis, consisting of countries such as Denmark alongside Peru. In their statistical model, they control for GDP per capita and estimate that if heterogeneous Peru would become as homogeneous as Denmark, spending on social welfare would increase from basically zero to 7.1 percent of GDP. AG's counterfactual statement becomes more graspable if we formulate their claim using concrete countries as examples: AG argue that if Peru were racially as homogeneous as Denmark, then Peru would spend about as much on social welfare (% of GDP) as the about five times richer country Australia.

From a counterfactual point of view, the problem is that in such a diverse sample – and in the absence of adequate control variables – the other countries in the sample do not provide adequate information for estimating the counterfactual for Peru: a country exactly like Peru but with low levels of racial fractionalization (antecedent) and medium levels of welfare state spending (consequent).

Obviously, with such a heterogeneous sample and so few control variables, AG's regression model violates the assumption of conditional independence. In a previous analysis, they used more control variables and arrived at rather similar conclusions (Alesina et al. 2001)[10], but even these models are very likely to violate the conditional independence assumption. Table 3 compares the control variables used in the models of AG, Alesina et al. (2001) and Huber and Stephens (2001). Although this is not a formal test for omitted variable bias, the comparison clearly illustrates that AG are likely to have missed some crucial control variables. Thus, in the absence of adequate controls for constitutional structure, unemployment rates or government composition, causal inferences about the effect of racial fractionalization on social expenditure are unlikely to be valid.[11]

Table 3: Control variables and conditional independence (excl. regional dummies and interaction effects)

Alesina and Glaeser (2004: 142, 144)	*Alesina et al. (2001: 231)*	*Huber and Stephens (2001: 68-69)*
GDP per capita	GDP per capita	**GDP per capita**
	Population aged 15-64 (%)	**Population aged 65+ (%)**
	Majoritarian representation (MR)	**Institutional veto points (incl. MR)**
		Left cabinet share
		Christian democratic cabinet share
		Female labour force participation
		Voter turnout (% adult population)
		Strikes (working days lost)
		Authoritarian legacy
		Consumer price index
		Unemployment rate
		Military spending (% GDP)
		Outward FDI (% GDP)
		Imports and exports (% GDP)

Notes: Significant coefficients, as observed in their analyses, are marked bold.

Second, AG do not acknowledge the possibility of second-order counterfactuals. For instance, in another paper, Alesina et al. (2003) show that ethno-linguistic fractionalization has a negative effect on long-term economic growth and some indicators of quality of government. Thus, heterogeneity seems to affect not only social expenditure but also other variables. This phenomenon is referred to as second-order counterfactuals. In this case, lower levels of heterogeneity would lead not only to higher levels of social expenditure, but also to higher levels of economic growth (and probably many more things). Since AG operationalize their dependent variable as social expenditure as a percentage of GDP, this second-order counterfactual has an offsetting effect on the relationship between heterogeneity and welfare state spending.

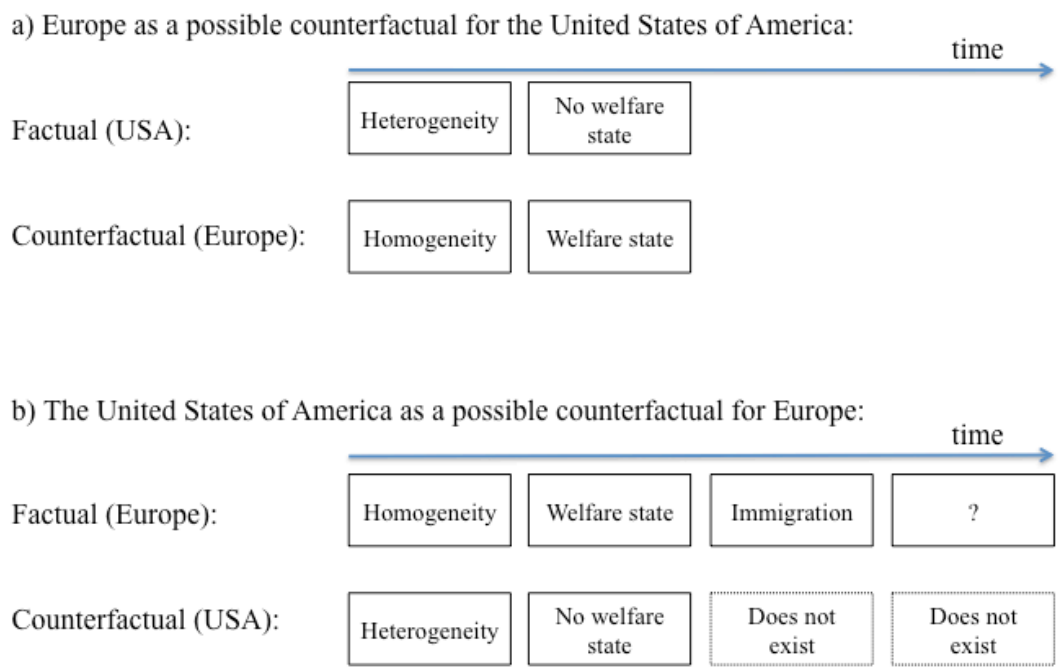
To give a numerical example, Alesina et al. (2003: 166) claim that going from complete homogeneity (score of 0) to maximum heterogeneity (score of 1) decreases annual economic growth by 1.9 percentage points. At the same time, AG (2004: 142) claim that going from Peruvian levels of heterogeneity (score of about 0.70) to Danish levels (score of 0.02) would increase social expenditure as a percentage of GDP by 7.1 percentage points. However, when

we take the higher levels of economic growth in the thirty years prior to AG's (2004) observation in 1998 into account, the positive effect of going from a heterogeneity score of 0.70 to 0.02 on social expenditure as a percentage of GDP decreases from 7.1 percentage points to 4.8 percentage points. This 32 percent reduction of the observed effect is simply the result of taking into account second-order counterfactuals.[12]

Third, AG ignore the longitudinal implications of their argument and the important role of the sequence of events.[13] In the United States, racial-linguistic fractionalization was historically prior to the development of the welfare state, while contemporary migration movements to Western Europe lead to higher levels of racial-linguistic fractionalization in countries with mature welfare states. Thus, the counterfactual case to Western European countries is not the historical development of the welfare state in the United States. Rather, a mature welfare state and constant or even decreasing levels of racial-linguistic fractionalization would characterize the appropriate counterfactual. The United States does not satisfy any of these two criteria. Thus, we are confronted with problems of logical and historical consistency.

Figure 1 displays the important difference. For the development of the US welfare state, we might be able to use European experiences to inform us about the counterfactual to the US case (upper half of Figure 1). While a heterogeneous population characterizes the US, most European countries have relatively homogeneous populations. Thus, following AG, European welfare states were able to take off, while the US welfare state remained lean. However, turning to the question of what will happen to European welfare states in the face of mass immigration, we cannot use the US case to inform us about the counterfactual to the European case. The US did not experience decreasing levels of heterogeneity *in the presence of generous welfare states.* Thus, by relying on the US case to inform us about a counterfactual Europe, AG ignore the sequence of events and ultimately rely on a 'miracle' counterfactual.

Figure 1: Possible counterfactuals for the US and Europe



This conclusion has important implications for research on the nexus between racial-linguistic fractionalization and welfare state generosity. Following AG, researchers have plotted indicators of diversity against indicators of overall welfare state generosity (e.g. total public social expenditure). However, this choice of indicators has clearly been affected by the idea that racial-linguistic fractionalization has a negative effect on all sorts of social expenditure. This makes a lot of sense when racial-linguistic fractionalization is logically prior to the development of the welfare state. However, in the case of countries with mature welfare states and increasing levels of racial-linguistic fractionalization due to immigration, this relationship cannot be expected.[14]

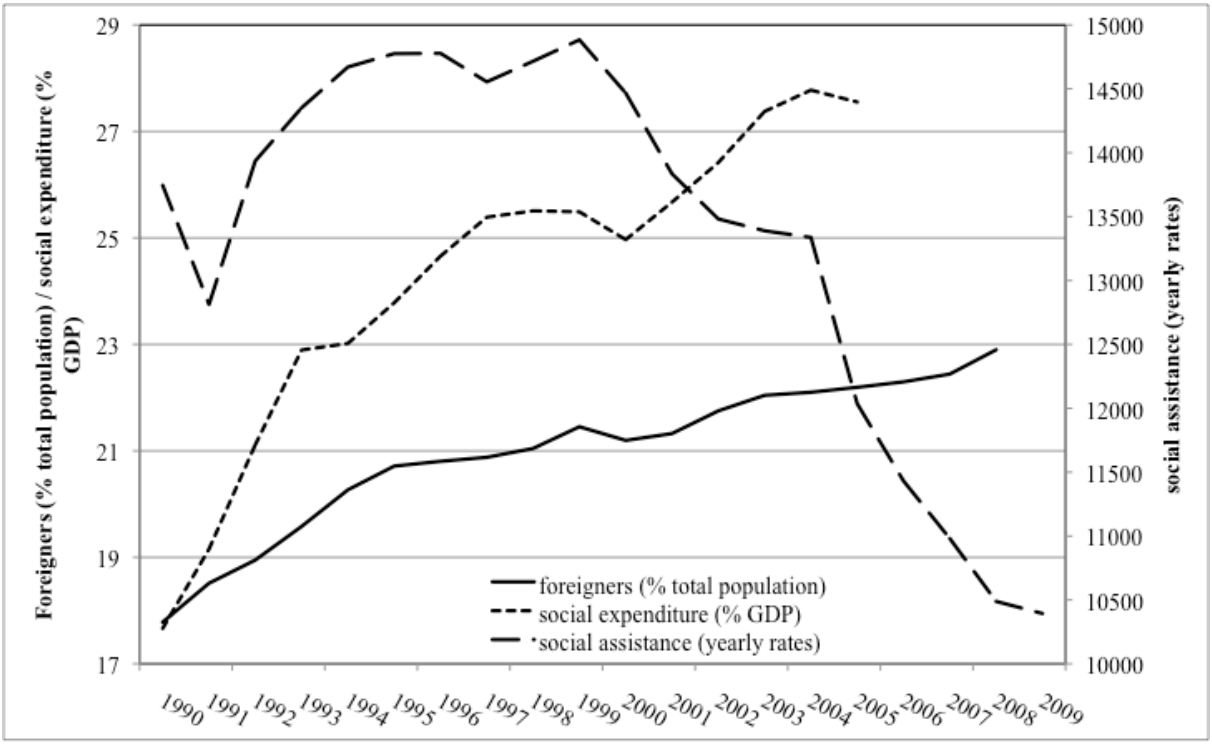
Mature welfare states spend most of their money on old-age pensions and health care. These programmes enjoy high levels of public support because a considerable share of the population benefits or expects to benefit from these schemes. Moreover, in many countries, these schemes pay benefits as a function of years of contribution and prior income, thereby disproportionately benefiting the native population. Consequently, if increasing levels of

racial-linguistic fractionalization really lead to welfare state reform, we should expect to observe the retrenchment of social security schemes that disproportionately benefit immigrants, and not of overall public social expenditure.[15]

In fact, looking only at total spending might be seriously misleading. Figure 2 displays public and private mandatory social spending as a percentage of GDP, the share of foreigners as a percentage of total population and social assistance yearly rates standardized for the development of wages for Switzerland in the period 1990 to 2009. As shown by Morissens and Sainsbury (2005: 650), migrant households are more dependent on social assistance than citizen households. We choose Switzerland because of its high level of racial-linguistic fractionalization and continuously high levels of immigration. Figure 2 shows that in the period 1990 to 2009 both social expenditure and the foreign population have increased in Switzerland. Thus, one might conclude that there is in fact a positive relationship between heterogeneity and social expenditure. However, looking at social assistance only, one can easily see that such a conclusion would be premature, as in the period 1990 to 2009 yearly social assistance rates have decreased by 24 percent. Since immigrants are generally overrepresented among social assistance recipients, they are likely to have suffered disproportionately from decreasing social assistance levels.

Thus, by simply evaluating the counterfactuals on which AG base their claims, we can show that their causal argument is flawed. However, AG are no special case. Quite the contrary, deficient counterfactual theorizing is legion in macro-comparative welfare state research. For example, Bergh (2005) has demonstrated that the commonly used indicator of the redistributive effect of welfare states suffers from a counterfactual problem. If we want to know the redistributive effect of a given welfare state, we need to know the income distribution after taxes and transfers and the income distribution in the absence of this welfare state. The latter is normally measured using the income distribution before taxes and transfers.

Figure 2: Development of social expenditure, yearly social assistance rates and the share of foreigners in Switzerland (1990 to 2009)



Sources: Data taken from OECD social expenditure dataset, URL: www.oecd.org (accessed July 6, 2010), the Swiss Federal Statistical Office, URL: www.statistik.admin.ch (accessed July 6, 2010) and Nelson's (2009) social assistance and minimum income protection interim dataset, URL: http://www2.sofi.su.se/~kne/ (accessed July 8, 2010).

This approach relies on the counterfactual assumption that a given welfare state does not affect the income distribution before taxes and transfers (otherwise the income distribution before taxes and transfers would not be a good indicator of the income distribution in the absence of a given welfare state). This assumption, however, is very problematic. Much of the contemporary debate on the upsides and downsides of welfare states centres on the question of negative incentives and moral hazard. Thus, there is very good reason to believe that welfare states affect the decisions of economic agents. As a consequence, the counterfactual is not valid (lack of theoretical consistency). The fact that different institutional setups may have different effects on labour supply responses, market incentives and the private provision of social welfare, thereby introducing a systematic cross-national bias, make matters even worse. Although Bergh (2005: 355) argues that these considerations are unlikely to dramatically

change the rank order of countries with regard to the distributional of effects of their welfare states, we should be nevertheless careful when interpreting these indicators and consider possible systematic biases. Again, this is not to argue that we should abandon these indicators. To our knowledge, these indicators are the best available. However, we should keep in mind that they may suffer from *systematic* bias.

While AG refrain from the explicit usage of counterfactual analysis, Huber and Stephens (2001) explicitly use counterfactual theorizing to explore possible alternative explanations. In the discussion of their methodological approach, they highlight the need for 'realistic' counterfactuals. To illustrate their point, they argue that it does not make any sense to ask whether the Australian welfare state would have been any different if Christian democracy and not the secular right had been in power in the period 1950 to 1972. In contrast, they argue that it makes sense to ask whether the Norwegian welfare state would have been any different in 1980 had bourgeois coalitions been politically predominant up to that point (Huber and Stephens 2001: 35).[16]

It goes without saying that Huber and Stephens (2001) are right. The Norwegian welfare state would look completely differently had bourgeois coalitions been predominant up to 1980. The problem is, however, that we have absolutely no clue how the Norwegian welfare state would look like in this case, and the real Norwegian welfare state in 1980 will not be any help in figuring this out.

From a counterfactual point of view, the problems are the enabling and the second-order counterfactuals. If mostly bourgeois coalitions had been in power up to 1980, not only the welfare state would look differently. It is very likely that after decades of rightwing instead of leftwing political dominance we would also observe numerous other differences in areas such as political institutions, industrial relations or education policy (second-order counterfactuals). Moreover, we further need to address the question of which factors we need to adapt in our counterfactual case in order to make bourgeois political dominance logically possible

(enabling conditions)? How is bourgeois political dominance possible in a country with Scandinavian-style industrial relations? Do we need to assume different electoral institution and/or a different socio-economic composition of the voting population? And how can we assume that after 1980, Norway would suddenly follow the factual trajectory?

One might agree that many things would have changed had bourgeois coalitions been predominant in Norway up to 1980. However, one might still insist that the welfare state would be less generous. Put differently, in this counterfactual case, Norway might look completely different, but welfare state generosity would be lower in any case, thereby supporting the causal statement. Yet, this conclusion is not necessarily true, either. If we accept the possibility of equifinality, i.e. if we believe that there is more than one path to generous welfare states, a bourgeois dominated Norway might still have a generous welfare state.

Social democratic-led governments ruled Norway for the most part of the period 1945 to 1980 (with three exceptions, lasting together about eight years). In contrast, the Netherlands experienced in the same period about 20 years of centre-right-led governments (mostly of a Christian democratic kind) and 15 years of social democratic-led governments. Still, in 1980, the Dutch welfare state was more generous in terms of social spending as a percentage of GDP (24.2 percent as compared to 16.9 percent in Norway; Armingeon et al. 2008). Thus, we cannot be sure whether our imagined 'alternative' Norway, one dominated by bourgeois coalition governments up to 1980, would spend less on the welfare state. In a nutshell: Although Huber and Stephens' (2001) Norwegian example is somewhat more realistic than their Christian democracy in Australia example, it does not help us identifying causal conditions.

Again, the point of this discussion is not to argue that there is no relationship between partisan politics and welfare state generosity. This discussion is simply to illustrate that part

of Huber and Stephen's (2001) argumentation is relying on evidence that does not allow for this conclusion.

5 Conclusions

All causal statements in studies using observational and historical data, thus all research in macro-comparative welfare state research, rely on counterfactuals (Fearon 1991; King et al. 1994; Levy 2008; Lebow 2010). For example, in macro-comparative welfare state researchers often argue that electorally stronger left parties in a given country, say Great Britain, would lead to higher levels of public social spending. Of course, we cannot know how much the British state would spend if left parties were electorally stronger than they actually are, as this counterfactual case *does not exist*. However, by claiming a causal relationship between electoral strength of left parties and social spending, we implicitly argue that social spending would be higher.

In quantitative research, such statements are typically based on the Neyman-Rubin-Holland theory of causation, which compares something that did occur (the factual) with something that did not occur (the counterfactual). Since we can never observe the counterfactual, we need to rely on assumptions about counterfactuals in order to identify causal effects. Most importantly, we need to specify our research design in a way that allows us to observe the causal effect independent of confounders. This can be best done in an experiment in which we introduce a treatment, while eliminating rival explanations by randomizing the assignment into treatment and control groups.

In quantitative macro-comparative welfare state research, we are almost never in a position to postulate that our research design justifies the assumption of independence of confounders. Rather than relying on experimental data, we use observational data. Rather than randomly assigning cases to treatment and control groups, we are facing a situation in which history has made the case selection for us. Thus, in macro-comparative welfare state research,

the assumptions of the Neyman-Rubin-Holland theory of causation are systematically violated and most causal statements are thus based on shaky foundations.

In this article, we do not propose to abandon quantitative macro-comparative research. In the social sciences, we have to work with the methods that are available and quantitative approaches are an important part of our tool-kit. However, we suggest being more careful in causal statements. Rather than simply assuming that the counterfactuals, on which our causal statements are based, are sound, we should scrutinize them and ask ourselves whether our analysis relies on reasonable counterfactuals or 'miracle' counterfactuals.

To this end, we have proposed several criteria that can be used to evaluate counterfactuals. We have demonstrated using the hallmark study by Alesina and Glaeser (2004) that simple reasoning based on the criteria for good counterfactuals can reveal fundamental problems in causal statements. However, rather than simply highlighting problems in the study by Alesina and Glaeser (2004), we have used the criteria for good counterfactuals to formulate *more realistic* hypotheses. For instance, we have demonstrated that rather than looking at total social expenditure, researchers interested in the relationship between racial-linguistic fractionalization and welfare state generosity in Europe should look at programmes that disproportionately benefit migrants and criteria that regulate the access to welfare state benefits. Thus, rather than using the criteria for good counterfactuals to refute causal statements based on quantitative macro-comparative research, we suggest using these criteria to *improve* macro-comparative welfare state research – a common goal for both qualitative and quantitative researchers.

References

Achen, Christopher H. (2002). 'Toward a New Political Methodology: Microfoundations and Art', *Annual Review of Political Science* 5: 423-450.

Alesina, Alberto, Devleeschauwer, Arnaud, Easterly, William, Kurlat, Sergio and Wacziarg, Romain (2003). 'Fractionalization', *Journal of Economic Growth* 8(2): 155-194.

Alesina, Alberto and Glaeser, Edward L. (2004). *Fighting Poverty in the Us and Europe: A World of Difference*. New York: Oxford University Press.

Alesina, Alberto, Glaeser, Edward L. and Sacerdote, Bruce (2001). 'Why Doesn't the United States Have a European-Style Welfare State?', *Brookings Papers on Economic Activity* 2001(2): 187-254.

Armingeon, Klaus, Gerber, Marlène, Leimgruber, Philipp and Beyeler, Michelle (2008). *Comparative Political Data Set 1960-2006*. Institute of Political Science, University of Bern.

Bergh, Andreas (2005). 'On the Counterfactual Problem of Welfare State Research: How Can We Measure Redistribution?', *European Sociological Review* 21(4): 345-357.

Brady, Henry E. (2004). 'Doing Good and Doing Better: How Far Does the Quantitative Template Get Us?'. In Brady, Henry E. and Collier, David (eds.). *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham/Boulder/New York/Toronto/Oxford: Rowman & Littlefield Publishers, Inc., pp. 53-67.

Brady, Henry E. (2008). 'Causation and Explanation in Social Science'. In Box-Steffensmeier, Janet M., Brady, Henry E. and Collier, David (eds.). *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, pp. 217-270.

Brady, Henry E. and Collier, David (eds.) (2004). *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham/Boulder/New York/Toronto/Oxford: Rowman & Littlefield Publishers, Inc.

Clarke, Kevin A. (2005). 'The Phantom Menace: Omitted Variable Bias in Econometric Research', *Conflict Management and Peace Science* 22(4): 341-352.

Collier, David, Seawright, Jason and Munck, Gerardo L. (2004). 'The Quest for Standards: King, Keohane, and Verba's Designing Social Inquiry'. In Brady, Henry E. and Collier, David (eds.). *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham/Boulder/New York/Toronto/Oxford: Rowman & Littlefield Publishers, Inc., pp. 21-50.

Czada, Roland (1998). 'Vereinigungskrise Und Standortdebatte. Der Beitrag Der Wiedervereinigung Zur Krise Des Westdeutschen Modells', *Leviathan* 26(1): 24-59.

Dunning, Thad (2007). 'Improving Causal Inference: Strengths and Limitations of Natural Experiments', *Political Research Quarterly* 61(2): 282-293.

Ebbinghaus, Bernhard (2005). 'When Less Is More: Selection Problems in Large-N and Small-N Cross-National Comparisons', *International Sociology* 20(2): 133-152.

Elster, Jon (1978). *Logic and Society: Contradictions and Possible Worlds*. Chichester/New York/Brisbane/Toronto: John Wiley & Sons.

Emmenegger, Patrick (2010). 'The Long Road to Flexicurity: The Development of Job Security Regulations in Denmark and Sweden', *Scandinavian Political Studies* 33(3): 271-294.

Emmenegger, Patrick and Careja, Romana (2010). *From Dilemma to Dualization: Social and Migration Policies in the 'Reluctant Countries of Immigration'*, Paper presented at the annual conference of the Council for European Studies, Montréal, April 2010.

Fearon, James D. (1991). 'Counterfactuals and Hypotheses Testing in Political Science', *World Politics* 43(2): 169-195.

Fearon, James D. (1996). 'Causes and Counterfactuals in Social Science: Exploring an Analogy between Cellar Automata and Historical Processes'. In Tetlock, Philip E. and

Belkin, Aaron (eds.). *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton NJ: Princeton University Press, pp. 39-67.
Fogel, Robert (1964). *Railroads and American Economic Growth:Essays in Econometric History*. Baltimore MD: John Hopkins University Press.
Freedman, David A. (2010). 'Statistical Models and Shoe Leather'. In Collier, David, Sekhon, Jasjeet S. and Stark, Philip B. (eds.). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge: Cambridge University Press, pp. 45-62.
Goertz, Gary and Levy, Jack S. (2007). 'Causal Explanation, Necessary Conditions, and Case Studies'. In Goertz, Gary and Levy, Jack S. (eds.). *Explaining War and Peace: Case Studies and Necessary Condition Counterfactuals*. London/New York: Routledge, pp. 9-45.
Goldthorpe, John (2000). 'Current Issues in Comparative Macrosociology'. In Goldthorpe, John (ed.) *On Sociology: Numbers, Narratives, and the Integration of Research and Theory*. Oxford: Oxford University Press, pp. 45-64.
Holland, Paul W. (1986). 'Statistics and Causal Inference', *Journal of the American Statistical Association* 81(396): 945-960.
Huber, Evelyne and Stephens, John D. (2001). *Development and Crisis of the Welfare State. Parties and Policies in Global Markets*. Chicago IL/London: University of Chicago Press.
King, Gary, Keohane, Robert O. and Verba, Sidney (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton NJ: Princeton University Press.
King, Gary and Zeng, Langche (2006). 'The Danger of Extreme Counterfactuals', *Political Analysis* 14(2): 131-159.
King, Gary and Zeng, Langche (2007). 'When Can History Be Our Guide? The Pitfalls of Counterfactual Inference', *International Studies Quarterly* 51(1): 183-210.
Lebow, Richard Ned (2000). 'What's So Different About a Counterfactual', *World Politics* 52(4): 550-585.
Lebow, Richard Ned (2010). *Forbidden Fruit: Counterfactuals and International Relations*. Princeton NJ/Oxford: Princeton University Press.
Levy, Jack S. (2008). 'Counterfactuals and Case Studies'. In Box-Steffensmeier, Janet M., Brady, Henry E. and Collier, David (eds.). *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, pp. 627-644.
Lieberson, Stanley (1985). *Making it Count: The Improvement of Social Research and Theory.* Berkeley: University of California Press.
Lieberson, Stanley and Hansen, Lynn K. (1974). 'National Development, Mother Tongue Diversity, and the Comparative Study of Nations', *American Sociological Review* 39(4): 523-541.
Morissens, Ann and Sainsbury, Diane (2005). 'Migrants' Social Rights, Ethnicity and Welfare Regimes', *Journal of Social Policy* 34(4): 637-660.
Persson, Torsten and Tabellini, Guido (2003). *The Economic Effects of Constitutions.* Cambridge: MIT Press.
Pierson, Paul (2001). 'Post-Industrial Pressures on the Mature Welfare States'. In Pierson, Paul (ed.) *The New Politics of the Welfare State*. New York: Oxford University Press, pp. 80-104.
Przeworski, Adam (2007). 'Adam Przeworski: Capitalism, Democracy, and Science'. In Munck, Gerardo L. and Snyder, Richard (eds.) *Passion, Craft, and Method in Comparative Politics.* Baltimore: John Hopkins University Press, pp. 456-503.
Ragin, Charles C. (2000). *Fuzzy-Set Social Science.* Chicago IL: University of Chicago Press.
Rubin, Donald B. (1978). 'Bayesian Inference for Causal Effects: The Role of Randomization', *Annals of Statistics* 6(1): 34-58.

Shalev, Michael (2007). 'Limits and Alternatives to Multiple Regression in Comparative Research', *Comparative Social Research* 24: 261-308.
Taylor-Gooby, Peter (2005). 'Is the Future American? Or, Can Left Politics Preserve European Welfare States from Erosion through Growing 'Racial' Diversity?', *Journal of Social Policy* 34(4): 661-672.
Tetlock, Philip E. and Belkin, Aaron (1996). 'Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives'. In Tetlock, Philip E. and Belkin, Aaron (eds.). *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton NJ: Princeton University Press, pp. 3-38.
Tetlock, Philip E. and Lebow, Richard Ned (2001). 'Poking Counterfactual Holes in Covering Laws: Cognitive Styles and Historical Reasoning', *American Political Science Review* 95(4): 829-843.
Tilly, Charles (1975). 'Reflections on the History of European-State Making'. In Tilly, Charles (ed.) *The Formation of National States in Western Europe*. Princeton NJ: Princeton University Press, pp. 3-83.
Weber, Max (1949). *The Methodology of the Social Sciences*. Glencoe IL: Free Press.
Weber, Steven (1996). 'Counterfactuals, Past and Future'. In Tetlock, Philip E. and Belkin, Aaron (eds.). *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton NJ: Princeton University Press, pp. 268-288.

Endnotes:

[1] Arguably, this limitation applies to all observational studies. However, the problem is accentuated in macro-comparative welfare state research because of the small number of cases and the strong empirical evidence for the existence of selectivity.

[2] The usage of pooled time-series cross-section data does not solve these problems. Apart from leading to new problems such as heteroskedasticity, spatial and serial autocorrelation, this method relies on units of analysis (e.g. Germany in 1980, Germany in 1981, Germany in 1982 etc.) that are obviously not independent from each other and that are not randomly assigned to treatment and control groups.

[3] Unfortunately, natural experiments are rarely an option in macro-comparative welfare state research. Natural experiments rely on observational data. However, unlike conventional studies relying on observational data, natural experiments use data that has been assigned to treatment and control groups 'as if' random (Dunning 2007: 283). This is a very rare feature of naturally occurring data.

[4] It is somewhat ironic that some critics of qualitative comparative analysis (QCA) and Mill's methods like to remind proponents of these methods that Mill himself considered his methods unfit for the study of social phenomena (e.g. Goldthorpe 2000: 49), when Rubin (1978) and Holland (1986) are critical of the usage of their counterfactual theory of causation in case of non-randomized data.

[5] Propensity score matching is not going to solve this problem (see Lieberson [1985: 38] for an early discussion of matching's inability to deal with these issues). In fact, if applied to small samples propensity score matching can lead to very odd 'couples'. For instance, in the well-known study by Persson and Tabellini (2003: 144-147), the UK is matched with Romania, the USA with Venezuela, Australia with South Africa and Zimbabwe with Fiji.

[6] Even though we focus on quantitative research in this article, these criteria can equally be used to evaluate counterfactuals in qualitative research and the analysis of singular events.

[7] There is a clear parallel between the criteria for good counterfactuals and the Neyman-Rubin-Holland theory of causation. The Neyman-Rubin-Holland model considers only treatments that can be manipulated possible causes. Otherwise, the independence assumption might be violated, for instance as a result of unobservable confounders. Thus, gender is no possible cause because researchers cannot simply change it and because gender is likely to be related to other variables that cannot be observed and eliminated through random assignment.

[8] This means that we cannot know for sure whether the observed redistributive effect is due to actual redistribution or due to the effects of welfare states on individual economic behaviour.

[9] This criterion overlaps with the aforementioned criterion of 'logical consistency'. However, here, the focus is more on the (unintended) effects of the 'enabling conditions' on the counterfactual consequent rather than on the 'enabling conditions' that are needed to make the counterfactual antecedent logically possible.

[10] The interpretation of the results of Table 9 in Alesina et al. (2001) is complicated by the fact that it is unclear whether the number reported in parentheses refer to t-values or standard errors. Although it is stated in the note to the table that the numbers refer to t-values, they are treated as if they would refer to standard errors (see asterisks and text on page 230).

[11] In a replication, Taylor-Gooby (2005: 669) shows that restricting the sample to advanced Western democracies and Japan leads to a weak bivariate relationship between racial fractionalization and public social expenditure, which is entirely dependent on the inclusion of the US case. If the United States is dropped from the sample, the relationship disappears.

[12] Own calculations using data provided by the Penn World Table 6.3. URL: http://pwt.econ.upenn.edu/ (accessed July 8, 2010).

[13] Lieberson and Hansen (1974) already extensively discussed the problems associated with using cross-national data to examine causal models of change. Coincidentally, they use the

example of the relationship between mother tongue diversity and national development to illustrate their points.

[14] Changing the level of analysis is no option. For instance, AG (2004: 146-148) show that in the United States there is a negative relationship between the share of blacks (% total population) and the generosity of some welfare programmes at the state level. But unless we have good reason to assume that there is no nation effect, these results have no meaningful implications for relationships on the nation-state level (see Lieberson 1985: 110-115). For instance, data for Switzerland – another federal state with high levels of racial-linguistic fractionalization – show that there is in fact a positive correlation between the share of foreigners and the gross cantonal income per capita (r_p=0.42).

[15] As Emmenegger and Careja (2010) demonstrate, recent cutbacks in social security programmes have indeed disproportionately affected immigrants.

[16] However, note that Huber and Stephen's (2001) quantitative model implicitly considers the possibility of Christian democracy in Australia, even though they rule it out as an unrealistic counterfactual in the qualitative part of their analysis.