WHAT YOU SEE AND WHAT YOU GET –

PITFALLS AND PRINCIPLES OF NESTED ANALYSIS IN COMPARATIVE RESEARCH

Typescript including all changes made during the review and editing process

The final, definitive version is available at http://online.sagepub.com/

Ingo Rohlfing
rohlfing@wiso.uni-koeln.de
Assistant Professor for Comparative Social Research
Universität Köln
Abstract
In a recent contribution to this journal, Munck and Snyder found that many studies suffer from a deficient application of qualitative and quantitative methods. They argue that the combination of small-n and large-n analysis represents a viable method for promoting the production of knowledge. Recently, Evan Lieberman proposed nested analysis as a rigorous approach for comparative research that builds on the complementary strengths of quantitative and qualitative analysis. In this paper, I examine the methodological potential of nested inference to advance comparative political analysis. I argue that the specific methodological problems of nested designs have not been fully appreciated. It will be shown that, under certain circumstances, nothing is gained from a nested analysis. On the contrary, one might lose more than one gains compared to single-method designs. I suggest specific methodological principles that take these problems into account in order to make nested analysis fruitful for comparative studies.
1. Introduction

In a recent contribution to this journal, Munck and Snyder (2007a) evaluated the state of comparative political research. One of their major findings is that many studies suffer from a deficient application of qualitative and quantitative methods. They argue that the combination of small-n and large-n research represents a viable method for promoting the production of knowledge (Munck & Snyder, 2007b). Their suggestion is echoed by the increasing trend toward the use of mixed-methods design in comparative research (Lieberman, 2005). In particular, regression analysis and case studies are frequently combined (e.g. Lieberman, 2003; Simmons, 1994).

Recently, Evan Lieberman (2005) proposed nested analysis as a rigorous method that brings together the complementary strengths of case study research and regression analysis.\(^1\) These two instruments serve to cross-validate the accuracy of the ontological modeling decisions that are made in a nested analysis. In a nutshell, the regression analysis determines the explanatory power of a model, the causal effects of the independent variables and their statistical significance. The residuals of the cases provide the basis for case selection. The inferential goal of the qualitative analysis is to discern whether the significant independent variables are linked to the dependent variable through causal mechanisms. A nested analysis of this kind is argued to yield synergistic value because of the opportunity to make integrated causal inference that is infeasible in single-method designs.

I do not deny that nested analysis is potentially superior to single-method research. I argue, however, that the specific methodological problems of nested designs have not been fully appreciated.\(^2\) It will be shown that ontological misspecification has methodological implications that are difficult, if not impossible, to detect in a nested analysis in its current form. The peculiar methodological effect of an ontological mistake is that it “travels”
through the design and undermines causal inference in the quantitative and the qualitative part, rendering its identification and elimination arduous. In this instance, nothing is gained from a nested analysis. On the contrary, one might lose more than one gains because a single-method analysis does not suffer from traveling mistakes. Thus, nested analysis in its current form cannot live up to its promises. The consequence is not to abandon nested analysis altogether because it has the potential to be superior to single-method designs. It is, however, necessary to follow specific guidelines that I introduce in this paper and go beyond what has been written about the conduct of nested inference so far.

In section two, I outline the basic elements of a nested analysis in order to familiarize the reader with this approach. In section three, I distinguish different types of ontological misspecification. The effects of the different variants of misspecification for causal inference are described in sections four and five. In addition, these sections involve a discussion of graphical, quantitative and qualitative tools for uncovering ontological misspecification. The sixth section concludes.

2. Nested analysis in a nutshell

A nested analysis always starts with a preliminary regression analysis (Lieberman, 2005, p. 438). This regression is preliminary inasmuch as the subsequent small-n analysis might point to a misspecification of the regression model, an incorrectly delineated population, or inappropriate measurement of variables (p. 449). It is suggested that regression models of competing theoretical approaches should be tested in order to separate the better models from the poorer ones. The question with which one should approach the regression results is whether the results are “robust/satisfactory” (p. 440). “Satisfactory” refers to the goodness of fit of a regression model, the congruence of the coefficients with the theoretical expectation, and the significance of the independent variables (p. 439). A satisfactory
model has high model-fit and statistically significant variables with signs that are in accord with the underlying theory. “Robustness” denotes the degree to which the obtained results are sensitive to slight changes in the specification.

One should plot the predicted scores of the dependent variable against the actual ones in order to assess the quality of the model. The cases are randomly distributed around the bisecting line when the model is well-specified. If there is reason to believe that the cases are systematically distributed, one should regard the estimated model with suspicion (p. 439). In addition to a visual inspection of the distribution, it is recommended to assess the quality of the model on the basis of the location of theoretically important cases. Assuming that one is examining revolutions, the French revolution should lie close to the bisecting line (p. 440).

Following the large-n analysis (LNA), one proceeds with a small-n analysis (SNA). The LNA is located at the cross-case level, while an SNA takes place on the within-case level (p. 440). Lieberman distinguishes between model-testing small-n analysis (Mt-SNA) and model-building small-n analysis (Mb-SNA). One should conduct an Mt-SNA when the model appears satisfactory/robust and an Mb-SNA in all other cases. Two questions should be asked in an Mt-SNA: is there a causal link between the hypothesized cause and effect, and does the cause precede the effect (p. 443)? In order to answer these questions, Lieberman recommends the selection of cases that have small residuals and cover a maximum of variation on the independent variables in order to test the model in various contexts (pp. 442-444). After the cases have been selected accordingly, one should examine the best-fitting model of the preliminary LNA (p. 445). Lieberman recommends that the focus of the within-case analysis should lie on those variables that were statistically significant in the LNA. Mt-SNA, however, is not confined to models that have been estimated in the LNA. It
can also be applied when a certain theory cannot be tested quantitatively due to a lack of appropriate data (p. 443).

When the Mt-SNA corroborates the results of the preliminary LNA, one can be confident that the underlying theory has much explanatory power and the nested analysis can be terminated. One should refine the model according to the insights gained from the case study if the Mt-SNA yields evidence that contradicts the results of the LNA. When the modified model can be tested against data that is different from that employed in the preliminary LNA, one should conduct a model-testing LNA (Mt-LNA) so as to gather cross-case evidence about the validity of the new model. Where no such data is available, the nested analysis ceases after the model has been refined.

An Mb-SNA is performed when none of the models tested in the preliminary LNA proves to be satisfactory/robust. Case selection in Mb-SNA is different in two respects. First, one should choose at least one case that deviates from the best-fitting model. The rationale for this is that it makes little sense to limit case selection to cases that are well-predicted by a badly performing regression model. Second, in contrast to Mt-SNA, the cases’ scores on the independent variables are of minor importance. The comparatively low model-fit indicates that the independent variables that were fed into the preliminary LNA are insufficient for explaining much of variation of the dependent variable. Therefore, the scores on the independent variables are irrelevant in case selection for Mb-SNA (p. 445).

The new model should be tested by an model-testing LNA (Mt-LNA) when the Mb-SNA suggests a modification of the original model. Lieberman recommends to estimate the new model, based on the full set of cases and on this set minus the case(s) used for the Mb-SNA. This makes it is possible to assess the influence of the examined cases on the basis of the regression results (p. 449, fn. 30). When the results of the Mt-LNA are satisfacto-
ry/robust, one can be confident in the validity of the model derived from Mb-SNA. Correspondingly, the nested analysis comes to an end without any satisfactory/robust model if the cross-case results fail to confirm the model (p. 450).

The nested analysis should be terminated when an Mt-LNA is infeasible due to lack of data or when the case is considered exceptional, meaning that it is useless for the development of a cross-case model (p. 449). Besides of the different roles of Mt-SNA and Mb-SNA in model-building, both types of within-case analysis have in common that they serve to understand the outcome in a particular case and the variables and processes that caused the outcome (p. 436).

3. Types of variables and ontological misspecification

The discussion of the specific problems of nested inference requires to distinguish between different types of variables and variants of ontological model misspecification. Ontological specification of a causal model involves two steps each of which is a potential source of error. First, it is important to determine the variables that are to be included in the model. Since the goal is to develop a general causal model, it should only cover systematic variables, given the theory and research question at hand. The opposite of systematic variables are non-systematic variables. Non-systematic variables are important in explaining the outcome in a particular case, but they are irrelevant to the development of a cross-case model. A model is overfitted if it contains a non-systematic variable, and underfitted when it excludes a systematic variable. Underfitting and overfitting are not mutually exclusive phenomena. It is possible that the model includes an irrelevant variable while simultaneously omitting a relevant one. In the practice of social research, systematic and non-systematic variables cannot always be easily delineated. As will be demonstrated below, the
uncertainty about the correct status of a variable is a major source of ontological misspecification.

The distinction between systematic and non-systematic variables and underfitting and overfitting is not new (e.g. King, Keohane & Verba, 1994, section 3.1). As will be shown below, however, the implications of these distinctions for nested inference have not been fully appreciated in the literature. In order to understand the consequences, it is important to explain how systematic and non-systematic variables influence the score on the dependent variable of a particular case. The score on the dependent variable is determined by the combined influence of systematic and non-systematic variables. To give an example, the vote share a party receives on polling day is the result of systematic variables, e.g., its party platform, and non-systematic variables, e.g., the weather. Each of these variables is, in turn, affected by a combination of systematic and non-systematic variables. If one could run a large number of hypothetical replications, one would obtain the mean causal effect of the systematic variables because the influences of the non-systematic variables would cancel each other out. In most instances, multiple replications of the same event are infeasible. The alternative is to run a regression on a group of comparable cases. If all of the statistical assumptions are fulfilled, there is no deteriorating impact of non-systematic variables on the mean causal effect (cf. King, Keohane & Verba, 1994, section 3.3). The score on the dependent variable of a specific case, however, remains the result of the influence of both systematic and non-systematic variables. This point will be of particular importance when it comes to the discussion of model specification and case selection. The second ontological task of model-building requires the specification of the causal effect of each independent variable (the variable’s parameter in regression models). It is important to consider, for ex-
ample, whether the effects of the independent variables are independent of each other or whether interaction effects are involved.

In the following sections, I will discuss how all of the elements of causal inference in a nested analysis are affected by ontological misspecification. Modifying Lieberman’s framework somewhat, I distinguish between regression based nested analysis and case study based nested analysis as two variants of a nested design. Regression based nested designs begin with the quantitative analysis. Case study based nested analysis, on the other hand, takes the small-n study as its starting point. I make this distinction because of three reasons. First, it is possible to begin a nested analysis with the case study. As Lieberman remarks, a couple of existing studies perform the qualitative analysis first (e.g. Lynch, 2002). Such studies are not covered by Lieberman’s nested analysis that always starts with the regression part. Thus, there exists no methodological guide for case study based nested analysis. Second, the combination of Mb-SNA with a subsequent Mt-LNA, as suggested by Lieberman, comes close to what I conceive of as a case study based nested analysis. There is, however, one important difference. Lieberman suggests to do an Mb-SNA if no model performs satisfactorily. It starts with the selection of typical and deviant cases on the basis of the best of all of the poorly performing regression models. It seems paradoxical to use the residuals for case selection when the model-fit is considered unsatisfactory. Implicitly, Lieberman acknowledges that the use of a deficient regression model for case selection is questionable. He argues that one should not choose cases on the independent variable because they apparently have insufficient explanatory power. If the scores on the independent variable should not be considered, this should apply to the residuals as well, because they are partially determined by the independent variables. Therefore, I contend that the methodologically superior procedure is to begin a new nested analysis that begins with the case
study component, which is a case study based nested analysis. Finally, Lieberman acknowledges that the preliminary LNA might start with “varying levels of background information about a specific case or set of cases” (2005, p. 438). In essence, this is a case study based nested analysis. As will be explained in detail below, one has to make a set of important methodological decisions when progressing from the case study to the regression analysis. It is necessary to be cognizant of these decisions and to make them explicit in order to avoid the existing pitfalls for causal inference.

I will discuss regression based nested analysis in the following section and turn to case study based nested analysis in the section after the next. Within each section, I will explain the consequences of underfitting and overfitting for all components of a nested analysis. In the regression part, these are the model-fit, regression coefficients, and their p-values as measures of their statistical significance. I will first explain how each component is affected by underfitting and how this influences causal inference in within-case analysis. Moreover, I will detail how visual, quantitative and qualitative instruments can be employed for detecting model misspecification.

It should be noted that it is difficult to detail the implications of a misspecification of the causal effect. The consequences hinge on the modeling decision that is made in combination with the nature of the true causal effect. Depending on the situation at hand, the misspecification of the causal effect may be similar to underfitting, overfitting, or both simultaneously (cf. Kohler & Kreuter, 2001, pp. 228-229). Since it is impossible to provide a general discussion of this issue, I will leave it aside in the following section and refer the reader to the econometrics literature for further details (e.g. Greene, 2003, chap. 8).
4. Regression based nested analysis

4.1. The consequences of misspecification

Model-fit tends to be underestimated because more variables are likely to capture more variation on the dependent variable. Underestimation of the model-fit biases the decision about what type of SNA to perform toward Mb-SNA. This effect is beneficial because one should do Mb-SNA in order to identify the left-out variable and construct the true model. It follows that a problem arises if the true model has much explanatory power. In this instance, it is unlikely that the underestimated model-fit will fall below the cut-off point that distinguishes Mb-SNA from Mt-SNA. The model will be deemed satisfactory and Mt-SNA will be carried out. In a sense, this decision is correct because one rightly considers the model to have great explanatory power. The point is, however, that the excluded variable will probably not be recognized in Mt-SNA because one is not looking for an omitted variable (this problem will be discussed in detail below). High model-fit, which generally is welcomed in quantitative analysis, thus constitutes a problem in nested analysis.

The estimator of a regression coefficient of a variable, serving as the measure of its causal effect, is biased and inconsistent if the omitted variable is correlated with this specific variable. One will almost always observe some degree of correlation between the excluded variable and each of the included variables, so bias and inconsistency are generally present. The estimator is biased, but consistent even if the variables are uncorrelated (Gujarati, 2004, p. 510). Therefore, underfitting is highly likely to lead to a misinterpretation of the causal effect. It is impossible to determine whether the estimated coefficient is larger or smaller than the true coefficient in multivariate regression models, which are the rule in quantitative analysis (Bartels, 1995).
In addition, bias and inconsistency may cause a misleading choice of cases. The status of a case as typical or deviant is determined by its residual, which is the difference between the score of a case on the dependent variable and the regression surface. The shape of the latter is affected by underfitting because of its effects on the estimation of the regression coefficients. Depending on the severity of the bias and inconsistency, the residuals thus may prove misleading indicators for the status of a case. Within-case analysis is undermined by bias and inconsistency inasmuch as the status of a case is deduced from its residual. When one believes that the case under scrutiny is typical on the basis of its residual, then one also believes that the observed causal processes are exemplary for a typical case. It is almost impossible to determine the generalizability of a causal proposition without considering the place of the case in the population (Lieberman, 2005, p. 441). Consequently, underfitting might result in the choice of a wrong case, which is selected when the status of the case derived from the underfitted model differs from the status one obtains under the true model. This means that a truly typical case appears deviant, while a truly deviant case seems typical.

The effect of wrong case selection for causal inference in within-case analysis tends to be more serious if one mistakes a deviant case as typical rather than the other way round. Underfitting is likely to go unrecognized if one erroneously conceives of a deviant case as a typical case. The reason is that all of the variables that are part of the model are causally connected to the dependent variable in an underfitted model. If one believes that the case is typical because of its small residual, there appears to be no reason to search for omitted variables, i.e., underfitting will go undetected. The situation is somewhat different if one mistakes a typical case as deviant. A large residual indicates that the case is unusual in some respect. One examines the case with the suspicion that a systematic variable might have
been omitted from the model. The question is whether the left-out systematic variable will be recognized as such. If one erroneously declares a non-systematic variable as systematic, one will overfit the model at hand (with consequences that will be discussed below). On the other hand, the model will improve when the disregarded systematic variable is recognized during within-case analysis and included in the model. The irony might be that, although the original model is underfitted and case selection flawed, the within-case analysis might produce an improved model that is correctly specified.

Underfitting renders the estimator of the variance of a regression coefficient biased, which makes the usual tests for statistical significance misleading (Gujarati, 2004, pp. 510-511). This means that underfitting affects the focus one takes in within-case analysis. According to Lieberman, within-case analysis should concentrate on significant variables and leave aside non-significant variables. Underfitting creates a problem in this respect because the causal significance of a variable is deduced from its statistical significance. The precise effect of underfitting on within-case analysis depends on its impact on the p-values. If the biased p-value is greater than the true p-value, a variable tends to appear as non-significant while it is significant in the true model. The consequence is that one erroneously excludes a variable from within-case analysis. When this neglected variable is systematic, the model becomes even more underfitted because one drops a systematic variable from an already underspecified model. On the other hand, the incorrect p-value may be smaller than the true p-value. The implication may be that a truly non-significant variable is considered significant and included in within-case analysis. The ironic consequence will be to introduce overfitting in addition to underfitting.

After having discussed the effects of underfitting, I now turn to the consequences of overfitting. When the model is overfitted, model-fit tends to be overestimated because more
variables capture more variance of the dependent variable. Overfitting biases the decision between Mt-SNA and Mb-SNA in favor of the former. On the one hand, there is a certain chance that overfitting will get recognized because the overfitted variable will not be causally linked to the dependent variable. On the other hand, one does not throw variables into a model at random at the beginning of a regression based nested analysis. On the contrary, the inclusion of a variable is always, at least partially, motivated by the knowledge of a couple of cases in which the variable has causal impact. Thus, there is a certain chance that one will discover a causal link between the independent and the dependent variable in within-case analysis. For this reason, overfitting is less easy to identify than one might first imagine. The apparent solution is to select other cases for within-case analysis than those on the basis of which one implicitly or explicitly developed the model at the outset (Lieberman, 2005, pp. 446-448). In the practice of case study research, however, this methodologically sound procedure might prove infeasible due to the lack of resources, e.g., in relation to the available sources or language skills.

The estimators of the regression coefficients are unbiased and consistent in overfitted models (Gujarati, 2004, p. 513). Therefore, the causal effect is correctly estimated and case selection is not undermined by bias and/or inconsistency in the ways detailed above. However, case selection is adversely affected by a dimensions problem. In the case of overfitting, the residual of a case is determined in k dimensions instead of k-1 dimensions. When variables are removed from the model, cases show, on average, a trend of moving away from the regression surface, meaning that there is a certain chance that cases that are typical in the overfitted model will be deviant in the true model. In effect, one may select the wrong case, having the consequences described above in the section on underfitting.11
Moreover, overfitting leads to inefficient estimators, so the significance of the included variables is underestimated (Gujarati, 2004, p. 512). Analogous to what was said above, this may undermine the decision regarding the variables to focus on in within-case analysis. One will make a wrong decision if the variable is significant in the true model and not significant in the overfitted model. Because of the apparent lack of significance, one erroneously excludes a variable from within-case analysis. Consequently, the adjusted model will be underfitted and overfitted if the non-systematic variable is not recognized as such in within-case analysis.\textsuperscript{12}

4.2. Instruments for detecting model misspecification

The discussion of the effects of underfitting and overfitting shows that mistakes “travel” through a nested analysis. The adverse consequences of misspecification for regression analysis that were detailed in the previous section are known in quantitative social science. It could be seen, however, that these effects acquire a new quality in a nested analysis because they also undermine causal inference in the case study part. The negative effects on within-case analysis would not exist if one would only do a case study. A nested analysis thus might do more harm than good when conducted improperly. The problem is that the factors on which the crucial decisions for within-case analysis are based, e.g., the residual, are statistical figures whose accuracy is difficult or even impossible to cross-validate in the small-n part.

Later in this section, I will demonstrate that the potential of within-case analysis to detect model misspecification is larger than its current role in nested analysis might lead one to believe. Nonetheless, I will also show that qualitative analysis has several limitations that require the application of graphical and quantitative instruments with which one can search for misspecification. In the following, I will discuss graphical, quantitative and qual-
itative tests for ontological misspecification. Graphical and quantitative specification tests have already received ample attention in the literature. In contrast, a discussion of how process tracing can contribute to model-building and model-testing is lacking so far. I will only touch briefly upon the first two instruments and refer to the references for a more detailed treatment. The baseline of my discussion is that it is necessary to test for misspecification through a mix of graphical and/or quantitative means and qualitative tools. Tests for model accuracy should follow the basic logic of mixed-method designs in order to maximize the opportunity to detect invalid ontological modeling decisions. Figure 1 visualizes the conduct of regression based nested analysis that I will now elaborate in detail.

Figure 1: Regression based nested analysis

Lieberman suggests the use of visual analysis for model-testing in the form of a plot of the actual scores on the dependent variable against its predicted scores (2005, pp. 439-440).
The quality of a model should be assessed in two different ways. First, one should determine the location of cases with great significance. For example, the French revolution is expected to be a typical case in a quantitative analysis of revolutions (Lieberman, 2005, p. 440). Such an approach is misleading inasmuch as it confuses substantive importance with theoretical and statistical typicality. A case is typical in terms of theory if the data-generating process one observes in within-case analysis conforms to the theoretical model. Admittedly, the best guess is that a typical case is statistically typical, i.e., it has a small residual. However, statistically typical cases are not necessarily theoretically representative. It may well be that a theoretically non-representative case has a small residual because of the impact of non-systematic variables (Lieberman, 2005, p. 448). Moreover, there is no reason to believe that a substantively important case has a small residual because the impact of a non-systematic variable might push the case far away from the regression surface. In this instance, the statistically deviant case remains theoretically typical if the large residual is due to a non-systematic variable. The baseline is that substantive importance is a misleading guide for the interpretation of the plot. It only matters whether statistically typical (deviant) cases are theoretically typical (deviant) or not. This cannot be assessed on the basis of a visual interpretation of a plot but deserves a within-case analysis.

The second recommendation that Lieberman makes is to interpret the distribution of all cases. A systematic distribution around the bisecting line is a strong indicator for a flawed model. This is a standard approach in visual data analysis, which, however, exhibits two weaknesses. First, it is not always easy to state by visual inspection alone whether a distribution is unusual or not. This uncertainty might implicitly bias the researcher’s decision in favor of the tested model. Second, and more importantly, the source of a suspicious distribution cannot be unequivocally determined because it may be due to ontological or
statistical misspecification. An example of the latter is the lack of control for autocorrelation in time-series models (cf. Beck, 2001). Thus, the finding of a non-random distribution of cases yields little guidance concerning the nature of the problem. I conclude that graphical data analysis alone is insufficient for detecting misspecification. Essentially, the basic shortcoming of graphical data analysis is similar to the problem of causal inference in regression analysis. There is uncertainty about whether the results obtained at the cross-case level are spurious or the consequence of a data-generating process at the within-case level. Therefore, it is necessary to use the results from the visual data analysis as a guide for the qualitative test of model misspecification.

Quantitative instruments are test statistics that calculate the probability of having a misspecified model at hand. One such test is Ramsey’s regression specification error test (RESET), which I will discuss in more detail in order to exemplify the problems of quantitative tests for misspecification (cf. Gujarati, 2004, pp. 521-523). RESET follows an F-distribution. The null hypothesis tested by RESET is that the model is correctly specified. To give a hypothetical example, imagine a bivariate model that has an F-score of 1.26 and a corresponding p-value of 0.29. This result does not reject the null hypothesis at conventional levels of significance, i.e., one concludes on the basis of RESET that the model is correctly specified.

A test result that allows one to reject the alternative hypothesis that the model is wrong does not necessarily rule out misspecification, however. This is because the quantitative tests for specification error are located at the cross-case level. In this view, quantitative tests for misspecification are similar to ordinary regression estimation. Consequently, such tests cannot determine whether or not the hypothesized data-generating process is in place. The basic problem of regression estimation thus remains. On the other hand, a test result
that leads to the rejection of the null hypothesis is insufficient to demonstrate the presence of ontological misspecification. A significant result may be due to ontological and/or statistical misspecification. For this reason, a rejection of the null hypothesis should not be automatically interpreted as the need to alter the model specification in terms of ontology. The uncertainty about the true source of a significant test result may be mitigated by applying additional instruments testing for various statistical problems, e.g., multicollinearity. Even if one can rule out all of the alternative sources of a significant result, the problem of spuriousness remains, however. It follows that quantitative instruments are insufficient to test for model misspecification if applied on a stand-alone basis. Similar to what was said about visual analysis, quantitative tests are best employed in combination with qualitative means. 17

Visual and quantitative tests for model misspecification should be complemented by qualitative analysis. More particularly, they can suggest where to focus on in qualitative analysis in order to assess whether the test results are spurious or not. Conversely, qualitative means alone are also inappropriate because of two closely connected reasons. The hunch derived from qualitative model-testing that the model might be misspecified needs to be cross-checked by applying graphical and quantitative instruments accordingly. There is good reason to alter the model when the test results improve and to leave it as it is otherwise. Second, one may also estimate the modified model and reach a decision about its accuracy by comparing the output of the old and the new regression estimations. The problem with this procedure is that the best performing and/or most robust model is not necessarily the true model (Hoover & Perez, 2000). Therefore, it is mandatory to supplement the comparison of the regression results with comparative graphical and quantitative tests. In this perspective, a lack of robustness of regression models is not necessarily a problem.
Lieberman argues that one should perform an Mb-SNA if the quantitative results are not robust (2005, p. 441). This does not appear to be necessary if one applies graphical and quantitative instruments. If one, for example, has two models that differ with respect to one independent variable, one can determine graphically and quantitatively which model performs better and perform a complementary Mt-SNA that particularly focuses on this variable.

It has become clear above that graphical and quantitative tests for misspecification are distinct from ordinary regression analysis. On the other hand, the qualitative tests for misspecification essentially capture the case study component of nested designs. If within-case analysis should test for misspecification, however, it is necessary to follow certain guidelines that go beyond what is commonly recommended regarding the conduct of within-case analysis.\textsuperscript{18} It is argued in the literature that case studies serve to elucidate the causal mechanisms linking cause with effect. As will be demonstrated in the following, this procedure unnecessarily restricts the potential of within-case analysis in nested designs. If one aims to use within-case analysis as a means for detecting misspecification, one needs to broaden the purpose of the case study beyond the hunt for causal mechanisms between the variables included in the model.\textsuperscript{19}

In order to elaborate this argument, I first need to distinguish three causal perspectives that one can adopt in within-case analysis: \textit{X-centered}, \textit{Y-centered}, and \textit{XY-centered} (cf. George & Bennett, 2005, p. 218; Gerring, 2001, p. 137). One is interested in the effect of a particular cause in X-centered designs, that is, the independent variable is the starting point for process tracing. A Y-centered analysis is interested in the causes of an effect. The starting point of the analysis is the dependent variable and within-case analysis moves backward in order to determine the relevant independent variables. The starting point and
end point are known at the outset in XY-centered designs. The prime goal of XY-studies is to discern if there are any causal mechanisms that link X to Y.

Lieberman recommends the application of XY-centered perspectives in Mt-SNA. He suggests the selection of typical cases and the search for the causal mechanisms leading from X to Y in order to test for spuriousness (2005, p. 441), which is similar to overfitting. This procedure makes it impossible to detect underfitting.\textsuperscript{20} The qualitative test for underfitting requires the scope of the within-case analysis to be extended beyond the variables that are included in the model.\textsuperscript{21} This can be done \textit{deductively} and \textit{inductively}. Theoretical reasoning and the development of alternative hypotheses is the deductive approach for the identification of omitted variables. Once an independent variable has been derived deductively, the analysis becomes an XY-centered design because one seeks to discern a causal process leading from a known X to a known Y. The pitfall of the deductive approach is that the alternative hypotheses only cover those independent variables that one would plausibly expect to be causally relevant. An excluded variable is not likely to be part of any alternative hypothesis if it is exceptional, given the theory under consideration. Moreover, the deductive approach is unsuitable to explain the outcome of the specific case under scrutiny, which is one major purpose of a nested analysis.\textsuperscript{22}

Both shortcomings of the deductive approach are absent under the inductive approach that requires to take a Y-centered perspective. In this variant of within-case analysis, one moves backward from the outcome of interest, i.e., against the direction of causality, in order to identify the systematic and unsystematic variables that contribute to the outcome. The inductive test for underfitting is probably more resource-intensive. Depending on the research question, there might be a lot of independent variables that are linked to Y which need to be assessed. This is the cost of minimizing the risk of underfitting the model.
In relation to the treatment of X- and Y-centered within-case analysis, the discussion can be expanded to the practice of using primary and secondary sources for causal inference. Lustick (1996) points out that historical studies are often biased inasmuch as they take a specific perspective on the empirical matter. The unreflected use of biased historical sources can create biased narratives and invalid causal propositions. Among other things, Lustick recommends that one should draw on multiple sources in order to mitigate bias (1996, pp. 515-516).

Lustick’s warnings and suggestions matter here inasmuch as a change in the causal perspective during the within-case analysis is interrelated with the use of historical work. Assume that one aims to examine the impact of liberal economic ideas on trade policy decision making. One decides to assess to what extent liberal thinking contributed to French trade liberalization in the form of the Cobden-Chevalier treaty of 1860. Suppose one somehow got to know the work of Arthur Dunham (1971 [1930]), who probably provides the most detailed treatment of the process leading up to the treaty. His study will give one a good idea of the role liberal economic ideas played in this process because he puts great emphasis on the role of Michel Chevalier, who was a liberal economist and close economic adviser to Napoleon III. Although Dunham deals with the foreign policy concerns of France and Britain, he focuses comparatively less on these factors, which are undoubtedly systematic variables if one aims to understand trade policy making (cf. Gowa & Mansfield, 1993). An exclusive reliance on Dunham’s work thus tends to lead to an underestimation of the impact of foreign policy concerns. For this reason, one is well advised to look for other historical studies than those used for the analysis of the XY-relationship of prime interest. The additional material also covers the event under investigation, but pays more attention to other systematic variables, e.g., concerns about external security (in this case for example
Iliasu, 1971). The historical material used in the analysis of the XY-relationship can serve as a guide when it touches on other variables that played a role in the process under scrutiny. Dunham, for example, mentions that foreign policy concerns played a role. This information can be used to search for other studies that focus on this issue. Even if Dunham had completely ignored foreign policy factors, the research question at hand and knowledge of the traditionally antagonistic relationship between France and Great Britain would have been sufficient to lead one to assume that this factor contributed to the Anglo-French agreement and to search for historical work on this topic.

5. Case study based nested analysis

5.1. Inductive model-building

In case study based nested analysis, process tracing aims to develop inductively a model whose explanatory power is subsequently estimated quantitatively. The model can be generated on the basis of the three causal perspectives distinguished above. Studies that explore the effect of a cause adopt an X-centered perspective. The risk of constructing a misspecified model in X-centered designs can be mitigated by taking the following four steps. First, one should perform a forward-moving within-case analysis that starts with the independent variable of interest and stops at a certain outcome that one deems interesting (cf. George & Bennett, 2005, p. 218). This event becomes the dependent variable. The originally X-centered perspective is transformed into a XY-centered analysis because the beginning and the end of the causal process of interest are now known. Given this XY-relationship, one needs to determine other systematic variables that ought to be included in the model. This process covers the second and third phases. In the second step, one empirically determines the variables that contributed to the dependent variable in addition to the independent variable of major interest. As was explained above, the chance to identify all
of the relevant variables is maximized in Y-centered process tracing. By moving backward from the effect to other causes, one is potentially able to determine all of the variables that influenced the outcome. The third phase involves assigning variables as either systematic or non-systematic on the basis of the XY-relationship of interest. Both the independent and the dependent variable influence what constitutes a systematic variable in a specific context. Finally, it is necessary to specify a causal effect for each systematic variable. This is the most difficult task in inductive model-building. The causal effect of a variable is a cross-case characteristic that can only be determined by considering the relationship between the independent and the dependent variable in a larger set of cases. Even a comparative case study design comprising three or four cases is unlikely to provide a sufficient basis for the correct identification of the causal effect. Nonetheless, its specification is one integral element in model-building. If there is no clear indication of what the causal effect of a variable may be, one can simply construct the default model, which is linear-additive, see how it performs in the regression part and employ graphical and quantitative test instruments in order to determine whether this causal effect is the correct one (see below).

The arguments made above on the use of historical sources apply to case study based nested analysis in a similar manner. The material one employs for the establishment of the XY-relationship is potentially biased and inappropriate for generating a correctly specified model. Thus, one should turn to additional sources once the XY-relationship of prime interest has been identified. The goal is to discern other systematic variables that received insufficient attention in the material that is employed in the first step of inductive model-building.

Y-centered analysis differs from X-centered studies only in respect of the first step. In Y-centered designs, the first step involves determining the cause of an outcome that one
deems interesting. If, for example, one aims to explain the conclusion of the Anglo-French agreement of 1860 as a case of bilateral trade cooperation, one might identify liberal economic ideas as an important factor in explaining the behavior of France and Britain. Once a factor like “liberal economic ideas” is set as a systematic variable, the Y-centered design becomes a XY-centered design and the first phase is concluded. The following three steps are similar to steps two to four of X-centered analysis. The second phase comprises the identification of the other variables that contributed to the outcome. In step three, one labels the variables as either systematic or non-systematic, and assigns a causal effect to each systematic variable in step four. XY-centered designs only cover phases two to four of X-centered and Y-centered designs because one starts with a XY-centered perspective at the outset of model-building. There hence is no need for the inductive identification of an effect or a cause that constitutes the XY-relationship on the basis of which one proceeds.

5.2. Consequences of and tests for misspecification

Even when within-case analysis is performed with the greatest of care, it cannot be ruled out that the inductively generated model will be misspecified. Three sources of misspecification exist in case study based nested analysis. First, one might overlook a systematic variable in the empirical analysis so that the model is underfitted. Second, the classification of the identified variables may be incorrect. The model will be underfitted when a systematic variable is considered non-systematic and overfitted in the reverse case. Third, one might assign the wrong causal effect to a systematic variable. The specification of the causal effect probably yields the greatest potential for erroneous model-building for the reasons mentioned above.24

The consequences of the different types of misspecification for regression estimation have been described in the previous section, so I will not reproduce these arguments.
here. What is worth mentioning is that the problem of traveling mistakes is also apparent in case study based nested analysis. Causal inference in the case study and the regression part of the nested design will be invalid when the inductively derived model is misspecified. Therefore, one should not place too much confidence in model accuracy simply because one has combined a case study with regression analysis. The application of instruments testing for model misspecification is indispensable in case study based nested analysis as well. Figure 2 depicts how one should proceed in model-building in case study based nested analysis.

**Figure 2: Case study based nested analysis**

It is unlikely that one will only have to test the specification of one inductively generated model. On the contrary, it is reasonable to expect that one will be uncertain about the validity of many modeling decisions, e.g., about which variables to conceive of as systematic.
Graphical and quantitative tools can be employed to compare the performance of the competing models and determine which performs best. During this process, it should be kept in mind that model-building is a theoretical endeavor and that neither regression output nor the test results will necessarily discern the true model. Thus, theoretical considerations should always predominate over the regression results and the tests for misspecification.

As is the case for regression based nested analysis, the visual inspection of a predicted-vs.-observed plot is not meaningful. One might think that the examined case should have a small residual because one should choose typical cases for model-building. This reasoning, however, fails to distinguish what a typical case is in terms of theory and statistics. There is no guarantee that the case selected for within-case analysis will have a small residual because there might be a non-systematic variable keeping the case away from the regression surface. This should not bother the researcher precisely because the variable is non-systematic. On the other hand, a small residual does not necessarily indicate that the case is substantively typical. Thus, one should employ graphical and quantitative instruments instead of a predicted-vs.-observed plot in order to assess the quality of the model in terms of ontology.

6. Conclusion

Munck and Snyder argued in a recent contribution to this journal that much comparative research lacks methodological soundness. They propose the combination of quantitative and qualitative analysis as a viable method for improving comparative political analysis (2007a, 2007b). I do not deny that mixed-method designs can and should play a central role in comparative research. However, my discussion of Lieberman’s nested analysis, which is currently the most elaborate methodological treatment of mixed-method inference, high-
lights that there are substantial pitfalls waiting for the researcher implementing a nested design.

I have demonstrated that ontological model misspecification leads to methodological mistakes that travel through a nested design. In consequence, causal inference in the quantitative and the qualitative component of a nested analysis is undermined. In these cases, nothing is gained from the combination of different methods. On the contrary, a single-method design might have been preferable, because the application of a nested analysis might produce the illusion of methodological sophistication that leaves one blind to the existing problems. Moreover, I demonstrated that it is difficult, if not impossible, to cross-validate the accuracy of the ontological modeling decisions made in the quantitative and the qualitative part respectively. Precisely because the two methods are closely integrated, it is problematic to use within-case analysis to assess the ontological quality of the regression model and vice versa.

In the face of these problems, I consider it strongly necessary to apply a mix of visual, quantitative, and qualitative tests for model misspecification in nested analysis. In addition to the cross-method validation that is at the heart of nested designs, I deem it indispensable to resort to within-method validation, meaning, for example, that the accuracy of the regression model needs to be determined through quantitative tests for model misspecification. This renders nested analysis even more demanding, given the increasing degree of sophistication of quantitative and qualitative methods. I agree with Munck and Snyder and others (e.g. George & Bennett, 2005) in arguing that the continuing methodological improvements in political science demand collaborative research among scholars who are experienced in different methods that can be fruitfully combined.
References


Notes

1 According to Lieberman, a nested analysis can combine any two methods as long as they are located at different levels of analysis (2005, p. 441). Later in his paper, however, he restricts the discussion to regression estimation and case studies (p. 439). This is important because the residuals play an essential role in nested analysis. Other cross-case methods, e.g., Qualitative Comparative Analysis (cf. Ragin, 1987), do not produce residuals, so it remains unclear how nested analysis can be applied beyond regression estimation. For this reason, and since case studies and regression analysis probably are the most frequently combined methods, I limit my discussion to this specific type of nested design.

2 Lieberman acknowledges that causal inference in a nested analysis might suffer from methodological problems. However, one should do a nested analysis in order “to gain a sense of what can be explained by the theory and data that are available” (2005, p. 439, fn. 9). I do not dispute this argument, but in order to be able to interpret the results derived from a suboptimal analysis, one needs to identify the potential problems related to causal inference, how to detect their presence, their potential effect on causal inference, and how the problems can be mitigated, if at all (King, Keohane & Verba, 1994, p. 6).

3 This section only touches on the elements of a nested analysis that are central for my purposes.

4 A more general discussion of combining small-n and large-n methods can be found in Coppedge (1999).

5 In this context, “n” refers to the number of cases.

6 Another problem is statistical model misspecification, e.g., in terms of the lack of control for serial correlation in time-series regression (cf. Beck, 2001). The effects of statistical misspecification depend on the nature of the mistake. For example, multicollinearity has
different implications than serial correlation. The arguments I make regarding the effects of ontological misspecification can be transferred to statistical misspecification errors if they are similar in their effects.

7 Another way of developing a regression model is formal modeling (cf. Bennett, 2002). With respect to formal modeling, the question also is in how far the model is implicitly or explicitly inspired by in-depth knowledge of particular cases.

8 If the measure for model-fit accounts for the number of independent variables, for example adjusted $R^2$, the model-fit does not increase when the absolute score of the t-value of the excluded variable is less than 1 (Gujarati, 2004, p. 537).

9 Leaving methodological problems aside, one could generally further argue that one should not use statistical significance as a guide for within-case analysis, since this procedure tends to confound statistical and substantive significance (cf. Ziliak & McCloskey, 2004).

10 The caveat mentioned above in the discussion of underfitting applies here too (see note 7).

11 The reverse dimensions problem exists for case selection on the basis of underfitted models. In this instance, however, this problem is not central because of the more severe effects of bias and/or inconsistency.

12 It may happen that one excludes the overfitted variable, but there is no guarantee that this will happen.

13 Lieberman’s (2005) discussion of Mt-SNA and Mb-SNA focuses on the role of process tracing in nested analysis and largely neglects its concrete application. Likewise, other treatments of within-case analysis mostly discuss its place in research designs, but not how to conduct it (e.g. George & Bennett, 2005).
As I explain below, testing the robustness of a model can be framed as a test of the model specification. Therefore, a test for robustness is not explicitly mentioned in this figure.

There exist specific plots that allow to test for underfitting, overfitting, and the misspecification of the causal effect (Schnell, 1994). However, the basic problem remains that a systematic pattern can have multiple sources.

Other tests that follow a similar logic and share similar problems are the Durban-Watson d-statistic, the Lagrange multiplier test (Gujarati, 2004, chap. 13; Greene, 2003, chap. 8).

Another instrument to test for the model misspecification is cross-validation. As is the case for all other quantitative instruments, one cannot rule out that the final model is misspecified. Consequently, the need for assessing the model specification in SNA remains.

Lieberman argues that a within-case analysis could also cover another regression analysis (2005, p. 441). In my view, such a nested analysis is not meaningful. A within-case regression is unable to perform the essential functions of a qualitative within-case analysis, e.g., to rule out spuriousness.

Careful case selection can support the assessment of model specification. For example, one can select a case that is typical in one model and deviant in a slightly differently specified model. SNA can then be used to determine whether it is more reasonable to consider this case a typical or a deviant case (I owe this argument to one anonymous reviewer) (cf. King & Zeng 2007).

Lieberman contends that underfitting would get recognized in the regression part and/or through the inspection of the plot comparing the predicted and the actual scores on the dependent variable. I have argued above that the regression output does not necessarily indi-
cate underfitting and that the interpretation of the plot is insufficient for regression diagnostics.

The question of course is, how many cases should one analyze in order to be sure that the model has been underfitted or not? The analysis of all cases is impossible for simple practical reasons. Given that research resources are limited, I think that one should emphasize depth at the expense of breadth. The lack of breadth of in the qualitative part can be compensated by the use of the visual and quantitative tests discussed above.

As was explained above, systematic and unsystematic variables jointly contribute to the outcome of a particular case. The deductive search for omitted variables exclusively pays attention to systematic variables, meaning that one fails to capture the complete bundle of variables that explain the score on Y.

The neglect of non-systematic variables is unproblematic because these variables should not become part of a general model. Evidently, in practice, the problem is that it is impossible to know whether a disregarded variable is non-systematic if one does not know that a variable has been overlooked in the first place.

The second and third sources of misspecification are not a problem if one is only interested in explaining the case under examination. The causal effect of a variable is irrelevant for the understanding of a particular case. The classification of variables is unimportant because this is only essential for the development of a cross-case model. Disregarding a variable in the empirical analysis is the only source of an incomplete understanding of the case. Isolated case studies, thus, are less demanding compared to case studies that are the basis for model-building in nested designs.