NHH

INSTITUTT FOR STRATEGI OG LEDELSE

DEPARTMENT OF STRATEGY AND MANAGEMENT

Ph.D. thesis

No. 2010/09

ISBN: 978-82-405-0225-6

Three papers on evaluations:

The "what if" in the evaluation of public programs

BY

OLAV ANDREAS KVITASTEIN

Acknowledgements

I am grateful to director Hungnes for his initiative and his efforts to read and comment previous versions of the first paper submitted for this dissertation. I have received several constructive suggestions from members of the EVA forum and participants in a seminar at Innovation Norway, spring 2009. Senior advisors Gry Elisabeth Monsen and Knut Senneseth of Innovation Norway have commented on ideas presented here on several occasions. The first paper is also commented by Dr. Arvid Raknerud and Dr. Morten Henningsen of Statistics Norway. I am grateful for their generous sharing of technical insights. The ideas for the second paper were presented at a visit to Ruhr Universität, Bochum several years ago and I have benefited greatly from the advice of Professor Rohwer and Dr. Ulrich Pötter.

I have received constructive comments from Professor Rune Lines, and openhanded moral support from Professor Sigurd Troye, Professor Sven Haugland, Professor Paul Gooderham, Associate Professor Lasse Lien, Associate Professor Aksel Rokkan and Professor Einar Breivik. I am grateful for their support.

Professor Kjell Grønhaug, Department of Strategy and Management, has been my supervisor for this dissertation. I thank him for his perseverance and efforts. I have later benefited greatly from advices from Professor Salvanes, Department of Economics, and I am very grateful for his support. Professor Espedal, whose generous backing made this dissertation possible, also deserves many thanks.

I also wish to express my gratitude to Vivienne Knowles for proof-reading the manuscript. Finally, I am indebted to Eva Reme for her encouragement and enduring support and to my mother Bjørg Eide Kvitastein for her thought-provoking inspiration. They saved me from the Business School myopia.

The usual disclaimer applies. Responsibilities for errors and omissions are my own.

Bergen, May 2010
Olav Andreas Kvitastein

Introduction

The topic of the dissertation is the evaluation of public programs. I am submitting three papers on related topics concerning methodologies and practices. The "what if" phrase refers to the state of affairs provided that the program under evaluation was *not* carried out. Thus, the "what if" signals a belief that the counterfactual, that is, the most likely situation to prevail without the program, should be the benchmark against which achievements are measured. This is the perspective employed in the first and second paper presented. Both papers present analyses of programs carried out by Innovation Norway for the Norwegian Government.

The first paper demonstrates the applicability of observational methods for the assessment of the program level contributions of two Financial Schemes and two Governmental intervention programs from the predecessors of Innovation Norway in the early nineties. The four initiatives under scrutiny are different in terms of intentions and organization. The Regional Venture Capital Loans program aims at compensating presumed regional funding disadvantages. The Investment Grant program is, as implied by the name, aimed at stimulating physical investments in buildings, machinery and equipment. The FRAM program aims at enhancing leadership skills while the Network program tries to encourage cooperation between companies. The raison d'être for all four initiatives is a presumed market failure and the public benefits from compensating an assumed funding gap when positive externalities are expected. Impact is assessed by means of counterfactual analyses that establish the differences between the factual situation and the counterfactual situation; the situation most likely to prevail provided that the business programs were not implemented. A covariate matching procedure is applied for the construction of the coun-

terfactual. A combination of covariate matching and difference-in-differences analysis is employed for the purpose of finding the best possible estimates for the effects of the programs under scrutiny. The basis for the estimates is two different but related analyses. The first are two-period difference-in-differences and the second is a panel data analysis, both based on matched data.

The second paper concerns the survival value of a leadership-training program, the FRAM program which also is analyzed in the first paper. The perspective, however, is different. Whereas the important issue in the first paper is the effects of the programs in money terms, the focus in the second paper is to what extent the training program has contributed to the companies' abilities to survive in competitive environments, i.e., whether it can be substantiated that the program has helped to keep the firms in business and avoid going bust.

The third paper concerns contemporary trends within the evaluation community and revisits the debate concerning qualitative or quantitative methods. Current developments signal a considerable split between the two camps which in many ways reflects the history of evaluations as a field of research. The remainder of this introduction gives a brief introduction to the development of evaluation as a field of research.

II. The origins and many meanings of evaluations

The three papers all concern *evaluations*, a highly varied field of activities usually viewed as interdisciplinary research. In the 1950s and 1960s evaluation was mainly something that concerned primary schools. The few people engaged in evaluation

research were schoolteachers. With increased attention to the problems of evaluating the effects of new reform programs for schools and new pedagogical techniques for making education more effective, the psychologist, trained in the experimental tradition entered the scene and introduced more scientific methods, and thus the seeds for a new academic field. Clearly, the concept of *policy-research* (Coleman, 1972) is closely related to the evaluation of public programs, but the ambitions throughout the 1980s that evaluations should be a more generic field and not limited to public policy problems left this relation more vague. In the U.S., the Government and Performance Act (GPRA) of 1993[1] provided legal obligation for evaluation of public projects above a specified cost. To my knowledge there are no equivalent legal devices in European countries although the U.S. initiative undoubtedly inspired evaluation activities and boosted what today has become big business, in particular within the European Union. Professional organizations, noteworthy the *American Evaluation Association* (AEA) and the *European Evaluation Society* (EES) are well established with refereed journals and yearly conferences. Public agencies like the Government Accounting Office (GAO) in the U.S. and the European Commission's various auditing offices and the professional organizations issue a substantial number of reports with guidelines and procedures for carrying out evaluation, although it is not obvious what coherent knowledge can be extracted from these numerous publications. The academic sphere of evaluation, assembled in various camps conditional on educational training and paradigmatic devotion, are more coherent but less in agreement with each other even on basic issues like what evaluations are supposed to be. Michael Scriven, philosopher of science and former president of the American Evaluation Association, tell us that "*Evaluation is the process of determining the merit, worth, and value of things*" (Scriven, 1991:1).

[1] Also known as the Roth bill after Senator William Roth

From a political science point of view the Swedish scholar Evert Vedung claims that evaluation is "*the careful retrospective assessment of the merit, worth, and value of administration, output, and outcome of government interventions, which is intended to play a role in the future, practical action situations*" (Vedung, 2000:3), thus restricting the reach of the concept to the consequences of government interventions. The Northwestern University econometrician Charles F. Manski (1996) maintains that: "*program evaluations are efforts to learn from experience in order to improve social decisions*" and the economist Robert L. Darcy (1981) declares that evaluation is "*the systematic collection and analysis of information to determine the worth of purposive organized activity*". Darcy also includes a footnote, which explains that: "*there are different views concerning the nature and purpose of evaluations*".

The tendency to desist from authoritative definition of the term *evaluation* has most likely been to the advantage for the field. An unambiguous, delimiting definition that marked what evaluation *should* be and what it should *not* be, would presuppose an authority that could warrant such a declaration. Fortunately, despite considerable disagreements over many issues, dogmatism of this kind is alien to the mainstream evaluation literature. The ambiguity of the term evaluation serves to lubricate the market for evaluations. The term evaluation has attracted so many meanings that it is appropriate to call it a "*semantic magnet*" (Lundquist, 1976) "that has come to signify almost any effort at systematic thinking in the public sector" (Vedung, 2000). Or, as stated by Carol Weiss, (1972) "evaluation is an elastic word that stretches to cover judgments of many kinds". Moreover, evaluations come in many *forms*. We have *formative evaluations*, typically conducted during the development or improvement of a program, *summative evaluations*, usually conducted after the completion

of a program, and *process evaluation* that focuses on the variables between input and output or upon the process components of a full evaluation, just to mention a few. The ambiguity of the term evaluation and its many subcategories contributes to the dictum that "anything goes" is the hallmark of evaluation practices.

It is the hope that the three papers on evaluations submitted here can contribute to bring down the "anything goes" impression which many think of as proper characterization of the field.

References

Coleman, J. S. 1972. ***Policy research in the social sciences***. Morrist., N.J.

Darcy, R. L. 1981. Value Issues in Program Evaluation. ***Journal of Economics Issues***, XV(2): 449-461.

Lundquist, L. 1976. Några synspunkter på begreppet politisk planering. ***Statsvetensklig Tidskrift***: 121|-129.

Manski, C. F. 1996. Learning about Treatment Effects from Experiments with Random Assignment of Treatments. ***The Journal of Human Resources***, xxxi(4).

Scriven, M. 1991. ***Evaluation thesaurus*** (4th ed.). Newbury Park, Calif.: Sage Publications.

Vedung, E. 2000. ***Public Policy and Program Evaluation***. New Brunswick: Transaction Publishers.

Weiss, C. 1972. ***Evaluation Research: Methods of Assessing Program Effectiveness***. Englewood Cliffs, N.J.: Prentice-Hall.

Contents

Measuring the Long-Term Effects of Active Industrial Policies

Counterfactual Accounts of the Causal Effects of Governmental Actions

Olav A. Kvitastein[1]

Norwegian School of Economics and Business Administration

December 2009

ABSTRACT

The essence of evaluations of governmental interventions is the appraisal of what is achieved by the actions, compared to the situation most likely to prevail in the absence of these initiatives. The purpose of this study is to suggest methods that make such comparison of the factual and counterfactual situations possible and to demonstrate that such analyses can be carried out at low cost by making use of information from available administrative records.

The paper demonstrates the applicability of observational methods for the assessment of the program level contributions of two Financial Schemes and two Governmental intervention programs from the predecessors of Innovation Norway[2] in the early nineties, The Regional Development Fund, and from 1993, The Norwegian Industrial and Regional Development Fund.

The four initiatives under scrutiny are different in terms of intentions and organization. The Regional Venture Capital Loans program aims at compensating presumed regional funding disadvantages. The Investment Grant program is, as implied by the name, aimed at stimulating physical investments in buildings, machinery and equipment. The FRAM program aims at enhancing leadership skills while the Network program tries to encourage cooperation between companies. The raison d'être

[1] Norwegian School of Economics and Business Administration, Department of Strategy and Management, Breiviksveien 40, NO-5045 Bergen, Norway, olav.kvitastein@nhh.no

[2] As of January 1st 2004 Innovation Norway took over the tasks of the Norwegian Tourist Board, The Norwegian Trade Council (NTC), The Norwegian Industrial and Regional Development Fund (SND) and The Government Consultative Office for Inventors (SVO). Innovation Norway is state owned and the objective is to promote private and- socio-economic profitable business development throughout the country, and to release the commercial opportunities of the districts and regions by encouraging innovation, internationalization and image-building. Innovation Norway has offices in all the Norwegian counties and in more than 30 countries world wide. The head office is located in Oslo.

for all four initiatives is a presumed market failure and the public benefits from compensating an assumed funding gap when positive externalities are expected.

Impact is assessed by means of counterfactual analyses that establish the differences between the factual situation and the counterfactual situation, the situation most likely to prevail provided that the business programs were not implemented. A covariate matching procedure is applied for the construction of the counterfactual. The contemporary matching literature is a cacophony of conflicting decisions concerning algorithms to use and procedures to employ. Recent research has revealed a need for modifications of established routines. The paper is based upon the view that matching routines are nonparametric pre-processing methods that facilitate further analysis. A combination of covariate matching and difference-in-differences analysis is employed for the purpose of finding the best possible estimates for the effects of the programs under scrutiny. The basis for the estimates is two different but related analyses. The first ones are two-period difference-in-differences and the second is a panel data analysis, both based on matched data.

The analyses suggest that all four initiatives produce positive contributions and produce lasting impacts that are observable for a considerable period following participation.

Keywords
Program evaluation, difference-in-differences, observational studies, matching models, panel data analyses

JEL Classification Codes
H43, C31, C33

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Causal inference in empirical science is based on counterfactuals. The "what if" statement about counterfactual outcomes is at the heart of evaluations which in turn examine what is potentially achieved by a given project or program. The obvious yardstick for comparisons is the situation most likely to prevail in the absence of the initiative. The central statistical model for the analyses of counterfactuals is the "Potential Outcome Model" (POM) which describes a setting in which one or more units, e.g. human beings, are potentially exposed to a particular treatment, e.g. taking an aspirin, and some response corresponding to this treatment e.g. getting rid of headache or not. The causal effect of interest is the outcome of this treatment relative to some other treatment, usually not taking an aspirin. Thus the source of the medical sounding jargon treatment, control, outcome jargon of the POM model is apparent.

The units under scrutiny in this study are private limited firms and the treatments are a number of financial schemes and programs executed by Innovation Norway. The outcome of interest is added value, defined as the sum of labor costs and net operating profit, in the companies that received the treatment. For the single unit, the individual firm, only one realization of the treatment can be observed e.g. receiving a treatment and observing an outcome in terms of added value. Clearly, this factual information is not enough to establish causation. To learn about the effect of the treatment it is necessary to answer the counterfactual: "What would have been the outcome for the firm provided it had not received the treatment?" The difference between the factual and the counterfactual then measures the causal effect of

treatment. The obvious problem is that in reality each unit can only be observed in one of the two states, as receiving treatment and as not receiving treatment. The conceptual point is that each unit has *two* potential outcomes associated with itself where one can actually be observed and the other is the counterfactual outcome. This state of affairs reveals a clear distinction between singular and general causal claims: If the CEO or other prominent representatives of a recipient company are asked about the effects of treatments the answer could be yes, or no, for a number of reasons. Since there would be no reliable counterfactual at the single unit level, judgments are unsubstantiated and possibly speculative. Adding the answers would most likely be meaningless. Thus, general causal claims have to be based upon the subclass of causal theories usually called probabilistic causation. The Potential Outcome Model (Neyman, 1923 [1990]) facilitates the statistical analysis of the case where two potential outcomes are associated with a single unit. Hence, general causal claims based on probabilistic causation is possible, the problem is that no single unit can be observed in two states, as both recipient and non-recipient of treatment. The firm that receives treatment has to be compared to *another* firm that does not receive treatment. This problem is a central theme of this study.

The main source of information for the study is publicly available administrative records. Recent developments in analyses based on such openly available data, collected for a variety of purposes, labels this observational studies (Rosenbaum, 1995). Observational methods are primarily developed within labor economics and medical testing. The success of observational studies in these settings and the availability of administrative records that constitute inexpensive information about the outcomes of

policies inspired this attempt to investigate the method's potential for policy evaluation applications.

Evaluation of public policies has long traditions within economics and political science. The dominant perspective on public policies is undoubtedly welfare economics, a diverse body of theoretical perspectives that dates back to the 19th century. Its contemporary version constitutes insight that is mandatory to take into account. Thus, reflections over the outcomes of this particular study of the causal impacts of two financial schemes and two intervention programs have welfare economics as a central frame of reference.

The purpose of the study is experimental in the sense that we want to test out the applicability of observational methods and to reflect over the potential pitfalls and fruitfulness of this approach. The available data has shortcomings and represents challenges that call for compromises between the ideal and the possible. If, however, we cannot ascertain the proximate impact of governmental actions, theoretical reflection may prove futile. Thus, the ambition of this study is to contribute by adding a piece of evidence concerning the impact of governmental actions.

Despite some departures from the conventional practices, this study is well within what we can call the econometric evaluation tradition (Heckman, Ichimura, & Todd, 1997; LaLonde, 1986). Thus, the ambition is to establish the causal effects of specific actions.

The phrase *industrial policy* does not indicate that we are solely concerned with actions directed at manufacturing industries. The term is employed due to traditions and pertains to all value generating business activities. Few phrases elicit such strong reactions from economists as industrial policy (Pack & Saggi, 2006). We define industrial policy as any kind of selective government intervention aimed at altering or encouraging business activities in a direction beneficial to society. Conventional wisdom from mainstream economics predicts that the initiatives most likely would make little difference; equilibrium effects would prevail, interventions may have unwarranted or even adverse effects (Pack, 2000) and neutral, not selective policies should be preferable (Orvedal, 2005). The discussions in this paper acknowledge both that the lessons learned from economics applies *and* that industrial policy in this broader sense can make a difference with respect to support and encourage economic activities. The central question of this paper is whether the industrial policies work according to intentions. This should predominantly be an empirical question. In the spirit of Tinbergen (1975) we can distinguish between "*aprioristic policies* " based on theories rather than empirical research and "*empirical policies*" based on experience with alternatives. We hope that this report can be a small increment to the latter category.

1.1 The Purpose and Structure of the Report

The evaluation methods employed are conventional in the sense that they are thoroughly discussed by influential econometricians. Econometricians are not always fond of real data because it tends to invite compromises that make models and solutions less elegant. In this report we have to accept that we are dealing with data that does not conform to what econometric models would ideally require. In

particular, program impact is easier to handle when intervention takes place at one point in time and the subpopulation of concern for the program is straightforwardly identified. If in addition, assignment to program participation was random, estimates of impact could be obtained by trustworthy procedures. In our case, none of these prerequisites are present. Program participation takes place over several years and both self-selection, i.e. you may participate if you qualify, and selection by authorities, i.e. you cannot participate since you do not qualify, are present. Moreover, participants may leave and re-enter programs several times.

At the sacrifice of elegance we have to respect the peculiarities of the data and accept that estimates are less than perfect. The estimates we present are mainly simple arithmetic means and differences between means. A number of compromises have to be made in order to provide the best guesses when accurate estimates are infeasible. The spirit of the analyses is to explain how compromises are made and how solutions are constructed.

The report is divided into fifteen sections. The first is the introduction. Sections 2 to 5 discuss the problem under examination, the data at our disposal for the purpose and the measurements used. Section 6 discusses methods and describes the matching strategies and the outcomes of the matching procedures. Sections 7 and 8 and 9 present the preprocessing that is carried out and the quality of the matched pair datasets that is the results of the preprocessing. Sections 9 to 11 describe the analyses of the matched pairs and section 12 provides scenarios for possible interpretations of the findings. Section 13 gives a statistical assessment of the findings and evaluates their worth as evidence. Section 14 includes the replies to invited com-

mentators and a general discussion of central choices and decisions. Section 15 concludes the report.

2. The Cases under Scrutiny

The two financial schemes *Investment Grants* and *Regional Venture Capital Loans* are directed towards the stimulation of various business activities and are governed by Innovation Norway. *Investment Grants* can be applied for by all firms located in designated regional development areas. Predominantly, grants are provided for investments in machinery and plant equipment, and in some cases buildings. Besides eligibility based on location, innovative projects that are believed to succeed and ventures that stimulate entrepreneurship are prioritized. Regional Venture Capital Loans constitute a kind of top-up finance primarily intended for SMEs and may cover up to 50% of investment costs. Interests are usually somewhat above standard credit market interest but risk is reduced by means of a contracted right to a 50% discount on reimbursement in the case of failure. Normally, Regional Venture Capital Loans are disbursed *after* investment. Regional Venture Capital Loans are directed towards both new and established firms and are not dependent on the location of applicants. Since a considerable percent of the recipients is involved in both programs, the two instruments are treated both as jointly coexistent and as separate instruments. Thus, the analyses of these programs are split into *three* sub-categories: (1) Regional Venture Capital Loans only; (2) Investment Grant only; and (3) the union of the set of members in the Regional Venture Capital Loans group and the Investment Grant group combined.

The first program under scrutiny, *The FRAM program,* was developed in 1992-1993 by The Norwegian Industrial and Regional Development Fund (SND) as a follow-up program of a technology transfer program (BUNT). The FRAM program is now

administered by Innovation Norway. The FRAM programs are tailored towards various activities such as culture (FRAM-Culture), the establishment of new businesses (FRAM-entrepreneur) and the development of leadership skills in small and medium-sized enterprises (FRAM-Strategy and Innovation). The last one, the FRAM aimed at developing leadership skills and strategic competence for SMEs is the one under scrutiny here. The program's ambition is to contribute to lasting transfer of competence, improved competitive power and profitability. Thus, evaluations require a longitudinal design that facilitates analysis of to what extent any lasting effect can be traced.

The program was offered to the SMEs as an individually tailored development program, and companies are invited to participate, provided that they are not be engaged in competition with other participants. Found eligible for participation, companies are "screened" by criteria concerning the company's ability to benefit from the program. The program is organized as separate projects for groups of similar firms, usually groups of 8 to 12 persons, who are leaders of their respective companies. The program is run by experienced process consultants, and concentrates on strategic development processes, much in line with the standard textbook theory of normative strategic management.

The second program we study, *The Network Program,* was established in 1991 by the Norwegian Foundation for Industry (Industrifondet) and continued by the Norwegian Industrial and Regional Development Fund (SND) and Innovation Norway. The basic idea behind the program is that cooperation between firms in the long run has beneficial effects in terms of profitability. This idea is anchored in

theories collected from economic geography, e.g. theories of agglomeration (Krugman, 1991) transaction cost economics (Williamson, 1985) and theories about innovation and economic development. The networks are organized in smaller groups of five to ten firms that focus on themes and activities of mutual interest. All group members (firms) get together a least 3 to 5 times a year and all firms carry out a least one project during the network period to ensure that the intentions behind the program are attended to. Otherwise however, it is left to the individual firm to get the most out of the program. The expected effects of the program rely heavily on the belief that cooperation between firms tends to develop clustering effects (Reve, 1994). The Network program was phased out in 1998 pursuant to a negative evaluation by the consultancy company Econ. Thus, the present study makes it possible to judge the correctness of that decision.

2.1 Theoretical Justifications: Why should the initiatives work?

Public venture capital initiatives i.e. programs that make equity-like investments in firms, in particular in young or technologically advanced firms, have been around for at least four decades and is prevalent worldwide in the more advanced capitalist nations. Program designs may differ, but usually hinge upon two shared assumption: 1) that the private sector provides insufficient capital to new firms, and 2) that the government either can identify investment which will ultimately yield high social and/or private returns or can encourage financial intermediaries to do so (Lerner, 2002). The Norwegian Regional Venture Capital Loans initiative clearly shares these assumptions but have the regional requirement as an additional dimension. The inclusion of the regional aspect reflects both politically decided priorities and the belief that the degree of market failure coincides with a regional dimension in the

sense that capital supply problems are more severe in the less populated rural areas in Norway. None of these underlying assumptions are thoroughly investigated; it is simply assumed that such is the case.

A considerable number of firms are eligible for application for Regional Venture Capital Loans; a fraction of the firms get their loan approval after a comprehensive screening procedure. Each individual project receives careful evaluation according to criteria similar to those of the private venture capitalists whereas other decisive factors such as the likelihood of positive spillovers are unique for public funding.

Investment Grants follow a similar screening procedure where the probability of allotment depends heavily upon the prospect of the projects in question and, as for Regional Venture Capital Loans, the possibilities of positive externalities. Thus Investments Grants are carefully evaluated *as investments* with an expectation of a positive payoff to society, which, by definition is also the investor of public funds.

Both Investment Grants and Regional Venture Capital Loans are, as the market failure assumption implies, contingent upon the availability of private investments and loans. So why can it be the case that public funding works where private venture capital organizations and investors do not find efforts worthwhile? Can it be because private venture capital funds are in short supply in Norway? One suggested proposition is the so-called *Certification Hypothesis*: Public venture capital awards can certify that firms are of high quality and thus reduce the information problem for private venture capital organizations. It can be argued against this hypothesis that asymmetric information should from the outset be a problem for the bureaucrats

responsible for selecting firms for grants and loans, and thus that moral hazard and opportunistic behavior would render the quality of the chosen firms less trustworthy. On the other hand, it can be argued that the comprehensive screening process carried out by Innovation Norway is considerably more thorough and costly than private venture capitalists and investors would ever care to undertake. Thus, part of the job that the private financier otherwise would have to do is now carried out by the government for free. Empirical research seems to support the certification hypothesis (Himmelberg & Petersen, 1994; Hubbard, 1998; Klette, Moen, & Griliches, 2000) but do not provide unambiguous answers to what extent governmental efforts can succeed in reducing market failures.

The FRAM program is justified by the hypothesis that there is a positive relationship between the quality of management and business performance. If this link holds true, public money spent encouraging and educating leadership would increase economic performance and thus provide a positive contribution to the society. Even though the FRAM program is aimed at SMEs where the distance between leadership and the value creating parts of the organizations may be assumed to be short, empirical studies point toward that the leadership – performance link is notoriously hard to establish (March & Sutton, 1997). Moreover, the FRAM program can be accused of falling prey to myths about the significance of leadership (March, 2005) and thus has created a program that does little more than reinforce these myths.

The Network program is based upon the idea that economic growth can be nourished by network creation is closely linked to the notion of industrial clusters (Britton, 2003; Kacirkova, 2009; Kvitastein, 1995; Romanelli & Khessina, 2005; Takeda,

Kajlkawa, Sakata, & Matsushima, 2008) and endogenous growth theories (Romer, 1986; 1991; 1994). Although it is well established that industrial clusters under positive circumstances may induce growth, the underlying mechanisms at work are not easily revealed (Johansson, Stough, & Karlsson, 2005; Roterud, 2005).

2.2 Hypotheses about the outcomes of schemes and programs

Clearly, we have no hope of uncovering the underlying mechanisms that will ultimately decide the success or failure of the financial schemes and intervention programs. The designs of our analyses do no provide any evidence for the lower level mechanisms but invite speculations concerning the causes of outcomes. We have to accept the effects at the molar level (Cook & Campbell, 1979) at which we operate as caused by the financial schemes of program we analyze i.e. as *molar causation*; an overall causal relationship between a treatment package and its effects, in which both may consist of several parts (Shadish, Cook, & Campbell, 2002). That is, our sole ambition is to evaluate the surface outcomes in terms of added value i.e. the reward to labor and capital in the firms involved.

Conventional wisdom from mainstream economics predicts that the initiatives would most likely make little difference while the official justifications assert that these actions should be beneficial to society. Neither the conventional wisdom concerning governmental intervention nor the theory-based good reasons for why the actions should work, provide much guidance with respect to the formulation of hypotheses concerning outcomes. We can, however, make some guesses based on the implications of the justifications for the different initiatives. *Loans* imply a higher commitment for the firms since they involve a contractual relationship regarding repay-

ment. *Grants* can be distinguished from loans by the higher degree of involvement by the bureaucrats of Innovation Norway. This higher concern is most likely due to an increased awareness towards the potential moral hazard involved and the embarrassment that follows from failure. Thus, compared to the FRAM program and the Network program we can expect that the screening process is substantially more thorough for Regional Venture Capital Loans and Investment Grants. Furthermore, we expect the selection process to be even more robust for grants than for loans. Lastly, we believe that the basis for the FRAM program is the weakest one of the four initiatives.

Following these lines of reasoning we can establish a kind of *ordering* of our expectations. Thus, assuming that the average effects at the program level are above zero for all four initiatives, and since it is a central goal to compare the four initiatives, we propose that we expect the following order to prevail:

Hypothesis 1: *The combined Regional Venture Capital Loans and Investment Grant allotments will produce the highest yield in terms of added value.*
Hypothesis 2: *Investment Grant will produce the second highest yield in terms of added value.*
Hypothesis 3: *Regional Venture Capital Loans will produce the third highest yield in terms of added value.*
Hypothesis 4: *The FRAM program will produce the lowest yield in terms of added value.*
Hypothesis 5: *The Network program will produce the second lowest yield in terms of added value.*

This ranking implies that we have very low expectations with respect to the FRAM program and the Network program. In fact, the convincing arguments about the

weak link between leadership and performance put forward by James G. March (2005) make us believe that we should expect very little or nothing from the FRAM program in terms of demonstrable effects on added value. The low expectations regarding the Network program have a less elaborate explanation; we have considerable trust in the theories of agglomeration and the growth-inducing effects of industrial cluster, but recent empirical analyses of the Network program (Econ, 1998) indicate that the program has no observable economic effects.

3. The problem addressed

The primary objective of this paper is to put forward statistically defensible judgments[3] about the outcomes of two financial schemes and two motivational programs aimed at correcting market failure and encouraging industrial activity in Norway. *Outcomes* are defined in terms that closely resemble the concept of *additionality.* That is, we intend to make judgments that are instructive with respect to the most likely *effects* of these schemes and programs, i.e. the outcomes for the firms that participated in these programs compared to the most likely outcomes for the same firms provided they *had not* participated in the programs. Strictly, the term *additionality* requires that the activities presumably set off by the programs would not have been undertaken without the existence of the programs. We can, however, never know the exact answer to questions concerning what a given firm would have done in the absence of the actions in question. Investigating the most likely outcomes means a reasonable relaxation of the strictest interpretation of additionality and confidence in observable indicators of likely effects. Changes in *value added*, defined as *the sum of operating result and labor costs* is an indicator of yearly variations in economic outcomes that is observable provided we have access to accounting records. Thus, the proportion value added, defined in this manner, *that can be attributed to impacts caused by the programs* is our central measure of outcome, and hence, of the additionality of programs. Terms like *cause* and *effect* are challenging and usually linked with experimental thinking. Thus, terms like *treatment* (intervention) and *control* (non-intervention) will be used throughout the paper although no therapeutic or other health metaphor is implied.

[3] The informational basis of a judgment identifies the information on which the judgment is directly dependent and-no less important asserts that the truth or falsehood of any other type of information cannot directly influence the correctness of the judgment Sen, A. 1990. Justice - Means versus Freedoms. ***Philosophy & Public Affairs***, 19(2): 111-121.

3.1 The Use of Observational data vs. Experimentations

The classical randomized experiment is the archetypical design for causal analysis. Valid causal inferences can be achieved via the classical experiment roughly due to three critical features: (1) Random selection of units to be observed from a given population, (2) random assignment of treatment to each observed unit, and (3) relatively large sample size. These three features combined generally render causal conclusions feasible. While features (1) and (2) facilitate conclusions about statistical inference, (3) guaranties the absence of omitted variable bias. The experimental design guides the data-generating process which decides the statistical properties of the data. The classical randomized experiment is an ideal type and most social science research usually fails to meet at least one of the three features. Failure to conform to this ideal type may, however, produce substantially biased conclusions. *We define observational data as data based upon data-generating processes or collection mechanisms that does not meet all three features of a classical randomized experiment* (Ho, Imai, King, & Stuart, 2007).

The consequences of this definition are far-reaching and some researchers may find it prohibitively strict. It is, however, important to realize the consequences of ignoring the rules of inference and the strenuous requirements for causal claims. In evaluation research causal claims may be imperative for summative evaluations. To refrain from causal claims may not be a good solution, particularly not when followed up by a rhetoric that points toward causality with some reservations (Scriven, 1993). When it is obvious that causal claims are in demand, the justification for such claims should not be ignored.

The empirical basis for this paper is administrative records. Thus, the data generating process deviates substantially from the classical experiment. We are dealing with observational data that have been generated by processes that do not justify any parametrically based inference but represent reliable and comprehensive sampling frames for many purposes. As such, administrative records represent an underutilized source of information (Roed & Raaum, 2003) about social and economic consequences of governmental actions and outcomes.

3.2 Research questions

The obvious question of most summative evaluations is; does the program work? Here the question could be rephrased to ask if the companies which received benefits from the government, either in the form of *Regional Venture Capital Loans*, *Investments Grants,* and participation in the *FRAM program* or the *Network program* do better than those who did not. And, provided that the companies that received benefits from the government *do* better than those which didn't; to what extent can it be substantiated that this improvement is *caused* by the program in question? These are the obvious research questions. *The purpose of this paper is simply to investigate whether the average outcome in groups of companies that received governmental benefits in the form of participation in one of the above mentioned programs is significantly higher than in comparable companies that did not receive such benefits.* Although this may seem like an overtly simple task, in essence a simple comparison of two means, the procedures that justify emulations of the properties of the classical experiment are somewhat involved. The procedures are, however a prerequisite for a proper answer to the fundamental question of summative evaluations; did the program work according to intentions?

The dominant problem in this kind of studies is the so-called *selection problem*. The firms that apply for support from Innovation Norway may be different from those that do not apply for support. In addition, internal procedures, i.e. various rules and criteria for eligibility may produce other selection processes among those that have chosen to apply. To be able to make comparisons between the firms that receive benefits and those that do not, the selection problem has to be dealt with. Thus, a central goal of this report is to compare recipients and non-recipients in ways that make it possible to extract the most likely difference between these two groups in term of outcomes, i.e. differences in added value generated.

4. The Evaluation Problem

The key to estimating the impact of a program is constructing the *counterfactual* outcomes, representing what would have happened in its absence (Heckman & Smith, 1999). Thus, the logic of empirical analysis follows the *potential outcome approach to causality* introduced by Roy (1951) and Rubin (1974). Although the notation and general framework are mainly the same, the present-day label for this approach is the *counterfactual approach to causality* (Heckman, 1999; Heckman & Smith, 1995, 1998b; Lewis, 1973, 1986; Pearl, 2000b; Pearl, 2000c; Winship & Morgan, 1999). In our context this means that on the basis of administrative records we measure outcomes of interventions as the difference between the *actual* outcomes that we can observe and the *counterfactual* outcomes that most likely would have been the case provided that no intervention had been carried out. The problem is that the counterfactual, by definition, is something that *does not exist*. Thus, it has to be constructed in the most plausible manner. A crucial feature necessary for feasible detection of the effects of an intervention is that we are able to distinguish between the group of people, companies or other distinct targets of the action that are *exposed to treatment* (the program) and those that are *not exposed to treatment*. No administrative record has information for more than one state so no unit can be observed in both states; as both receiving the treatment and not receiving it. If this was possible, the effect could be calculated by comparing the two states for the same unit. A major problem is that the effect of treatment has to be calculated by comparing a unit that received the treatment with *another* unit that did not receive the treatment.

The potential outcome framework and the counterfactual approach to causality shares the common denomination *causal analysis*. This line of research is, however, distinct from path analysis and structural equation modeling *(SEM)* (Bollen, 1989; Jöreskog & Sörbom, 1989; Long & Bollen, 1993; Wold & Jöreskog, 1982) despite the fact that both areas are labeled causal analysis (Winship et al., 1999). Moreover, the terminology is distinct between the two strands of research. While *SEM*-modeling has no explicit reference to the terminology of experiments, the counterfactual causality framework permanently follows these terms. For the remainder of this paper we follow the conventional notation of the *potential outcome framework* (Rubin, 1974) that has its roots in the works of Neyman, 1923 [English translation (1923 [1990])] Quandt (1972) and Roy (1951). Thus, we use jargon like "treatment" for the groups of firms subject to intervention and "control" for firms that we use for constructing comparison groups. This terminology is consistently used in the literature and also throughout this paper.

Formally, say a unit can be in either a treated state, denoted state "1" or an untreated state, denoted state "0" and say outcomes Y_1 and Y_0 are associated with each state. The gain from treatment could then be calculated as the difference $Y_1 - Y_0$. Because we cannot determine impact of treatment for the individual unit, we have to rely on the distribution of impact across units, call it $F(\Delta)$ or on certain features of this distribution. The expected gain to a randomly selected unit in the population, denoted $E(\Delta)=E(Y_1 - Y_0)$ refers to the expected value or population average. Given that the entire population was included, as could be the case for e.g. a tax reform, this parameter provides information necessary to carry out benefit-cost analysis when combined with information about average cost (Heckman & Smith, 1998c). For a pro-

gram that targets certain groups, it is more informative to focus on what happens to those who actually participated in program. Denoting participation *d=1* and non-participation *d=0* we can write the distribution of gains for participants as $F(\Delta|d=1)$ and the impact for participants as $E(\Delta|d=1)=E(Y_1 - Y_0|d=1)$. The problem is that we do not know $E(Y_0|d=1)$,; it has to be estimated, and this is not straightforward. We cannot use the mean outcome among non-participants as a proxy for what would have happened to participants had they not participated. This is seen by subtracting the mean outcome among non-participants from the mean outcome of participants, $E(Y_1|d=1) - E(Y_0|d=0)$, *yields*

$$\{E(Y_1 \mid d=1) - E(Y_0 \mid d=1)\} + \{E(Y_0 \mid d=1) - E(Y_0 \mid d=0)\}$$

The first term in the curly brackets gives the mean impact of participation, and the second term represents the *selection bias* caused by the fact that non-participants may differ from participants in the non-participating state. That selection bias may be different from zero is easily seen in, say, a program that involves a small business firm where economic difficulty is the criterion for being eligible for participation. In such a case, non-participants could be expected to have outcomes higher than participants, and hence, a negative selection bias could produce incorrect estimates.

Randomization solves this problem, provided that randomization does not alter the pool of participants or their behavior and that close substitutes for the treatment are not available. Given that randomization is used both for the treatment (participant) group and the control group (non-participants) and that the control group is denied access to the treatment group upon completion of the selection procedure, the outcomes of both groups in the zero state before treatment would be equal. That is,

$E(Y_0|d=1) = E(Y_0|d=0)$ and the right hand side term in the curly brackets in the equation above would cancel out, implying no selection bias.

It is well known that randomization in general is hard to implement in practical settings due to ethical reasons such as fairness or social justice. In the case of financial schemes and programs for encouraging or inciting business activities, random assignment of benefits is probably politically unacceptable and likely to stir up reactions. Moreover, as pointed out by Heckman (1992), randomization does not remove selection bias. It balances the bias between the treatment group and the control group (Heckman et al., 1995). Selection bias may arise from sources that are hard to level out by means of randomization. It can be generated by missing data on the common factors affecting participation and outcome, or it may occur when random assignment causes the kind of units participating in the program to differ from units participating in the program as it normally operates (Heckman et al., 1995). This phenomenon is usually labeled *randomization bias.* Another cause of randomization bias is changes in participant behavior that operate via reactions towards participation and is measurable prior to treatment. Furthermore, *substitution bias* may occur when members of the control group gain access to close substitutes of the treatment under consideration. In e.g. training programs for small business firms, this phenomenon is likely to happen when someone in the control group recognizes that they are denied a service and react by seeking similar services offered elsewhere.

Even though it is nearly impossible to rule out all sources of bias, randomized experiment is traditionally regarded as the queen of quantitative evaluations. It is said that the late Donald T. Campbell expressed some regret over his celebrated book on

quasi- experimentation (Cook et al., 1979) because it may have mislead some researchers carry out a quasi-experiment when randomized experiment was a feasible and better solution. Experiments are, however, no universal remedy. Longitudinal studies cannot be based upon experiments (Weiss, 1982) and bias may be present even in an experimental setting (Heckman et al., 1995).

The selection problem has for a long time (Heckman, 1977; Heckman, 1979) been known as one of the major obstacles for evaluation studies. The hopes for unbiased estimates of impacts necessitate that this problem is attended to.

4.1 The Counterfactual account of causality

Recent work within so-called "observational studies" (Rosenbaum, 1995) that explores the "potential outcome" or "counterfactual" model of causality has produced valuable insights to the understanding to the understanding of governmental interventions. The philosophical underpinnings of these lines of reasoning is the groundbreaking work of David Lewis (1973) which provides the logic of counterfactual reasoning. Counterfactual thinking is common and probably unavoidable in ordinary languages. The "what if" question is prevalent in reflections over what could have been the case. "If I were you" opens empathetic reflections based upon a counterfactual in the same manner as the question "what would have happened if the government had not bailed out those banks"? The alternatives that trigger thinking are the non-existing counterfactuals. Noteworthy, statistical thinking does not offer an easy access to the empirical world in a way that effortlessly reveals the counterfactual in question. Statistical modeling does, however, offer a variety of approximations and compromises that facilitate a calculus of interventions (Pearl, 2000a)

that makes it possible to make judgments concerning *the most likely* counterfactual. Thus, differences between the observed factual and the most likely counterfactual may provide an estimate of the most likely effects of the intervention under scrutiny. Noteworthy, this way of thinking does in fact reverse conventional thinking about the relationship between cause and effect; instead of looking for the causes of observed effects *we search for the effects of known or assumed causes* (Holland, 1986). Noteworthy, there is considerable debate over the *meaning* of causation and the logic of counterfactuals (Choi, 2007) in general philosophy (Schaffer, 2007) and the philosophy of science e.g. (Kvart, 1994; Lee, 1986; Schweder, 1999; Shalkowski, 1992; Yablo, 1992) and in psychology and brain sciences e.g. (Arokiasamy, Robertson, & Guice, 1993; Chang & Herrmann, 2007; Mandel, 2007; Pollard, 1983). These debates may be of considerable interest for evaluations, in particular for the interpretation of results and the translation of results into policy implication. In this paper we focus on the retrospective *measurement* of effects at the molar level (Cook et al., 1979) where effects can be observed i.e. we accept causality at the level observed even though the causal mechanisms involved remains unknown. More explicitly, using data based on administrative records, we can identify three levels of causal inference; 1) unit-level causal inference, 2) subpopulation causal inference and 3) population-level causal inference (Holland & Rubin, 1988).

4.2 Parameters of interest for evaluations

Unit-level causal inferences can be defined as the difference $Y_t(u) - Y_c(u)$ and assumes homogeneity for every pair of matched treatment-control units. Clearly, inferences at the individual level may provide the basis for a variety of meaningful parameters of interest at the subpopulation level. Heckman & al. (2001) discuss four

parameters of interest , LATE, MTE, ATE and ATT, where LATE means Local Average Treatment Effect, MTE denotes Marginal Treatment Effect, ATE stand for Average Treatment Effect and ATT is Average Treatment Effect on the Treated. A fifth parameter ATC, Average Treatment Effects on the Controls could be added to the list. LATE (Imbens & Angrist, 1994) is defined as a change in the treatment effect for the treated that is induced by a change in the instrument (e.g. the intensity of the treatment) whereas MTE (Bjørklund & Moffitt, 1987) investigates treatment effects at specific levels of the instrument. Both LATE and MTE assume that changes in the instrument may affect the treatment decision (the selection into the program) and that changes in the instrument are independent of the outcome variable. ATC is of interest only if the goal is to investigate possible violation of central assumptions, such as the absence of equilibrium effects. ATE and ATT are by far the most common in applied research. ATE is simply the mean causal effects for the units whose characteristics are represented by $\boldsymbol{X}$, (our vector of control variables, see section 6.3, page 61) averaged over all units, whereas ATT is the mean causal effects averaged over those units that actually received treatment. In most applications, the estimation of treatment effect for each observation is not a central goal. This paper focuses on the average effects of treatment on the treated *(ATT)* and is thus solely concerned with effects at the subpopulation and population level.

4.3 Administrative Records and the Timing of Evaluations

Clearly, the question of effects is at the core of the concerns of policy makers. When large amounts of taxpayers money are spent for a purpose that is legitimate because it is beneficial to society the demand upon governmental *accountability* may require that outlays are justified i.e. that the actions taken produced the desired re-

sults. The asynchrony of events complicates matters in the sense that while effects are observable only after an (unknown) amount of time-elapse, the policy process is future oriented and interventions may be set off by a perceived urgent need for action. A common pseudo-solution to this problem is premature request for evaluations where the demand for legitimating the actions undertaken dominates and the researcher is asked to look for something that is yet to occur. With increasing acceptance of evaluations as an integral part of the policy process this phenomenon has become an inseparable part of the evaluation problem that has the potential of corrupting the evaluation process. As indicated in Figure 1, the manifestation of effects may not necessarily crop up when it is most convenient for the researcher or the authorities that want to have an evaluation done. In the worst case the research is undertaken right before effects are observable (exhibit *b* and *c*) or at a point in time where effects by coincidence are below average for a certain time span (exhibit *e*) or when effects are no longer observable (after time P_4 in exhibit *d*). The configuration in exhibit *a* is mostly wishful thinking and highly unlikely to happen.

The various configurations of effects in Figure 1 cast doubts over the usefulness of cross-sectional evaluation procedures. Excluding the time aspect may lead to erroneous conclusions. Thus, we advocate the use of repeated cross-sectional data or better, panel data whenever accessible. In this paper we rely on administrative records that are openly accessible to everyone, a source of information that is said to be underutilized (Roed et al., 2003). Administrative records offer the opportunity to construct panel data where the individual company can be tracked over time.

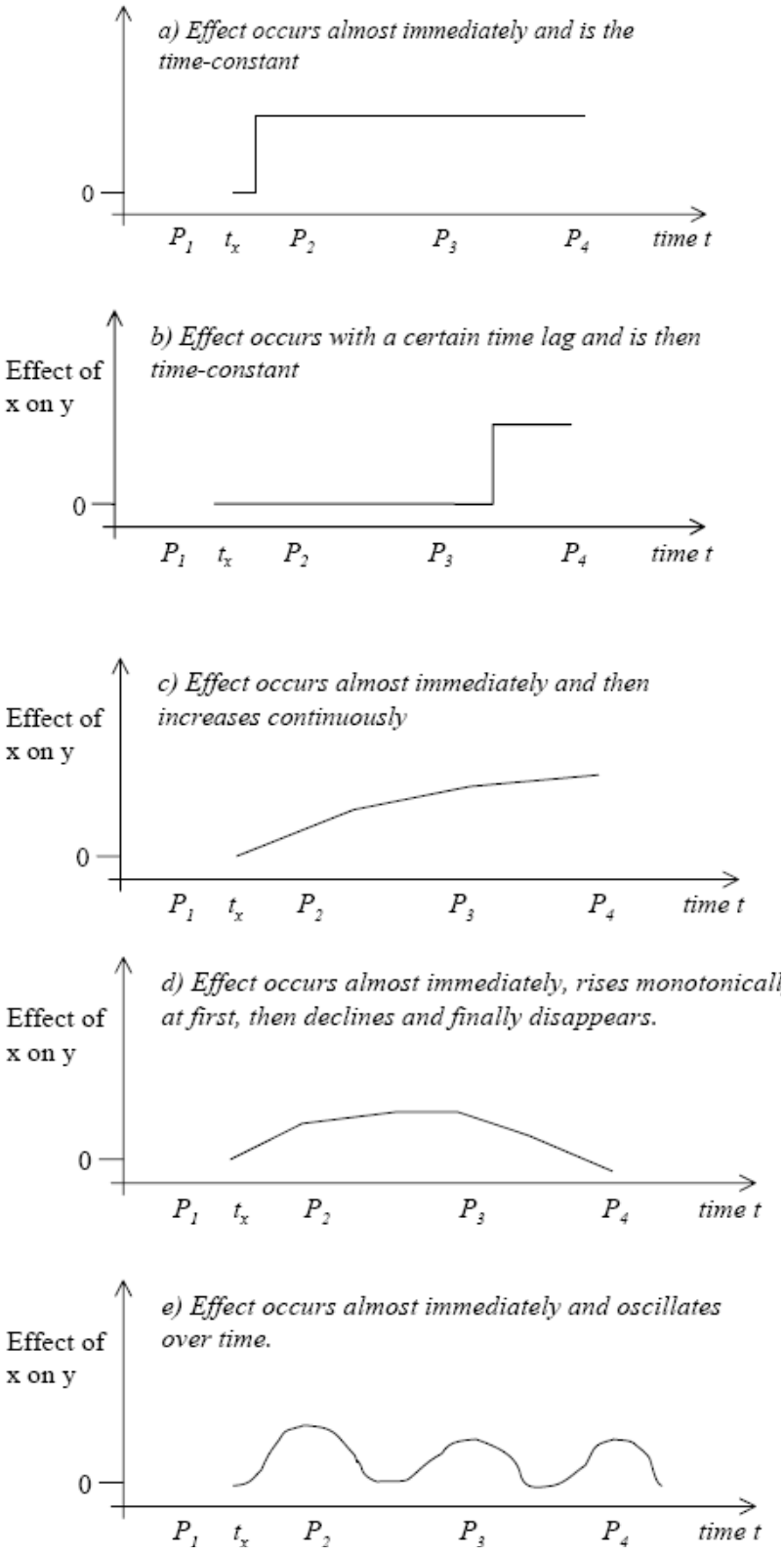


Figure 1 Temporal shapes of how a change in variable x, at time *tx*, effects a change in variable y[4]

[4] Adopted and modified from Blossfeld and Rohwer Blossfeld, H. P., & Rohwer, G. 1997. Causal inference, time and observation plans in the social sciences. ***Quality & Quantity***, 31(4): 361-384.

5. The Cases and Data

The data are derived from accounting records collected by Dun & Bradstreet, Innovation Norway's internal data-warehouse *BUSTER* and administrative records made available by Innovation Norway. The administrative records provide information concerning projects and programs; the accounting date supplies necessary financial information and data on location, time of establishment, number of employees and time for bankruptcies and liquidation. In addition, Innovation Norway made available a dataset that facilitates the exclusion of companies involved in other projects or programs administrated by Innovation Norway or other governmental agencies.

5.1 The Administrative Records

The data for the *Investment Grant* and *Regional Venture Capital Loan* programs are collected from administrative records on 5831 individual decisions on allotment of capital involving 67 individuals, and 3298 companies, predominantly privately owned limited companies over the period 1990 to 1994. Thus for this dataset we can identify the amount of money assigned to the individual firms. We were able to identify acceptable accounting records from Dun & Bradstreet for approximately 66% of the limited companies. Records for institutions and companies with special allotments such as SIVA, and Norwegian Pipelines ltd together with allocations to municipalities and counties were deleted from the data. Table 1 shows that we were able to identify the pre-program data for all records and that we have a large pool of potential controls for all five treatment periods.

Table 1 Participants & potential controls for the Investment Grant and Regional Venture Capital Loans programs

Year		Treatment		Control
t-1	***t***	*t-1*	*t*	*t-1*
1989	1990	425	425	20007
1990	1991	337	337	26653
1991	1992	237	237	29775
1992	1993	184	184	15733
1993	1994	252	252	41317
N =		*1435*	*1435*	133485

The *FRAM program* data set consists of 425 firms that have participated in the program over the period 1992 to 1997. To be able to carry out the matching procedure we need information about the company in question for *the year prior to the first time it entered the program.* As we can see from Table 2, there is a slight difference between the column *t* and column *t-1* indicating that for a few firms, data for the year prior to entering the FRAM program could not be identified. The loss of data is not substantial. The FRAM program aimed at developing general leadership skills and strategic competence for SMEs. We should, however, bear in mind that the dataset we have, includes only a minor portion of all companies that have participated in the many variants of the FRAM program. Thus, the conclusion from this study concerns outcomes from the group of included companies only.

Table 2 Participants & controls for the FRAM program

Year		Treatment		Control
t-1	***t***	*t-1*	***t***	*t-1*
1991	1992	18	19	16783
1992	1993	19	21	17306
1993	1994	105	110	17595
1994	1995	118	131	18287
1995	1996	126	142	18069
1996	1997	39	46	17748
N =		*425*	*469*	105788

Table 2 shows that we have a considerable pool of potential control companies for the six years' treatment period we look at. For the years 1992, 1993 and 1997, however, the treatment groups are small and may not be well suited for stand-alone analyses.

The data for the *Network program* is not as complete as the other data sets. We have an exact identification of the records before and after the program but unable to establish the year of participation. Thus, we do not know whether a company participated in the Network program over the entire period 1992 to 1996 or if it was involved in the Network program for a shorter time only, at the minimum one year. This is a serious problem in the sense that we run the risk of comparing a company that did not participate in a particular year with another company that did participate in that year, the only difference being that the first company either has participated previously or will be participating later on in the period between 1992 and 1996. Provided that the Network program generates a positive, additive payoff, this lack of exact mapping of the year of treatment should be expected to produce an upward bias in the estimates. Thus overestimating the differences between the treatment group and the control group is a more likely outcome than underestimating the difference.

Table 3 Participants & controls for the Network Program

Year		Treatment		Control
t-1	***t***	*t-4*	***t***	*t-1*
1991	1992	1575		44895
1995	1996		1636	
N =	Total	*1575*		44895

The risk of a too optimistic estimate is however, based on the premise that the Network program gives a positive payoff. If, on the other hand, companies that were in the treatment group at one or other time between 1992 and 1996 *performed better when not participating in the program* within the same period, we have committed the error of incorrectly attributing effects to the program when this is not the case. Table 3 shows more participants by the end of the period than in the beginning, but we have no clear notion of the turnover within the period, i.e. the number of participants entering or leaving the program between 1992 and 1996. Low turnover within the period would make our estimates more reliable and high turnover would do the opposite. We believe this should give a slight reservation only with respect the trustworthiness of our results. Better knowledge of the turnover between 1992 and 1996 would make it easier to judge probable impacts.

5.1.1 The Regional Dimension: Participation and Allocation of Funds

The data available for the two financial schemes are fairly complete. As indicated in Table 4, approximately 4.3 billion NOK is employed by the Government for Regional Venture Capital Loans and Investment Grants over the period 1990 to 1994. Although the total sums allocated to Regional Ventura Capital Loans and Investment Grants are comparable in magnitude, the two financial schemes involve different governmental costs. Whereas Investment Grants are directed towards preferably innovative or other interesting investments this are, as the name implies, grants, and as such motivated by their presumed payback to society in term of increased economic activity. Regional Ventura Capital Loans are loans with slightly higher interests than conventional loans. The loans are meant for projects that may carry risks which exclude conventional loans and thus, entail a calculated risk of

loss. In the past, losses have been about 30%, based on data from the 1970s and the 1980s.

A closer look at the regional allocation of funds reveals that Northern Norway has been prioritized with respect to allocation of funds, in particular in the earlier part of the period under scrutiny. Moving down the first column in Table 4 from Oslo to Svalbard implies going from the urban areas to the more rural areas of Northern Norway. Inspection of the sums allocated county by county tells us that there is an increasing tendency as we move towards the less densely populated areas with a marked peak in *Nordland*. This tendency is apparent both for Regional Venture Capital Loans and for Investment Grants. A striking feature of Table 4 is the central areas of Norway, in particular Oslo, and Østfold has close to zero allocations while other urban areas such as Western Norway (Hordaland, Sogn og-Fjordane and Møre og Romsdal) receive substantial allocations. Clearly, Western Norway contains both major cities and large rural areas which may partly account for this difference. Thus, there is a distinct regional profile in both Regional Ventura Capital Loans and Investment Grants with respect to the allocation of funds.

We are, however, not able to identify acceptable administrative records for all 3298 companies, mainly due to incomplete records and because we wanted to restrict our analyses to limited business companies[5]. Thus, as should be expected, we have to accept analyses based on less than the entire set of available administrative records for the financial schemes.

[5] Not that this is a limitation due to missing information in the administrative records from Dun & Bradstreet, it is not primarily imposed by shortcomings in the databases received from Innovation Norway.

Table 4 Funds Employed for the Period 1990 to 1994, by County

	Regional Venture Capital Loans					Investment Grants				
Fylke/Year	**1990**	**1991**	**1992**	**1993**	**1994**	**1990**	**1991**	**1992**	**1993**	**1994**
Østfold	5,1	0,0	2,3	1,7	0,0	0,0	0,0	0,3	0,0	0,0
Oslo	0,0	0,0	0,0	0,0	0,0	0,0	2,4	0,0	0,0	0,0
Hedmark	37,1	18,7	42,1	10,1	33,9	10,2	12,1	11,8	7,6	48,0
Oppland	25,4	39,2	30,3	14,5	6,5	16,4	10,4	12,5	16,2	28,0
Buskerud	9,6	4,1	11,5	1,6	5,5	3,3	1,1	3,6	1,0	12,9
Telemark	18,0	2,0	13,5	6,2	2,6	17,2	11,4	9,8	19,1	31,5
Aust-Agder	7,6	17,8	18,1	6,5	1,1	2,7	1,4	3,6	8,5	2,2
Vest-Agder	10,5	18,4	11,6	3,6	0,4	2,6	5,5	4,5	9,8	1,3
Rogaland	21,5	16,4	13,8	2,5	2,2	3,8	2,8	4,8	2,7	1,5
Hordaland	26,6	25,2	19,2	15,4	6,8	10,0	10,1	14,9	12,8	17,8
Sogn og-Fjordane	49,8	50,9	61,7	19,8	47,7	42,1	27,4	71,1	21,5	35,9
Møre og Romsdal	78,5	47,8	69,4	24,5	32,2	25,9	47,2	27,6	69,4	49,7
Sør-Trøndelag	25,5	25,6	25,9	48,0	6,2	21,2	18,0	17,4	14,9	18,7
Nord-Trøndelag	35,5	15,4	46,4	16,7	8,0	24,8	24,1	24,0	19,0	25,0
Nordland	101,3	102,0	90,4	35,5	29,7	265,4	113,7	95,7	111,7	128,2
Troms	32,6	20,2	16,0	6,5	5,9	49,4	48,1	44,9	72,5	56,9
Finnmark	68,5	35,3	75,3	16,7	2,9	63,8	60,0	71,1	46,6	47,8
Svalbard	0,0	0,2	0,0	0,0	0,0	0,0	1,1	1,0	8,7	0,5
Sum	553,3	439,1	547,6	229,8	191,6	558,9	396,7	418,5	442,0	505,9
Total = 4 283,4	Total Lending Funds = 1 961,4 mill. *Expected losses on loans is 30%*					Total Investment Funds =2 322,0 mill.				

As shown in Table 5, the overall coverage in terms of what proportion of allocated resources we are able to analyze is 31.6%[6]. Clearly, the proportions we are able to analyze vary somewhat across regions. If we ignore Oslo (which can safely be done since almost no allocations exist) the minimum coverage is 18.5% for Investment Grants to Northern Norway and the maximum coverage is 46.4% for both kinds of allotments (both Venture Capital Loans and Investment Grants) to Agder and Rogaland.

Clearly, since conclusion concerning the two financial schemes is supposed to be valid for the two schemes as nationwide initiatives it is of importance that the data coverage embraces all regions. Table 5 indicates a coverage that justifies nationwide conclusions.

[6] These figures are based on reasoning that is explained in more detail later.

Table 5 Coverage of Allocations, Total & Proportion of Funds Analyzed

	Total Information available - Mill. Nok			Information used for analyses - Mill. Nok			Percentage used for analyses - Percent		
Region	Regional Venture Capital Loans	Investment Grants	Both Kinds	Regional Venture Capital Loans	Investment Grants	Both Kinds	Regional Venture Capital Loans	Investment Grants	Both Kinds
Oslo and Akershus	0.0	2.4	0.0	0.0	0.0	0.0	0.0 %	0.0 %	0.0 %
Hedmark and Oppland	96.7	77.6	256.8	26.3	24.1	74.8	27.2 %	31.1 %	29.1 %
South Eastern Norway	52.3	88.8	58.6	15.0	28.5	21.8	28.6 %	32.1 %	37.3 %
Agder and Rogaland	96.4	36.2	74.5	27.1	8.0	34.5	28.1 %	22.2 %	*46.4 %*
Western Norway	247.7	314.2	497.2	70.2	99.9	184.6	28.4 %	31.8 %	37.1 %
Trøndelag	115.3	106.6	238.4	24.0	31.3	98.3	20.8 %	29.4 %	41.2 %
Northern Norway	239.0	896.8	790.1	105.8	165.5	314.7	44.3 %	*18.5 %*	39.8 %
Total	847.4	1522.5	1915.6	268.4	357.3	728.8	31.7 %	23.5 %	38.0 %
	Sum total = 4 285.5			Sum analyzed = 1 354.5			Percent analyzed =31.6%		

It is, however, reasonable to question whether adequate within-region coverage is sufficient if certain regions account for disproportional amounts of the allocations analyzed. Table 6 displays a simple comparison of the percentage regional distribution of allocated funds based upon the total sums in Table 5 (column 1 to 3) and the corresponding percentages for the analyzed sample (column 4 to 6 in Table 5). The difference in percentages (column 7 to 9 in Table 6) indicates mostly insignificant discrepancies, except for Regional Venture Capital Loans and Investment Grants in Northern Norway where the former is relatively underrepresented and the latter overrepresented. Noteworthy, balanced coverage with respect to the distribution allocated funds is easily altered when a few or only one large allotment is left out. Furthermore, the filtering out of deviant cases is an inevitable consequence of procedures that are used to preprocess the data for analyses. That is, since all analyses are based on comparisons of matched pairs of firms, companies with exceptional features e.g. very large turnover, tend to be left out.

Table 6 Coverage of Allocations, Deviations from Percentage Distributions

	Total Information available - Percent			Information used for analyses -Percent			Difference between total and used		
Region	Regional Venture Capital Loans	Investment Grants	Both Kinds	Regional Venture Capital Loans	Investment Grants	Both Kinds	Regional Venture Capital Loans	Investment Grants	Both Kinds
Oslo and Akershus	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Hedmark and Oppland	11.4	5.1	13.4	9.8	6.7	10.3	1.6	-1.6	3.1
South Eastern Norway	6.2	5.8	3.1	5.6	8.0	3.0	0.6	-2.1	0.1
Agder and Rogaland	11.4	2.4	3.9	10.1	2.3	4.7	1.3	0.1	-0.8
Western Norway	29.2	20.6	26.0	26.2	27.9	25.3	3.1	*-7.3*	0.6
Trøndelag	13.6	7.0	12.4	8.9	8.8	13.5	4.7	-1.8	-1.0
Northern Norway	28.2	58.9	41.2	39.4	46.3	43.2	*-11.2*	*12.6*	-1.9
Sum percent	100.0	100.0	100.0	100.0	100.0	100.0	-0.1 %	-0.1 %	0.0 %

Noticeably, although we find coverage with respect to the funds allocated to be satisfactory, i.e. the regional distribution laid out by Innovation Norway is not altered by the fact that we analyze less than the total number firms that actually have received grants or attractive loans, this imposes a limitation. The limitation that has to be kept in mind is that the reduction from population to sample *is not due to a random sampling procedure*. Thus, it is still necessary to corroborate that the firms selected out do not deviate substantially from those analyzed. Table 5 and Table 6 suggest that the differences between the analyzed sample and the population are insignificant. The remaining part of the data, the companies left unanalyzed, may still be different from the selected sample. These issues will be addressed in chapter 14.

The regional distribution of *participation* (firms, cases subject to treatment) does, of course, not entirely overlap the regional distribution of allocations of grants and loans. Also, the regional distribution of money is relevant for the Regional Venture Capital Loans and Investment Grants only. For the FRAM program and the Network program the distribution of funds is not an issue since the regional dimension was never an integral part of the intended policies behind these programs. Com-

parisons of the financial schemes and programs reveal these differences. While the financial schemes show distinct regional profiles, the FRAM program and the Network program appear to have participants haphazardly scattered over the country.

Table 7 Distribution of Participation – All financial schemes and programs

Region	Regional Venture Capital Loans	Investment Grants	Grants & Venture Loans Combined	*Total Grants & Loans*	The FRAM Program	The Network Program
Oslo and Akershus	2,1	3,4	0,8	2,2	8,9	17,0
Hedmark and Oppland	11,8	6,9	7,8	8,4	11,3	7,9
South Eastern Norway	5,9	6,9	3,5	5,5	*22,8*	17,9
Agder and Rogaland	10,0	4,8	4,3	5,9	15,2	14,7
Western Norway	*42,6*	21,2	29,1	28,9	20,0	*19,4*
Trøndelag	9,7	12,5	11,9	11,6	8,9	12,3
Northern Norway	17,9	*44,3*	*42,6*	37,5	12,9	10,8
Sum percent	100,0	100,0	100,0	100,0	100,0	100,0
Total N	340	609	486	1435	425	1575

The differences in regional profiles clearly reflect the intention of the initiatives. Regional Venture Capital Loans, as the name implies, have an explicitly given obligation to serve those regions where access to capital is expected to be limited. The Investment Grants follow much of the same pattern mostly since the two instruments often are employed jointly. The FRAM program (the part we analyze) aims at small and medium sized business while the Network program is intended to stimulate and foster industrial clusters. Thus, the regional dimension is not an integral part of the intentions behind these programs. The regional distributions simply reflect and confirm the intended policies of the initiatives.

5.1.2 Changes over time in the use of the Financial Schemes

Regional Ventura Capital Loans and Investment Grants are often combined. As shown in Table 8, the joint use of the two instruments amounts to close to 50% of all

allocations in 1990 while the combination of the two is down to 33.8 percent in 1994. Table 8 also reveals a trend away from Regional Ventura Capital Loans and towards more use of Investment Grants over the period.

Table 8 Venture Capital Loans & Invest. Grants - Percent of Allocations

	1990	**1991**	**1992**	**1993**	**1994**
Regional Venture Capital Loans	21.70	20.30	28.90	14.20	8.90
Invesment Grants	29.60	24.00	25.70	51.20	57.30
Jointly Both Kinds of Allotments	48.70	55.80	45.40	34.60	33.80
Total percent	100 %	100 %	100 %	100 %	100 %

The change from the use of Regional Ventura Capital Loans to the use of Investment Grants is even more evident when we look at percent approvals. As shown in Table 9, while the two instruments *and* the combination of the two all shared approximately a third of all corroborated decisions in 1990, Investment Grants alone stood for almost eighty percent of all approvals in 1994.

Table 9 Venture Capital Loans & Invest. Grants - Percent of Approvals

	1990	**1991**	**1992**	**1993**	**1994**	*Total*
Regional Venture Capital Loans	34.54	26.86	27.74	12.67	7.01	969
Invesment Grants	30.44	36.12	39.54	62.53	78.44	2030
Jointly Both Kinds of Allotments	35.02	37.02	32.73	24.80	14.55	1254
Total percent	100 %	100 %	100 %	100 %	100 %	
Total N	1025	886	822	750	770	4253

Note: N includes all firms involved over the five year period

5.1.3 Initial Exclusion of potential confounders

The varying numbers of potential controls in the three datasets (Table 1 to Table 3) reflect both the exclusion of records with missing data and the routine of excluding

records that fall outside the convex hull i.e. the polygon bounded by the extreme points of our matching variables (King & Zeng, 2007). The *chull* routine of the *R* package was pragmatically used for the purpose of reducing the number of cases. Because the matching algorithm applied (**nnmatch**) (Abadie, Drukker, Herr, & Imbens, 2001) is slow, this was a necessary data-preparing step[7]. More important, we had to exclude companies that had participated in other programs or received other governmental subsidies or allotments. Thus, although we do not have the complete number of participants for all three programs under investigation, we do have a substantial part of the participants in each scheme and program and we have succeeded in excluding companies involved in other projects and programs. Thus, to the extent it is possible; records that are potential confounders are excluded, both from the lists of participants in the schemes/programs we investigate and from being included in the control groups. However, as indicated by the (**X**) in Figure 2, participants in one of the schemes/programs that also participate in another of these programs were not excluded. These are few. Inclusion of cases that participate in more than one program does not affect estimation results and facilitates the examination of potentially additive effects of participating in several programs. The size of the circles in Figure 2 does not reflect the number of firms. While the number of treatment cases vary from 425 to 1575, the number of potential control cases remaining after removing possible confounders is more than sufficient for our purposes. After exclusion of cases due to missing data, being outside the convex hull and other reasons, we still have large amounts of data, as shown in Table 1 to Table 3. Data for participants in other programs than those under investigation here were supplied by Innovation Norway. We believe these data provide the best available dataset for the exclusion

[7] You never run a matching routine only once. Generally, it may take a considerable number of runs and in this case a single run could last more than three hours due to slow convergence.

of confounders. However, it is clear that this list can be incomplete. Thus the existence of potential confounders in the control groups cannot be entirely ruled out.

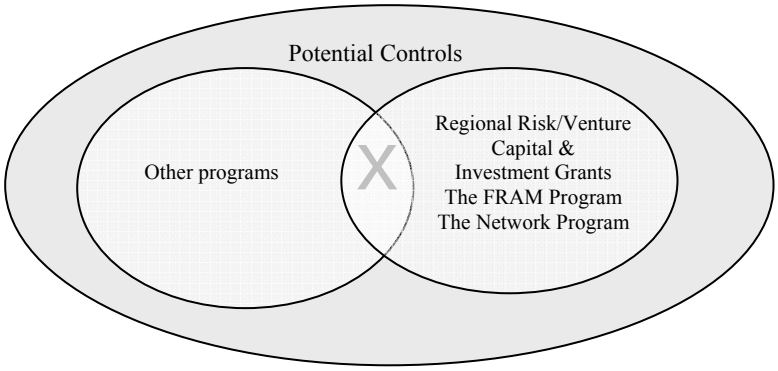


Figure 2 Data preparation – elimination of confounding controls

5.2 Measurements

5.2.1 The dependent variable

Our central measure is *added value*, defined as the sum of labor costs and net operating result. This is our dependent variable in most models. The measure can be thought of as the sum of the reward to labor and capital respectively. The two variables that constitute the construct are obtained from the financial records from Dun & Bradstreet. There are many definitions of measures intended to characterize the value creation processes of firms (Kay, 1993; Kay, 1995). The most common generic term is *value added* which measures the increase in the value of goods as a result of the production process. We deliberately try to avoid this definition by switching the words and use the term *added value*. *Added value*, defined as the sum of labor costs and net operating result does not make much sense as a stand-alone term. It takes its meaning as a comparative term; when the outcome of a governmental program is to be measured as the difference between the entities included in

the program and comparable entities that are not included in the program. Using the sum of labor costs and operating result has its advantages and disadvantages. Changes in the ratio of labor costs to operating result over time may indicate equilibrium effects, which in our models, by definition, is assumed away by the SUTVA[8] (stable-unit-treatment-value-assumption). From an orthodox economics point of view equilibrium effects is at the heart of the arguments against governmental interventions. In this report the existence of such effects is an empirical question. From the society's point of view increase in labor costs due to the hiring of more workers is beneficial. Thus, a change in the ration of labor cost to operating result is not a problem. The aggregated differences in added value that can be identified as caused by the program in question are interpreted as the *additionality* that can be credited the program.

5.2.2 The matching variables

Whereas the dependent variable within the causal counterfactual framework can be any measure of interest, this is not the case for the matching variables. All matching variables have to be pre-treatment variables, which are collected before treatment assignment. Post-treatment variables should not be used to predict the treatment assignment. The inclusion of such variables may result in biased estimates. Matching variables can be related to both the treatment and the outcome. The outcome variable can *not* be included among the matching variables. Moreover, according to Bryson & al. (2002) variables known to be related to the treatment assignment but not to the outcome should *not* be included. In the propensity score setting such variables may reduce the balancing properties of the covariates (Caliendo & Kopeinig, 2005). Whether this is also the case in the covariate setting is generally not known. Note-

[8] SUTVA implies that the treatment on one particular case should not affect any other case in the analysis.

worthy, there is no complete consensus on these issues. Dependent upon whether matching is looked on as an estimation technique or mainly a preprocessing tool, opinions may differ. The ultimate goal of the matching procedures is to arrive at a situation where participation, given a set of conditioning variables *X*, is independent of potential outcomes (see CIA, Conditional Independence Assumption, section 6.3 page 61) and thus facilitates the ceteris paribus i.e. assuming that full unit homogeneity is achieved we can estimated the treatment effect as $\Delta_i = Y_{1i} - Y_{0i}$ for unit i. Furthermore, given that the outcome equations $Y_{1i} = f(X_i) + \varepsilon_{1i}$ and $Y_{0i} = f(X_i) + \varepsilon_{0i}$ where f is one or other kind of function of *X*, can be matched in a way that gives $E[(\varepsilon_{1i} - \varepsilon_{0i}) \mid X] = 0$ where i.e. ε_{1i} and ε_{0i} are i.i.d. with zero conditional means, then the average program effect is $\Delta_{ATE} = E(\Delta_i)$ and the program effect for any subgroup S $\Delta_S = e[\Delta_i \mid i \in S]$ of a sufficient sized can be estimated as the difference in outcomes between participants and non-participants. We do not assume that we can ever achieve $\varepsilon_{1i} - \varepsilon_{0i} = 0$ for any participant so interpretations at the unit level may be meaningless[9].

We use two continuous and three discrete variables for matching purposes. In addition we use the variable *year* for exact matches. The first variable concerns the *size of the firm*; we want to balance firms of comparable sizes. As a proxy for company size we use the total income of firms. Clearly, this variable is related to outcome; generally, both allocations and returns are larger in the bigger firms. The variable is constructed from accounting records from Dun & Bradstreet. Due to the skewness of the empirical distribution of this variable we use the logarithm of total income for

[9] Full homogeneity is of course not fully warranted since this is matching on *observables* only.

the actual calculations. Since production technology may decide the ratio of total income to the size of the workforce, we include *labor costs as a percentage of total income* as our second matching variable. This variable is more of a semi-continuous variable since it has eleven categories where 0 means that labor costs amount to between 0 and 10% of total income, 1 means from 10 to 20% and 11 means that the company uses between 90 and 100% of total income for labor costs. The third matching variable is a coarse indicator for *industry* that can take three values; one for *trade, transportation & other industries,* one for *services & real estate* and one for *manufacturing, mining & construction*. The coarse division of this variable is due to the fact the more detailed NACE-codes are not very informative with respect to the nature of the activity of the firm and also that firms may be involved in several different business areas and are thus just as well described by a gross description of their general activity. The fourth matching variable concerns the relative *newness* of the company, i.e., approximate time since the firm was established. The variable can take three values; *1* if the company is less than two years old, *2* if it is from three to five years since establishment, *3* if it is from six to nine years old and *4* if the company was established more than ten years ago. Our fifth matching variable is a *location* variable that can take seven values; *1)* Oslo and Akershus, *2)* Hedmark and Oppland, *3)* South Eastern Norway, *4)* Agder and Rogaland, *5)* Western Norway, *6)* Trøndelag and *7)* Northern Norway.

Industry, newness and location are all related to the decision about treatment assignment. Although Innovation Norway has no distinctive policy with respect to industry for the financial schemes and intervention programs under scrutiny, industry matters. Also, regional priorities are imperative political goals for Innovation

Norway's activities. Thus, as well as for location there is a deliberate policy for newness: New establishments have a certain priority since stimulating the creation of new activity is an explicit goal for Innovation Norway. Moreover, new firms entail the liability of newness (Stinchcombe, 1965) and thus are more inclined to failure which may imply a rather high probability of losing matched pairs over time while matching firms from the same locations may be a prerequisite for proper use of pairwise matches (Heckman, Ichimura, & Todd, 1998a). The simultaneous match on size, technology, industry, newness and location may ensure comparisons of firms that are commensurable with respect to expected outcomes on our dependent variable *added value.*

The sixth variable employed in the matching process is *year.* This variable is used solely for the purpose of ensuring that schemes and programs that go on over several years have correct matches. With algorithms other than the **nnmatch** (Abadie et al., 2001), this is far from a trivial problem.

Thus, as a summing up; when we assert that two companies are comparable we contend that they are close to equal along the dimensions; *company size*, *labor costs relative to company size*, *industry classification*, *newness* and *location*.

5.2.3 The difference-in-differences estimator

The dependent variable added value shows considerable volatility from year to year. The difference-in-differences estimator (Ashenfelter & Card, 1985) reduces this problem. The combination of matching techniques and difference-in-differences analysis constitutes our central strategy for bias reduction. Manual correction of out-

liers in the dependent variable is carried out by removing the upper and lower 1% of the empirical distribution, dependent on the extent of extreme values. All estimators introduced are differences in means between various categories. The difference-in-differences estimator reduces volatility by using the difference in means before and after treatment in the treatment group and comparing this magnitude with the corresponding magnitude in the control group i.e. $DiD = (Y_A^T - Y_B^T) - (Y_A^C - Y_B^C)$ where Y is outcome, T denotes treatment, C denotes control and B and A denotes before and after respectively. As shown in Figure 3 the difference-in-differences is a simple before-after comparison of ΔT and ΔC which implies that we measure changes in the two groups in a way that allows them to evolve independently over time.

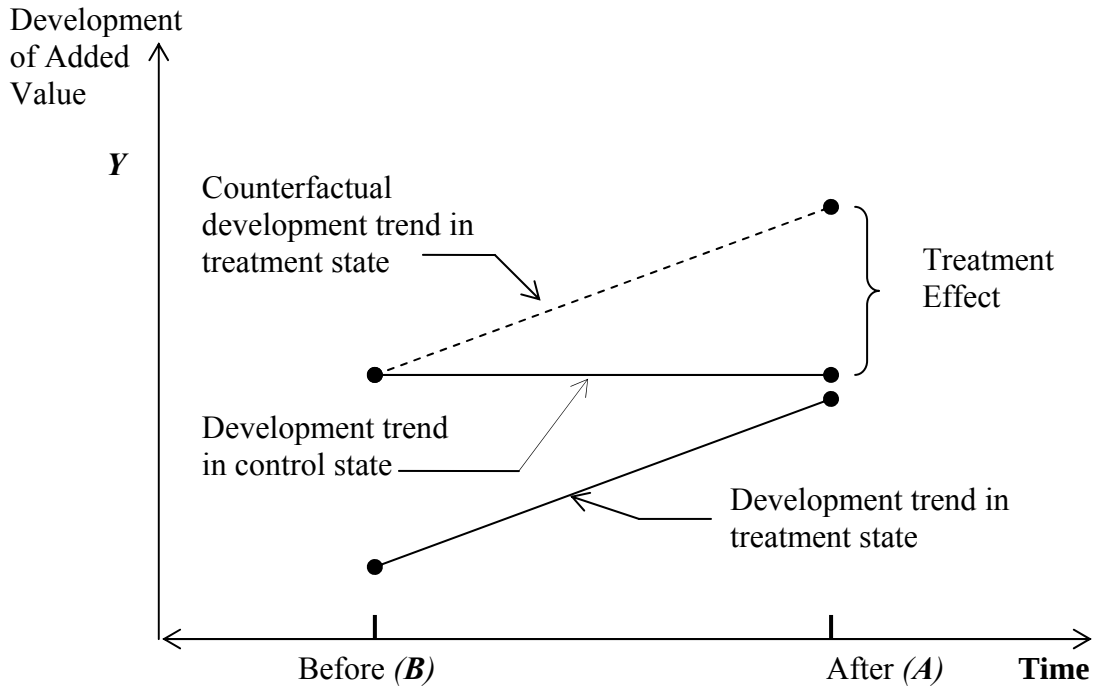


Figure 3 Causal Effect in the Difference-in-differences model

Clearly, the DiD estimator by itself does not signify that differences in changes between the two groups can be attributed to the treatment in question. The DiD estimator assumes that that the two groups follow approximately the same path over

time which indicates that a *ceteris paribus* clause can be defended. A quick glance at Figure 4 indicates a common increasing trend in both the treatment group and the control group for all three sets of intervention data used for matching.

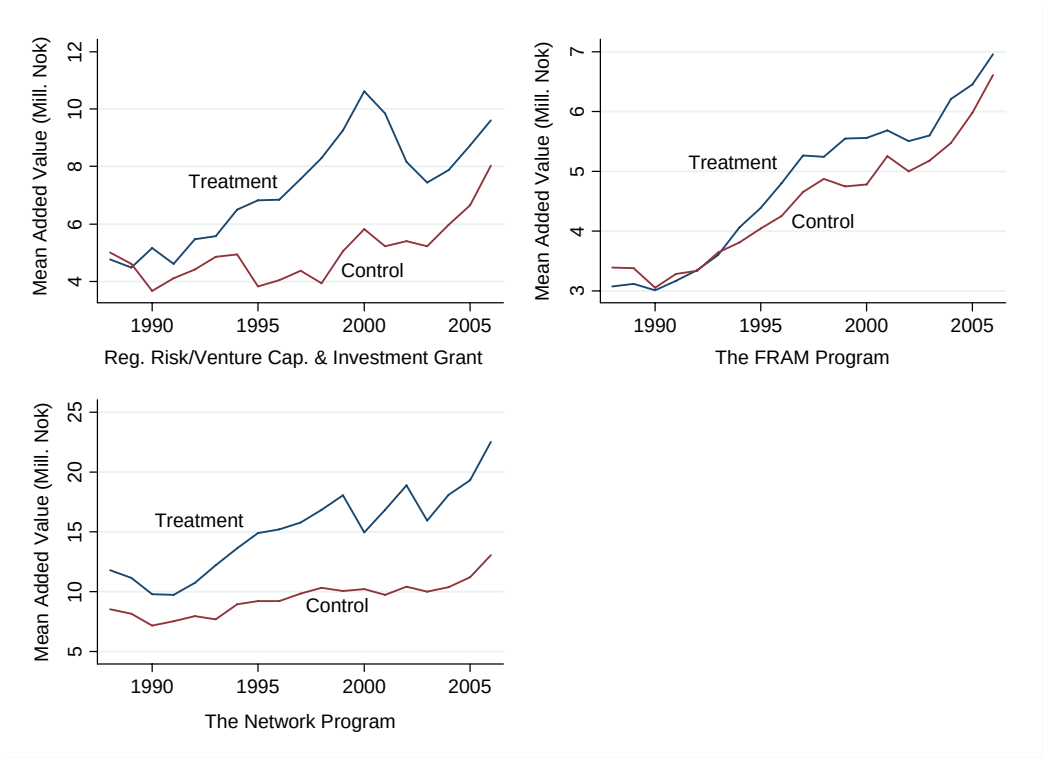


Figure 4 Trends in the development of Added Value over time – all cases

There is an inherent ambiguity in the common trends assumption since the key identifying assumption is that the trends should be the same in both states (control and treatment) *in the absence* of treatment. As pointed out earlier, the reason for applying a matched sample strategy is the fact the no single unit can be observed in two states, as treated and as controls. Thus, after match we have to regard the degree of common trend by roughly counting in what looks like the effect of treatment. The graph for Regional Venture Capital Loans & Investment Grant indicates a peak around the millennium followed by a sharp drop in the treatment group. The Network Program graph shows an atypical toothed pattern in the treatment group right after the year 2000. All three graphs show the treatment group above the control

group over most of the period thus indicating that a selection effect is present. Under the common trends assumption the DiD estimator in combination with matching potentially remove or substantially reduce the effect of selection into treatment.

6. Methods

6.1 A brief overview of the section

In this section we introduce perspectives on matching methods. The way we look at matching methods has bearings upon how it is applied. We maintain that matching is best applied as a preprocessing method i.e. a set of procedures that prepare the data for further analyses. Matching is not a single method; it comes in a multitude of variants with options that sum to a cacophony of possible choices. We therefore take the most widely applied approach, *propensity score matching* (PSM) (Rosenbaum & Rubin, 1983; Rubin, 1997), as the point of departure. The frequent application of PSM has crystallized a step-by-step standard that provides a helpful tool for matching procedures. We argue that, due to specific features of the data at hand, PSM is not the optimal choice and that covariate matching (CVM) (Abadie & Imbens, 2002) is a better choice. Also, since we have a large pool of potential controls, one-to-one matching is preferable because it makes the datasets easier to handle and further analyses less complicated.

6.2 Matching as nonparametric preprocessing

We think of matching as a *nonparametric preprocessing method* (Ho et al., 2007) rather than as an estimation method. The preprocessing prepares the data for further analysis in the sense that it facilitates the construction of a dataset that resembles *some* features of the classical experiment. Viewed as a preprocessing method, matching offers a way of adjusting for as much of the information in the control variables as possible *without making any parametric assumptions.* In the preprocessed data, the treatment variable is closer to being independent of the

background covariates, hence, subsequent parametric adjustments are less important. Moreover, the researcher is free to go on with the analyses using his/her favorite parametric methods. Even with a task as simple as comparing two averages, that is exactly what we do in this paper: we make distributional assumptions in order justify the use of sampling statistics to qualify our comparisons. Also, to be able to incorporate the time dimension, we make further distributional assumptions.

Another important aspect of the preprocessing view of matching methods is that it makes analyses less *model dependent.* Most applied social science quantitative research does not reach their ideal causal model in just one run, as assumed by statistical theory. The sequence: – collect data – decide model – run program – evaluate model – write up findings, never happens. In practical research, numerous, may be hundreds of runs are carried out with different control variables and models in order to find the model that the researcher finds acceptable, most likely in accordance with his/her theory. Hence, estimates depend heavily on their corresponding modeling assumption. A minor adjustment of the model may yield a very different estimate. Most causal models in the social sciences are therefore model dependent, at least to some degree (Ho et al., 2007). The statistical properties of the model do, however, depend on the assumption that *we know the single correct model.* Thus, any effort that can reduce model dependency is advantageous.

The preprocessing view of matching also implies that the notion of causality is untied from the statistical model applied. We consider causal effect as a theoretical quantity, *defined independently of any empirical method* that has been used to estimate it from the data.

There is little contradiction between this view of matching procedures and the point of view that matching is an estimation technique and that the estimate that is the outcome of the matching procedure is the ultimate goal. Most likely, the impression that the latter view exists and is held by many researchers is a product of the considerable theoretical effort aimed at refining the matching techniques. Since such work implies investigating the properties of *estimators* it may appear as it is the *estimates* that is under investigation and hence, this is an estimator technique and as such a causal model. In practice, the nature of the data available may have a bearing upon the question of causality. In this paper we want, after preprocessing, to exploit the data at hand to the extent necessary for the problem we are investigating.

6.3 The General Framework of Propensity Scores

The parameter of interest for this evaluation is the effect of treatment on the treated, in the literature usually labeled *ATT*. In recent years (PSM) propensity score methods (Rosenbaum et al., 1983; Rubin, 1973a) have dominated the econometric evaluation literature and we will use PSM as an exemplar for the step-by-step procedures that are more or less generic to all matching procedures. Propensity score is defined as the conditional probability of being exposed to treatment. The probability of participation in the treatment group is *modeled* by careful selection of matching covariates and can in principle be estimated by many different methods; the by far most usual is the logistic regression of the form:

(1) $p(X) \equiv \Pr\{D = 1 \mid X) = E\{D \mid X\}$ where $D = \{0,1\}$ is the indicator for participation, (1 for treatment, 0 for control), X is a vector of relevant covariates, i.e., covari-

ates believed to influence the probability of participation but not the outcome. Thus, the probability that a unit is exposed to the treatment in question is the propensity score. As demonstrated by (Rosenbaum et al., 1983), provided that exposure to treatment is randomly distributed within the categories defined by the multidimensional ***X***, then exposure to treatment is randomly distributed over the one-dimensional *p(X)*, the propensity score.

With the individual unit denoted *i*, ATT can be estimated as:

(2) $\tau \equiv E\{Y1_i - Y0i \mid D_i = 1\}$

$= E\{E\{Y1_i - Y0_i \mid D_i = 1, p(X_i)\}\}$

$= E\{E\{Y1_i \mid D_i = 1, p(X_i)\} - E(Y0_i \mid D_i = 0, p(X_i)\} \mid D_i = 1\}$

The expectation applies to the distribution of propensity scores for *the treatment group* $(p(X_i)|D_i = 1)$, usually referred to as the *region of common support* where *Y0* and *Y1* are possible outcomes for the effect-variable for treatment and control respectively.

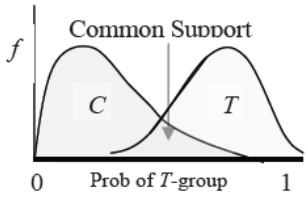


Figure 5 Propensity scores and the region of common support

These results require that:

(3) $D \perp X \mid p(X)$ i.e., that assignment to treatment is independent of *X*, conditional on *p(X)* and that

(4) $Y1, Y0 \perp D \mid X$ i.e., the outcomes on the effect-variables are independent of the assignment to treatment or control, and that:

(5) $Y1, Y0 \perp D \mid p(X)$ i.e., that the outcome on the effect-variables are independent of the assignment to treatment or control, given equal propensity scores.

Provided that the conditions (1) to (5) are satisfied, observations with approximately equal propensity scores will have similar distributions of observables (***X***) independent of assignment to treatment or control. The requirements (3) to (5), often called *CIA*, the *conditional independence assumptions* are central and ensure the independence of control variables and treatment assignment and of control variables and the outcome variable under investigation. Condition (2), the *region of common support* restricts the range of the propensity scores in the control group to the range of the score within the treatment group. This is a source of bias since cases that are outside the region of common support have to be deleted. Moreover, treatment cases and control cases may not be evenly distributed over the common range interval, causing an excess number of cases over some parts of the interval and a shortage over other ranges, which in sum may cause bad matches over the region of common support. A variety of techniques and matching algorithms are, however, developed to minimize or control the effects of such problems. If a substantial proportion of the treatment cases are outside the region of common support it is of course important to verify whether these deleted cases are different from the included ones.

An important additional condition, the *stable unit-treatment assumptions* (*SUTVA)* is considerably harder to control for. This condition demands that the effect of treatment on a unit *i* should be independent of the effects of treatment on any other units. This assumption implies no equilibrium effects.

6.4 Matching strategies and data features

The data at our disposal does not conform to what is ideal for the most commonly used techniques for the analysis of observational data. First and foremost, we do not have programs that take place at one point in time and thus makes it easy to identify the exact before and-after intervention periods. On the contrary, interventions occur at several points in time in the sense that a company may be subject to benefits at several points in time and in some (rare) cases, may enter the same program repeatedly. For some years the treatment groups are too small to be well suited for analyses. The number of treatment cases for the FRAM program is only 18 and 19 for the years 1991 and 1992 respectively (Table 2). In the case of the Network program we have no exact information concerning when participants first became engaged in the program. We have data that tells us exactly what firms participated in the program between 1992 and 1996 but we cannot identify participants who left or re-entered the program within this period. Moreover, in a few instances firms enter programs in the year following the year of founding and thus, for these firms we have no information about the year prior to intervention; the central piece of information for the matching procedures.

In the case of the Investment Grant and Venture Capital Loans programs the two programs are interwoven in the sense that about a third of the firms in our data participate in *both* programs. In the matching procedures we therefore consider Regional Venture Capital Loans, Investment Grants and the combinations of the two as *three subsets* within the matches of *one* dataset. Thus we have *two* financial schemes and *two* programs, but *three* datasets to analyze. Also, participants may enter programs at various times over a six-year period and may enter one of the

programs prior to entering both programs and also participate and receive benefits over more than one year. No participants enter one of the programs after a period of being involved in both programs, but some participants may re-enter one of the programs after a year or more of absence from the program. Thus, *the before program year*, which carries the central information for the matching procedure, the information about the pre-program status for our matching variables and our outcome variables is *a moving window* that ensures that the pre-program year for the treatment cases matches the pre-program year for the control cases. For all three datasets we can identify firms that left the market due to bankruptcy or other reasons for closure. In the treatment group, less than ten percent of the cases in all three datasets are lost over the observation period 1989 to 2006. Since we have a large pool of potential controls, control cases that are lost before 2006 are deleted from all three datasets,

The complexity of the three datasets and their peculiarities calls for matching strategies that attend to both their irregularities and what they have in common. We want to have the best possible estimate of the effects of interventions for all three programs and we also want to evaluate effects on an equal footing that makes comparisons possible. This implies compromises that may sacrifice exactness for comparability. The use of matching as a nonparametric preprocessing method that invites further analysis (Ho et al., 2007) has some consequences for how matching is carried out, noticeably a preference for *one-to-one exact matching.*

One-to-one exact matching is a feasible option for all three datasets because we have a large pool of potential controls. Thus, analyses that go beyond the simple difference between the means in the treatment and control group, notably analyses that

attend to the advantages of having longitudinal data, are facilitated. The question of comparability, i.e. to what extent we can compare the outcomes of the schemes and programs in question is a different one: We can argue that outcomes are comparable because *a)* all entities under scrutiny are from the same population *b)* the majority of firms are on the average within the same range in terms of size, although firms in the FRAM- program may be a bit below average compared to the other two interventions (most of them are small and medium-sized enterprises, SME) *c)* we use the same covariates for matching for all three programs.

We follow the principle that the *features of the data decide our choice of procedures* as long as this line of proceeding does not compromise the applicability of methods.

These overarching considerations place heavy demands on the implementation of operative procedures for matching. Clearly, when the goal of the procedures is to identify a non-treated firm that is as similar as possible to the treated firm, the solution is to search in the direction of the "twin" firm. Of course, identical "twin" firms cannot be found. Moreover, administrative data provides a limited number of firm traits and we have decided that five traits, *firm size*, the logarithm of total turnover, *technology* understood as the ration of employees to turnover, a coarse categorization of *industrial sector* based upon NACE codes, degree of *newness*, understood as the time elapsed since establishment, and geographical *localization* (region) provide a sufficient and feasible characterization of the firm. Thus, the achievable goal of the matching procedure is, for each firm exposed to intervention; to find the non-exposed firm that most closely resembles the exposed firm, along the five above

mentioned dimensions taken together. This requires a matching procedure able to account for all the peculiarities of our three datasets.

The structure of the data for Investment Grants and Regional Venture Capital Loans and the FRAM program has in common that interventions take place at various points in time over a limited time span. As illustrated in Figure 6, this implies that we can identify the before and after periods. It does, however, imply a couple of less obvious problems:

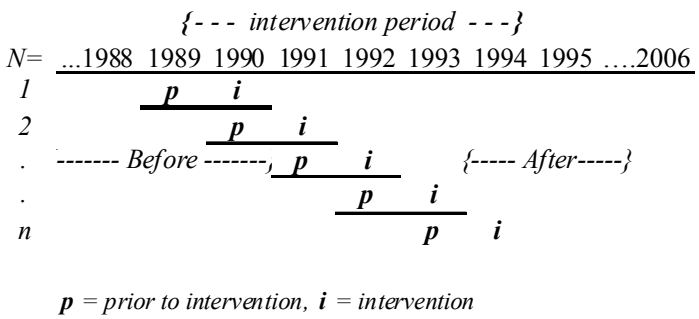


Figure 6 The timing of interventions

Provided that a match for case n_i in the first occurring intervention is identified, that case has to be reserved for all succeeding time periods in order to preserve the possibility for one-to-one exact match for the consecutive time periods. Furthermore, any case that has found its match has to be protected from being confused with cases from both preceding and succeeding periods. This situation is different from the multiple treatment case (Lechner, 2001) where a multinomial logit model can be applied. Clearly, *within* the intervention period we have no clear distinction between *before* and *after* intervention without an *explicit inclusion of the time variable*. This feature is important to our choice of matching procedure.

6.5 Matching methods – guidelines and procedures

The choice of matching algorithm is something that is usually discussed *within a propensity score framework* (section 6.3, page 61). The great divide in the matching literature is, however, between propensity score matching (PSM) and covariate matching (CVM) or matching on ***X*** in Heckman's (1998a) terminology. Until recently, only a few applied papers based upon covariate matching have been available. Thus, guidelines and procedures for covariate matching are rare. On the other hand, guiding principles and methods for propensity score matching (Caliendo et al., 2005) are now so well developed that they provide good instructions and checklists for covariate matching as well. Thus, we follow the assumptions in section 6.3 and use propensity score matching (PSM) as the standard for comparisons of alternative procedures.

6.5.1 Matching in one dimension – the propensity score

With PSM as reference, the algorithm decision is a question about how to best match two vectors of probabilities ranging from zero to one. Alternatively, we can use a number derived from this probability (e.g. $p'=ln((p/(1-p))$) in order to improve the working of the matching algorithm employed. The baseline technique for PSM matching is the nearest neighbor method algorithm, which is available in many variants. Since PSM operates by reducing the dimensionality problem from many dimensions to *one* dimension, the problem is usually that there are more cases close to one in the treatment group than in the control group. Correspondingly, there are (usually) more cases close to zero in the control group. Since the endpoints zero and one are not included in the range of probabilities, cases near the extremes should be deleted (Abadie et al., 2002). Clearly, the matching of propensity scores close to zero may not contribute much to the balancing properties of the covariates since the

substantial meaning of this is to compare two units that both have a low probability of being included in the treatment group.

As implied in section 6.3, the widespread use of propensity score matching (PSM) has via practice crystallized an almost standardized six-step procedure which starts with the choice between PSM and CVM (covariate matching) and ends with sensitivity analysis. Since PSM almost invariably is the preferred method, CVM is mentioned only as a part of the initial choice in Caliendo's (2005) excellent outline of the PSM procedure. Figure 7 also implicitly indicates that cross-sectional data is assumed. This most basic implementation assumes an identifiable binary treatment i.e., a treatment regime where it is clear-cut to decide who received treatment and who did not. It also assumes that treatment takes place at one point in time. Step 2, the estimation of propensity scores can be carried out in many ways. The most commonly applied is the binary logit model. Step 3 involves the choice of matching algorithm; an issue that depends heavily on sample issues such as the extent of common support region (Figure 8, page 71). Thus, step 4 may have consequences for step 3 in the sense that the amount of overlap between the propensity score for the treatment group and control group, the common support, may determine what matching strategy is best suited. Thus, with a control group that is large relative to the treatment group, one-to-one matching may be the preferred action while a relatively undersized control group may call for a one-to-many match and even matching *with replacements* of the comparison units. One-to-one matching permits unit weights in further analyses while one-to-many assumes one or other kind of weighting regime to account for the relative influence of the various comparison units, usually in the form of weights based on the distance to match, adjusted to sum

to unity. One-to-many matching may reduce bias, but weights may be cumbersome to handle in further analysis. Replacement of comparison units will also reduce bias and is recommended when the control group is small. There is, however, a trade-off between bias and precision: Compared to matching *without* replacement, replacement will generally lower the bias and increase the variance (Abadie et al., 2002).

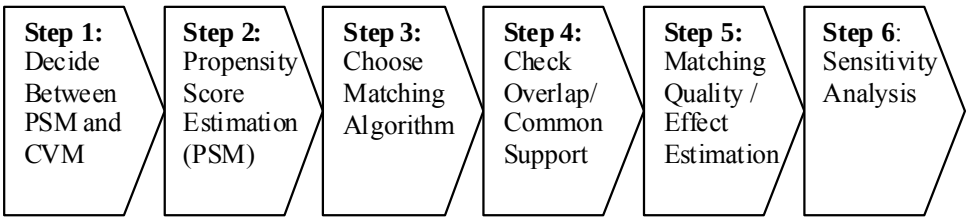


Figure 7 The stepwise procedures for practical propensity score matching[10]

The worst case for matching is the situation where a substantial number of treatment cases are outside the region of common support (to left in Figure 8). For this situation there is no proper cure. Another problem is different distributions of propensity scores in the treatment and control groups *within* the region of common support. One remedy for this is the so called caliper-matching which substitutes the notion of the closest match for an idea of an approximately evenly distributed distance between treatment and control cases over the region of common support. This is usually achieved by use a "caliper" radius around a point instead of using the nearest neighbor principle thus accepting a fixed difference in propensity score.

[10] The figure is modified from Caliendo, M., & Kopeinig, S. 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. Bonn: Forschunginstitut zur Zukunft der Arbeit. pp 2.

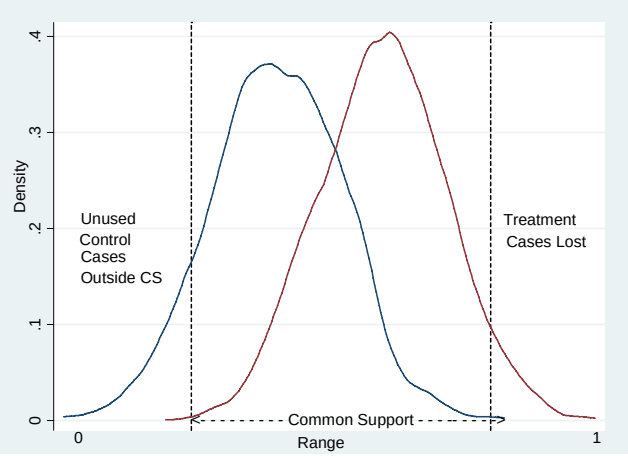


Figure 8 The Region of Common Support and Loss of Cases (Hypothetical)

Caliper matching is implemented in some matching programs and can be a reasonable compromise when data deviate from what is ideal, in particular when the number of propensity scores is scarce in the mid-range of the region of common support whereas cases cluster in the upper and lower part of the region. There are few, if any, rules or research based knowledge concerning the width of the caliper. Even though it is frequently applied, few scholars manage to come up with a credible justification for their choice of bandwidth.

A different strategy is to break the treatment sample up into intervals and estimate treatment effect separately in each region. Rosenbaum and Rubin (2006) have demonstrated that even a few groups reduces bias substantially. A more sophisticated strategy is to use some type of distribution to account for the distance between propensity scores. Kernel estimators, e.g. Heckman, Ichimura and Todd (1997) put one or other distribution (e.g. normal) around each treatment unit and weight closer control units more heavily than farther control units. Ongoing research (Sekhon & Diamond, 2005) implies that the list of matching methods will be extended.

6.5.2 Matching in many dimensions – Covariate Matching

We have decided to make use of covariate matching (CVM) as the basis for the construction of control groups. So why not use the well established procedures of propensity score matching? Our reasons for the decision to use CVM are based both on practical and theoretical grounds. We will first discuss the practical reasons.

An algorithm that has an opening for exact matches for at least one variable, the time indicator, is convenient. Without such an option, matching becomes very laborious and the likelihood of errors may increase. To demonstrate the problem we can return to Table 1, page 41, where we have altogether 1435 pre-treatment cases for the treatment variables and 425 cases that are registered in 1989 (*t-1*) which means that the decision concerning allotment was done in 1990. Regardless of the chosen matching method, these 425 cases have five pre-treatment covariates that have to be matched to the corresponding control group pre-treatment covariates also registered in 1989, not to covariates belonging to any other year. Thus, we can split the dataset into five datasets, one for each year, and estimate propensity scores separately for each set and then execute the matching procedure. Because almost all cases will exist over all five periods we then have to exclude all matched control cases from the remaining data before we estimate propensity scores and match the data for the next period (year-dataset). A problem with such a procedure is that it will produce different results dependent upon which year we take as our starting point for the procedure, and we would not know the optimal sequence of years until all ordered permutations (5!/(5-5)! = 120) are attempted.

Another problem is that, for some years, the treatment groups are small. As shown in Table 2, page 41, in the FRAM dataset we have only 18 and 19 treatment cases for the years 1992 and 1993 respectively. This may cause a problem with empty cells in those comparisons cases that would otherwise give the best matches. For the Network program (Table 3) the problem is that we do not know the exact years when treatment took place (what years companies participated in the projects). Thus, common features of the datasets calls for procedure that makes it possible to have exact match for at least one variable, namely years, while simultaneously matching the other variables according to one or other measure of minimum distance between treatment and control cases conditional upon the chosen covariates. Because matching is predominantly looked upon as a cross-sectional method, such options are not included in the most commonly used programs for propensity score matching such as Leuven and Sianesi's (2003) excellent Stata programs **psmatch2**, and **pscore** (Becker & Ichino, 2002), Sekhon's (2005) versatile ***R***+ program **GenMatch**, and the ***R***+ programs **MatchIt** (Ho et al., 2007) and **CEM** (Stefano M. Iacus, Gary King, & Porro, 2008). To our knowledge the only known program that has this option is the Stata[11] ado-program **nnmatch** (Abadie et al., 2001) for covariate matching. The feature of the data is a weighty reason, but convenience is not the only reason for using the program **nnmatch**. We have categorical covariates such as *region*, *newness* and *industry* that are not well suited for logistic regression unless they are recoded as dummy variables.

On the theoretical side, as we discuss later, some properties of covariate matching may be preferable to the comparable properties of propensity score matching e.g. efficiency (Frölich, 2007a). Clearly, covariate matching is more cumbersome to ap-

[11] StatCorp ©1984-2009

ply and the six steps that guide discussions of propensity score matching (Figure 7, page 70) cannot be fully followed.

6.5.3 Covariate matching vs. Propensity Score Matching

With many covariates, especially continuous ones, matching on covariates runs into the *curse of dimensionality*, the problem that propensity score matching so elegantly circumvents (Rosenbaum et al., 1983). Also, while matching on propensity score uses the absolute difference between the score in the treatment group and the control group as the basis for matching observations, covariate matching relies of other metrics for the judgments concerning differences between observations in the treatment and control groups. This implies that a clear-cut notion of the *region of common support* is lacking for covariate matching. Clearly, this is an instance where we cannot entirely follow Caliendo's (2005) excellent six-step procedure. While propensity matching (usually) relies upon a fine-grained number between 0 and 1 (excluding the endpoints) covariate matching relies upon a metric that is the outcome of matrices of the form $(\boldsymbol{X}_{D=1} - \boldsymbol{X}_{D=0})\boldsymbol{W}(\boldsymbol{X}_{D=1} - \boldsymbol{X}_{D=0})^T$ where $\boldsymbol{W}$ is a weight matrix and $\boldsymbol{X}$ is matrices of covariates. The program **nnmatch** offers several alternatives for $\boldsymbol{W}$; the *Euclidean* metric, the *Mahalanobis* metric and two variants of what we can call the "*Abadie-Imbens* weight matrix". The Euclidean weight matrix is the identity matrix while the Mahalalobnis weight matrix is the inverse of the variance-covariance matrix of $\boldsymbol{X}$. The Abadie-Imbens weight matrix is a diagonal matrix with the inverses of the variances of $\boldsymbol{X}$ as its elements. The result from the program **nnmatch** can be either ATE, the average treatment effect, the ATC the average treatment effect on the control group, or ATT, the average treatment effect on the treated. The program has several useful options that will not be discussed here. We consider only the

ATT and the bias correction option for the ATT that is based on the regression adjustment suggested by Rubin (1973b). The nnmatch algorithms minimize biases by allowing for more than one control case to match to each treatment cases. The program works in this manner regardless of whether the option for one-to one match is selected or not. The bias adjustment usually adds more matching cases to achieve the optimum result[12]. Despite many differences, the basic template of the propensity score procedure should be kept in mind in the discussion of results.

[12] Abadie et al. (2002) show that, besides retrial, this is the only way to reach a less biased estimate.

7. Preprocessing

7.1 A brief overview of the section

This section outlines the strategies applied for the preprocessing. The central goal of the preprocessing is to establish for later use the initial datasets and their respective control groups. We introduce additional assumptions required for the use of the difference-in-indifferences estimator and the basic approaches to the construction of new datasets. The pragmatic strategies introduced reflect both the limitations imposed by the complexity of the data at hand and our preference for taking advantage of the longitudinal structure of the data. The section presents the construction of match pairs and preliminary results for the intervention periods. These preliminary results should be considered as nothing more than an integral part of the prepossessing stage. They constitute the point of departure for necessary adjustment and corrections for the construction of datasets made up of one-to-one matched pairs of treatment cases and control cases.

7.2 The Construction of Pairwise Control

In line with the perspective that we use matching as a nonparametric preprocessing method we will present various analyses we believe can account for the effects of the interventions in question. All suggested analyses are based upon the *difference-in-differences* (DiD) estimator and are solely concerned with estimates of the average treatment effect on the treated (ATT).

The first step in our estimation strategy is to construct the control groups. For each treatment case we seek to identify the closest possible non-treatment case in order to construct three datasets; one for Regional Venture Capital Loans and Investment Grants and the combination of the two, hereafter called *the schemes dataset*, one for the FRAM program and one for the Network program. Each dataset is made up of the most similar pairs of treatment and control cases, i.e. similar with respect to the five matching variables measured in the year prior to treatment. Thus, we construct three datasets made up of such pairs for all years within the intervention periods. Note that we construct only one *schemes dataset* and consider the three variants of the financial schemes as *subgroups* within this dataset.

The upper part of Figure 9, the *pre-intervention* and the *intervention period*, shows the general structure of the *schemes* dataset. The corresponding datasets for the FRAM program and the Network program have a similar structure but different intervention periods. Note that the *post-intervention period* is merged to the datasets after cleaning and corrections of the datasets for the pre-intervention and the intervention period. This is easily done since every treatment case has a unique identifier (*1…i*) which after one-to one matching has found its pair which also has its unique (*1…n*) identifier. Perfect match means that any $X^{T}_{i,j,year}$ equals $X^{C}_{n,m,year}$. The unique *ij*-combinations facilitate the construction of a new unique identification number to represent matched pairs. The one-to-one match structure of the data matrix provides flexibility that permits rearrangements that facilitates many kinds of longitudinal analyses.

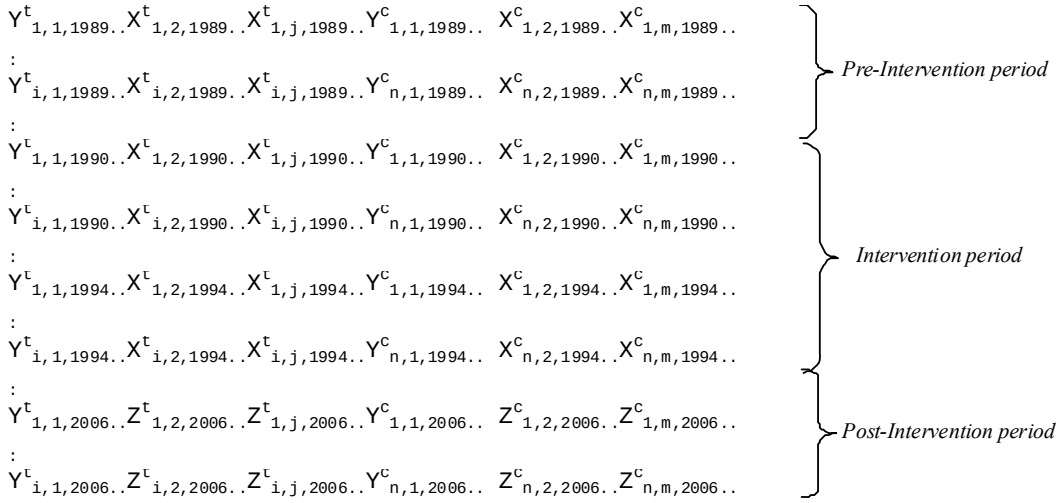


Figure 9 General Structure of the matched files

Since matches are based upon the values of the *X*s in the year *prior to intervention* (see Figure 6, page 67) it important that we do not lose cases over the analyses period. Missing cases may generate unbalanced panels that may affect estimates.

7.3 Difference-in-differences for Matched Data

In this section we present the initial results from the matching procedure based **nnmatch**. The purpose of this section is to explain the rationale for the adjustments of the initial matches and to introduce the initial *average over the intervention period estimators* for the treatment groups. The estimators are based upon the casewise after-before difference $(Y_{i,after} - Y_{i,\ before})$ constructed for all cases *before* running **nnmatch**. After program execution this difference becomes the difference-in-differences. Note that the algorithms of the ado-program are constructed so that the program can perform matching *without* entering any outcome variable. The advantages of running the program with the outcome variable included is that we obtain the standard errors suggested by Abadie & al. (2002) which is otherwise not easily calculated. Also, we can report their bias-adjusted estimates. This section is

also a prelude to two estimates we ultimately intend to report, the *two-period estimates* and the *panel estimates*.

The two-period estimates we intend to report are the difference-in-differences based upon matched data for the consecutive years after the first year after treatment. The panel structure of the data i.e. that we can follow the individual firm over time, facilitates these analyses. Thus the estimates we report are the traditional difference-in-differences from year to year based on the pairwise matched data. We are aware that these analyses are vulnerable to bias due to influence by unmeasured factors. These issues will be addressed by means of sensitivity analyses. The purpose of these analyses is primarily to get an idea of the shape of the non-linear development over time of the presumed effects (see Figure 1). The two period estimates will also be used in order to get an idea of how the magnitude of the cumulative effects evolves over the time-span observed after treatment.

The panel data estimates are more robust to hidden bias than the two-period estimated and make better use of the information in the data but give us an only overall estimate of the presumed effect over the time period subject to analysis. These analyses are based upon solutions suggested by Imbens and Woolridge (2007) concerning difference-in-differences estimation for panel data.

7.3.1 Additional Assumptions for DiD estimation for Matched Pairs

Traditionally, *the natural experiment* was the setting for difference-in-differences estimation. Challenging assumptions such as that the same process should generate the observed outcomes of the target variables in both the treatment group and the control group and, that the generated outcomes follow approximately the same pat-

tern over time in both groups, was central (Ashenfelter, 1978) to the analyses of the natural experiment. Problems like selection bias (Heckman, 1976, 1979), well known to be of importance in natural experiments was, however, less emphasized. In general, difference-in-differences estimation is a technique developed for settings that do not involve matching and where the before – after- differences in the treatment and control groups are estimated independently. More recently, (Dorsett, 2005; Eichler & Lechner, 2002) the combination of DiD estimation and matching techniques has become more common, often with an explicit reference to, among other things, the sample selection problem. Usually, the assumptions for difference-in-differences estimation in analyses that do not involve matching (Ashenfelter, 1978; 2005) are invoked for matching based analysis as well.

Noteworthy, matching involves a range of assumptions concerning the independence between treatment assignments and the variables used for matching, and the independence between treatment assignments and the outcomes on the effect-variable.
The assumptions needed to use difference-in-differences after matching is that (**6**) $Y_{ot} - Y_{ot'} \perp D|X$ and $0 < prob(D = 1|X) < 1$, where *t'* means the period before treatment and *t* means the period after treatment. Note that this concerns only the potential non-treatment outcome and that this in fact is a weaker assumption than those listed as assumptions (**4**) and (**5**) in section 6.3. Since we construct our panel by following the same units over time we must assume that the before-after differences in the outcome variable in the control group are independent of the treatment assignment, conditional upon our observed covariates used for the matching procedure and, that no control cases are systematically excluded in ways that affect this before-after difference. Clearly, a breach of this assumption is possible; it could for instance

occur as a result of a systematic dropping of cases over the time span *after* intervention. If so, this would imply a violation of the standard assumption of MCA[13] (Little & Rubin, 1987; 1989) which is a prerequisite for deleting missing cases. Throughout the analyses of the matched pair datasets, MCA is assumed and if any case is missing in the control group, the corresponding (pair) case is also dropped in the treatment group. Thus, although the panel data analyzes are based upon unbalanced panels[14], equal number of treatment and control cases is maintained over time for all pairs over the entire time span used. Other violations of assumption (**6**) would most likely also imply infringement of the SUTVA[15] assumption which says that the potential outcome on one unit should be unaffected by the assignment of treatment to any other unit (Cox, 1958). SUTVA goes beyond the concept of independence but is, according to Rubin (1991) not needed for defining causal effects.

7.4 Initial estimates and the adjustments of matched pairs

The average effect of treatment on the treated, ATT, is measured as the mean value of the difference between treatments and controls the year before the intervention periods contrasted with the corresponding difference the first year after the end of the intervention periods:

$$\Delta_{ATT} = (\tfrac{1}{t}\sum_{i=1}^{t} Y_i^{After} - \tfrac{1}{t}\sum_{i=1}^{t} Y_i^{Before}) - (\tfrac{1}{c}\sum_{n=1}^{c} Y_n^{After} - \tfrac{1}{c}\sum_{n=1}^{c} Y_n^{Before})$$

Where t is the number of observations in the treatment group and c is the number of observations in the control group. Clearly, ATT, as expressed above, can be calculated without any matching procedure. The result would most likely contain a large

[13] MCA = Missing Completely at Random
[14] Panel datasets with unequal number of cases over the time-span under examination e.g. due to missing data are called unbalanced panels
[15] SUTVA = Stable Unit Treatment Value Assumption

amount of unobserved heterogeneity since we would not know a lot about the observations (firms) we compare. After matching, however *(t = c)* and every *(i = n)* along a set of specified criteria, in our case, five carefully chosen observable characteristics, our five matching covariates. Dependent on the goodness of match, we now have difference-in-differences for any selected subgroup and any such group would have an equal number of treatment cases and control cases. Moreover, we obtain averages controlling for our five matching covariates.

Clearly if we estimate three ATTs, one for the financial schemes, one for the FRAM program and one for the Network program, these three simple measures of difference-in-differences for the groups that received treatments also add in the effects of being in the treatment group in any year prior to the last year before the end of the intervention period. For the Venture Capital Loans & Investment Grant dataset such effects will be present for the period 1991 to 1994, for the FRAM program, such effects will apply to the period 1993 to 1996 and for the Network program the period 1994 to 1996 will be affected. For a case that receive treatment only in the last year before the end of an intervention period that lasts, say 5 years, this implies that this case is viewed as receiving treatment in four consecutive years despite the fact that it did not. The direction of bias in such cases is most likely downwards, towards underestimating the true effect of intervention. If, on the other hand, a case receives treatment in the first year of the intervention period and is included as a treatment case in the remaining four years, the direction of bias is less clear. If effects of intervention are instantaneous and then fade off, extending the length of the intervention period may produce downward biased estimates. Effects that are additive over the years may work in the opposite direction. In any case, any extension of the in-

tervention period beyond a single point in time is, from a statistical point of view, problematic. From a more pragmatic point of view we do not expect the various lengths of the intervention periods to be a problem in this setting since the estimator obtained is not our most central measure of effects; it should rather be regarded as an estimator that is motivated by the matching procedure, i.e., an estimator that facilitates simultaneous calculation of the best matches for control cases over the entire intervention periods. As such, the estimator should be interpreted as the average over the intervention periods difference-in-differences effects.

The estimates presented in this section are the direct (which only means that estimates are not bias adjusted) outcomes of the program **nnmatch** (Abadie et al., 2001) based on the difference in added value before and after the treatment calculated separately for the treatment group and the control group *before* running the program. Thus, the estimates are difference-in-differences. We present the direct and the bias-adjusted estimates of the average treatment effects on the treated (ATT). All the estimates are based upon the uncorrected datasets; which means that all duplicate cases produced by the program are kept and there are no corrections for outliers. Thus, the peculiarities of the data are handled by the program and no manually adjustments are made. This implies that *more* cases than the number of existing treatment cases are used for constructing the best possible matching pairs. We estimate treatment effects (intervention effects) using the program option for one-to-one[16] matching. One of the side effects of using this option is that the difference between the direct estimates and the bias-adjusted estimates is less than if we allow for more than one control case to be used as matches for each treatment case. It turns out,

[16] One-to-one matching should not be confused with *exact matching* although the concepts occasionally are used interchangeably. Exact matching should be reserved for what it literally expresses, a casewise perfect match.

however, that the program **nnmatch** in most cases does not converge[17] to the one-to-one solution. Thus, one-to-one matching does, in most cases, imply that a minimum number of duplicate cases are used in the matching procedure.

For the dataset for Regional Venture Capital Loans and Investment Grant we present the initial results for the entire group only. In later analyses we will split this data set into three subgroups, (1) Regional Venture Capital Loans (2) Investment Grand (3) the group that received both kinds of allotments. The initial results presented below are for all three categories lumped together.

7.4.1 Preliminary Estimates for the Financial Schemes

Table 10 shows the estimates for the average treatment effect on the treated (ATT) over the entire intervention period as estimated by the algorithms introduced by Abadie et al. (2001). The bias corrected estimate suggests that the mean of the difference-in-differences between the treatment group and the matched control group between 1989 (the year before intervention) and 1995 (the first year after the period of interventions) is on the average 2.264 million NOK per company that received treatment. Thus, after choosing between 134920 minus 1435 potential controls, 1453 cases have been selected as acceptable controls and yield an estimate above 2.2 mill. NOK, which is our initial estimate of the difference-in-differences (DiD) between 1989 and 1995. Thus, the joint effects of the allocation of funds to limited companies in the form of Regional Venture Capital Loans and Investment Grant are, over the intervention period, at the average, about 2.2 mill NOK per company. The

[17] Convergence is in general a substantial problem for **nnmatch**.

two estimates are both significantly different from zero with relatively narrow confidence intervals ranging from 1.2 mill NOK to about 3.2 mill NOK.

Table 10 Preliminary ATT estimates for the Financial Schemes

Coef.	Std. err.	Z	P>|z|	[95% Conf.	Interval]	
2223.41	513.941	4.33	0.00	1216.10	3230.71	*Direct estimate*
2264.12	515.934	4.39	0.00	1252.91	3275.33	*Bias Adjusted*
Total number of matched pairs			= ***1453***			
Total number of treatment cases			= *1435*			
Total number of potential control cases			= *133485*			
Total N			= *134920*			

Note that both estimates are very similar and that the bias adjustment procedure results in a slightly higher estimate and a small increase in the standard errors of the estimate. Note also that 1453 matched pairs means that 18 treatment cases are added during execution, i.e. there must be duplicate cases in the treatment group.

7.4.2 Preliminary Esimates for the FRAM program

Table 11 shows the average treatment effect on the treated (ATT). The effect of the FRAM program is at the average about 170,000 NOK per company. The bias adjusted estimate is slightly higher than the direct estimate and the standard errors of the estimate is somewhat increased. The confidence interval contains zero, thus indicating that the estimated effects are not significantly different from zero.

Table 11 Preliminary ATT - estimates for the FRAM Program

Coef.	Std. err.	Z	P>|z|	[95% Conf.	Interval]	
169.27	164.551	1.03	0.304	-153.24	491.79	*Direct estimate*
172.55	164.469	1.05	0.294	-149.80	494.91	*Bias Adjusted*
Total number of matched pairs			= ***530***			
Total number of treatment cases			= *525*			
Total number of potential control cases			=*105788*			
Total N			=*106213*			

Thus, we have positive, insignificant outcomes and five cases are added during execution; hence, we know that we have duplicate cases in the treatment group.

7.4.3 Preliminary Estimates for the Network program

The Network program shows very strong effects. As shown in Table 12 the average treatment effect on the treated is close to 5 mill. NOK per company. The estimates are both highly significant with relatively narrow confidence intervals at the 95% level. The table shows that the bias-adjusted estimate exceeds the non-adjusted while the change in the standard error is only minor. The increase in estimates after regression adjustment is reassuring since the adjusted estimate is supposed to be the more accurate one. Even though we know that these estimates may reflect the additive effects over the five-year period the dataset covers and thus clearly overstate the true effects, the results are surprising, in particular since previous evaluations indicated no effects of this program. Surprising results does, however, ask for thorough further inspections. Since our analyses are based upon matches on *observed* covariates we should check for the influence of other unobserved variables.

Table 12 ATT - DiD - direct and bias adjusted estimates – Network Program

Coef.	Std. err.	Z	P>|z|	[95% Conf. Interval]		
4776.45	1188.488	4.02	0.00	2447.06	7105.84	*Direct estimate*
4813.78	1188.836	4.05	0.00	2483.70	7143.85	*Bias Adjusted*
Total number of matched pairs			=	***1582***		
Total number of treatment cases			=	*1575*		
Total number of potential control cases			=	*44895*		
Total N			=	*46470*		

We observe that 7 cases are added during execution and thus, we know that duplicate cases in the treatment group are generated during execution of the program.

7.4.4 Preliminary Estimates: A Summing Up

We decided to construct one single dataset for the financial schemes and analyze each individual scheme as a subgroup. Separate matching procedures for the Regional Venture Capital Loans and Investment Grants were also carried out. Matching results were almost identical, regardless of whether we matched the two groups separately or together. With close to 40% of the companies receiving *both* kinds of allocations the best match for all three (the third being the combination, i.e. close to 40%) groups were found by using *one* match that included all three groups. The reason for this result is probably that there are small differences between those firms that received one kind of allotment, Regional Venture Capital Loans *or* Investment Grants and those who received both. Analyses of subgroups depends on the sum of the quality of the matches for the individual cases within the subgroups only, and lumped together the mean of the distance metrics was more even distributed across the three groups and about the same size; thus we consider the one dataset strategy justified.

For all three datasets, the algorithm added cases in the treatment group. These surplus cases have to be removed in order to construct one-to-one matches. It also turned out the algorithm produced duplicate cases in the control group, even when the option for one-to-one matches was specified. The duplicate cases are of concern since the necessary deletion of cases may be a source of bias.

8. Matching quality and bias reduction

8.1 A brief overview of the section

The purpose of this section is to establish the *quality* of the outcomes of the matching procedures. This is primarily done by comparing the degree of similarity between the treatment group and the control group, measured in terms of the scores on the covariates selected for matching. The matching algorithms of the **nnmatch** program do not provide a perfect one-to-one match. The program generates duplicate cases, both in the treatment group and the control group, mostly in the control group, and these cases should be deleted. Moreover, we do not want our matched datasets to contain extreme outliers. Both outliers and duplicates have to be removed manually. Maximum bias reduction is achieved under perfect one-to-one match and deviation from perfect match implies a loss of bias reduction. Thus, the central goal of this section is the evaluation of the quality of the outcomes from the matching procedures. Furthermore, matching on observables does not fully exclude the potential influence of *unobserved* variables. We cannot entirely rule out the latent influence from unobserved confounders, but we can make judgements concerning the amount of influence they potentially can exert. Such considerations are the subject matter of the latter part of this section where we discuss the sensitivity of our measures.

8.2 Regional Venture Capital Loans & Investment Grants

With propensity score matching as the template, it is easy to see that some central points of reference are missing. With covariate matching we have no unambiguous *region of common* support (step 4, Figure 7, page 70) as with propensity score

matching. Each matched pair has *one* distance metric in common which informs us whether the two parts (the treatment case and the control case) that constitutes the pair have a good match or a less good match. Thus, although there is no such clear cut criterion as the region of common support we can distinguish between better and worse matched pairs. We have chosen to use the "Abadie - Imbens" (Abadie et al., 2001) distance metric which assumes a weight matrix which is a diagonal matrix with the inverses of the variances of $\boldsymbol{X}$, our chosen matching covariates, as its elements. This metric depends on the properties of $\boldsymbol{X}$ and has no standardized scale that provides an unambiguous criterion for what magnitude that should be considered a good match. The distance metric is however, useful in many respects. We remove the duplicate cases with the greatest distance metric and keep the best ones, those with the smallest distance metric. We use outlier analyses to remove extreme pairs.

There are many ways to evaluate the quality of the matching procedures. A common, not very recommendable practice is to look at the *t*-value for the difference between $\overline{X}_T$ and $\overline{X}_C$ before and after matching and consider the improvement. Since the control group may be, as here, very large compared to the treatment group a significant improvement may be meaningless because the significance depends on the number of cases in the larger group. A better, but not unreserved recommendable procedure is to use the *t*-statistics *after* match. Clearly, a low *t*-value would be taken to indicate a good match. The treacherous aspect here is that it may lead to the deletion of cases in order to improve match and such a procedure is self-fulfilling in the sense that it will invariably give the illusion of a better match since *t* decreases as the number of cases decreases. Clearly, deleting cases creates a new source of bias.

Another common procedure for checking overall matching quality (step 5 in Figure 7, page 70) is to look at the percent reduction in bias based on the

formula bias = $\left| \frac{100(\overline{X}_T - \overline{X}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}} \right|$ where $\overline{X}_T$ and s_T^2 are the sample mean and variance for the treatment group and $\overline{X}_C$ and s_C^2 are the comparable statistics for the control group (Rosenbaum & Rubin, 1985). For propensity score matching, this measure is usually reported for each control variable. The percent bias reduction is, however, less appropriate for discrete variables. A more obvious measure is the percent of cases correctly matched as shown in Table 13.

As noted above, another peculiarity we have to take into consideration when using **nnmatch** is that fine-tuning the matches means that more than one control case is usually matched to each treatment case. This is very inconvenient when we intend to use the matched datasets for further analysis and try to avoid cumbersome weighting procedures. Thus we delete all duplicate cases even though the loss of cases also is a potential source of bias. As shown in Figure 10, the dataset for Regional Venture Capital Loans and Investment Grants is reduced from 1453 cases to 1125 cases due to duplicates and extreme cases. In particular, cases that are extremes in terms of the matching metric are removed so the maximum distance metric is down to .25 compared to the original matched sample that had cases with distances above 10. Note that the original number of cases was 1435 while the matched sample has 1453 cases. Outlier detection is based on Hadi's Stata program for detecting extremes (Hadi, 1992; 1994).

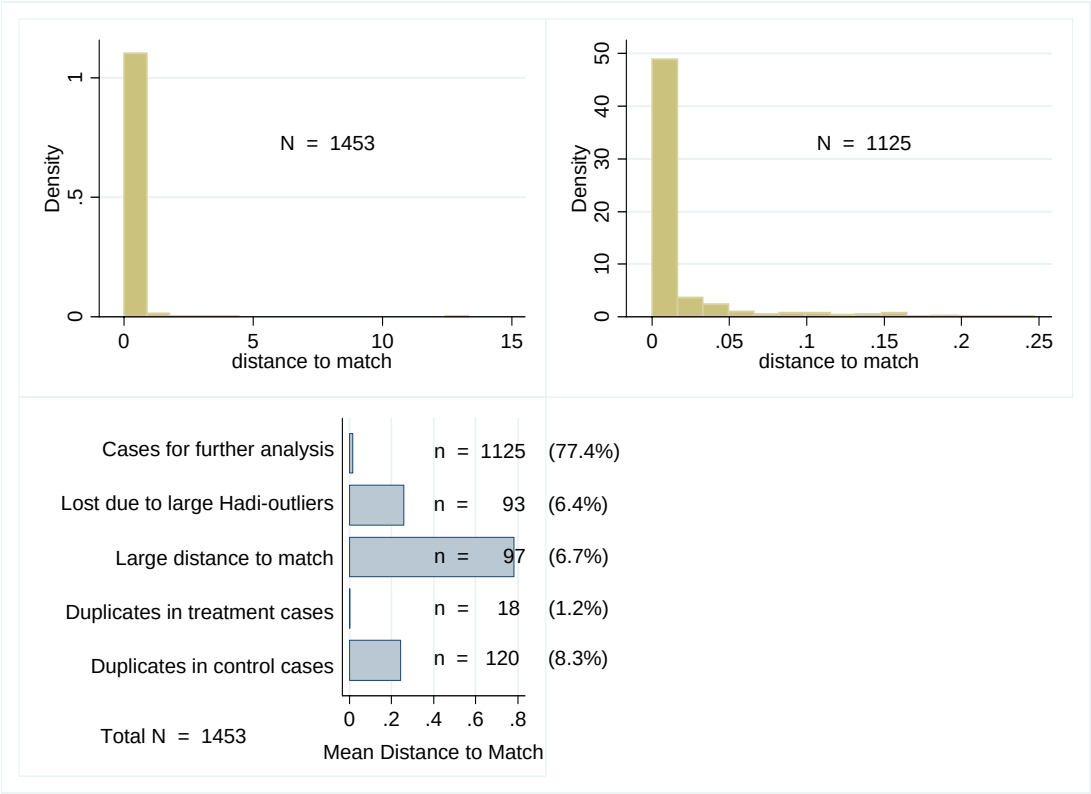


Figure 10 Outlier & duplicate correction for the Venture Capital & Investment Grants dataset – Matched pairs

Table 13 shows that the original match gave a bias reduction of more then 50% for the continuous variable and a correct match for well above 90% for all the discrete variables. The table also shows that outlier and duplicate corrections improved matches and brought all discrete variables up to a 100% correct classification.

Table 13 Matching statistics for the dataset for Venture Capital & Inv. Grant

		Cont. variable Ln Total Income - improved balance					Discrete variables - percent correct			
	Sample	**Obs.**	**Mean**	**Std. Dev.**	**Diff. C-T**	***t*-value**	**Industry**	**Region**	**Newness**	**Labor/ ratio**
Unmatched	Control	133485	8.27	1.360	-0.780	-21.57				
Data	Treatment	**1435**	9.05	1.636						
nnmatch	Control	1453	9.05	1.629	0.052	6.95	99.8 %	92.9 %	99.2 %	91.3 %
results	Treatment	1453	9.00	1.523						
Corrected	Control	1125	8.82	1.326	0.020	3.94	100.0 %	100.0 %	100.0 %	100.0 %
Match	Treatment	1125	8.80	1.293						
Percent of cases lost		*21.6 %*								
Percent bias reduction		*51.9 %*								

Even though Table 13 shows significant difference between the means for the treatment and the control group, $t = 3.94$ after corrections, this can safely be ignored as consequence of the large number of cases, the difference, 0.02 is substantially insignificant. A 21.6% loss of cases is a problem, but should not invalidate conclusions that are drawn solely on the basis of the subset of cases that are kept for further analysis.

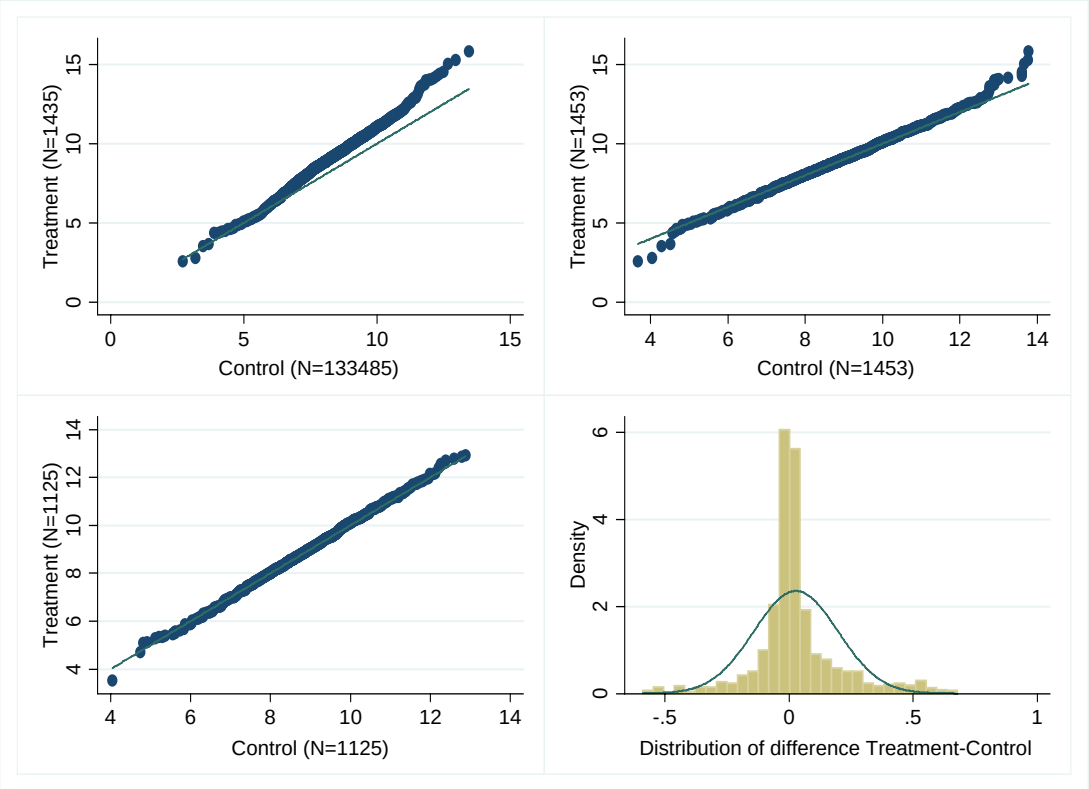


Figure 11 Q-Q plots of Total Income (log scale) and the distribution of the difference between treatment and control groups (Venture Capital & Investment Grants)

The practice of using the mean as the standard for assessing bias reduction may be deceitful since equal means may have completely different distribution. As pointed to by King (2007) a quantile-quantile plot gives a better picture of the improvement achieved by matching. As shown in Figure 11

the matching procedure has made the distribution of total income (on logarithmic scale) in the control group (upper left corner) more comparable to the distribution in the treatment group after match (upper right corner). Outlier and duplicate correction have further improved the equality of the distribution of income in the treatment in control group (lower left corner). The lower right corner of Figure 11 shows that the distribution of the difference between the two groups is approximately normal.

8.3 The FRAM Program

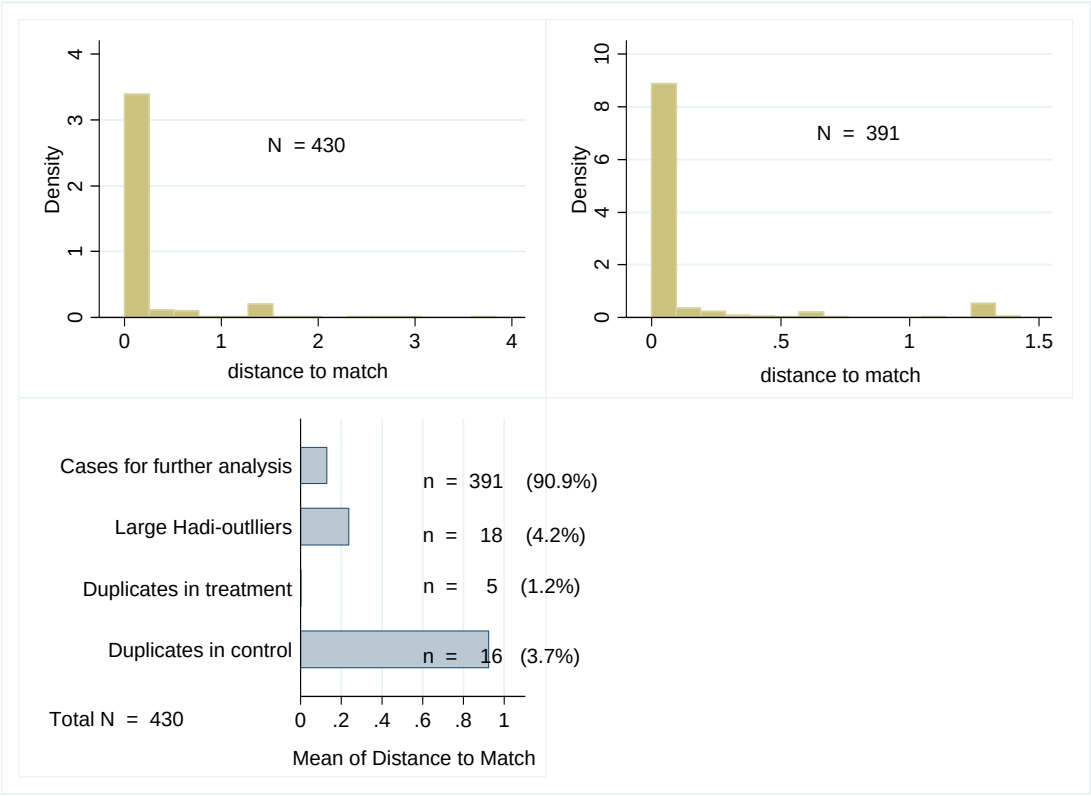


Figure 12 Outlier and duplicate correction for the FRAM program – Matched pairs

Figure 12 shows the number of cases lost due to outliers and duplicates in the control and treatment groups. The major reason for loss of cases is duplicates in the control

group due to the matching procedure. More than 90% of these are kept for further analysis, but for a few cases the magnitude of the metric matches is relatively high.

Table 14 Matching statistics for the dataset for the FRAM program

		Cont. variable Ln Total Income - improved balance					Discrete variables - percent correct			
	Sample	**Obs.**	**Mean**	**Std. Dev.**	**Diff. C-T**	***t*-value**	**Industry**	**Region**	**Newness**	**Labor/ ratio**
Unmatched	Control	105788	8.95	1.129	-0.127	-2.31				
Data	Treatment	**425**	9.08	0.850						
nnmatch	Control	430	9.10	0.821	0.010	1.06	99.8 %	92.8 %	93.7 %	95.3 %
results	Treatment	430	9.09	0.848						
Corrected	Control	391	9.06	0.823	0.001	0.08	99.8 %	94.9 %	94.4 %	97.0 %
Match	Treatment	391	9.06	0.847						
Percent of cases lost		*8.0 %*								
Percent bias reduction		*12.7 %*								

The matching statistics for the FRAM program indicate a moderate bias reduction due to the continuous variable Total Income. The percent correct classification of the discrete variables is well above 90% for all variables. This is the smallest dataset and the relatively low bias reduction and the fact that we have kept cases with a relatively large matching metric reflects a compromise between the quality of match and the loss of cases. Both low quality of match and loss of cases may cause bias. With small datasets the trade off between these two factors becomes more apparent.

As indicated by the quantile-quantile plots in Figure 13, matching causes a major improvement in the comparability of the distribution of Total Income between the treatment and the control group (upper left corner). The improvement due to the exclusion of duplicates and outliers is, however, less clear. Even though the Q-Q plot shows close to a straight line, the lower and upper regions of the plot reveal at least a few deviant cases. Given that a small dataset may require a compromise between matching quality and the loss of cases, this turns out to be the best possible adjustment so far. It does, however, signal that additional corrections may be necessary.

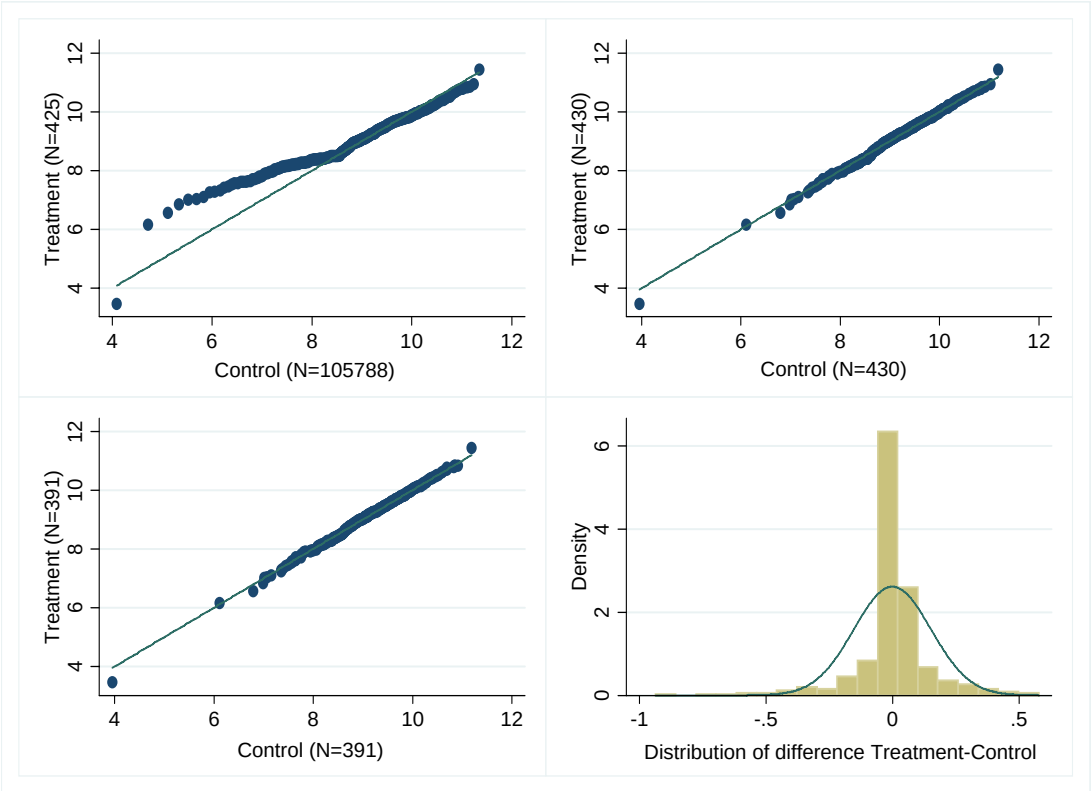


Figure 13 Q-Q plots Total Income (log scale) and the distribution of difference between treatment and control groups (FRAM)

8.4 The Network Program

The dataset for the Network program is relatively large with 1575 treatment cases that result in 1582 cases after matching. As can be seen in Figure 14 the dominant reason for loss of cases is duplicates in control cases (146) while only a few cases are excluded due to large distance metrics (24) or duplicates in treatment cases (7). Moreover, the maximum distance to match in terms of the matching metric is considerably reduced after corrections. The loss of cases after corrections is 10.8% only, leaving 1405 cases for further analysis. As shown in Table 15, the bias reduction due to the income variable is close to 90% and the percentage simultaneously correctly classified discrete variables is close to 100%.

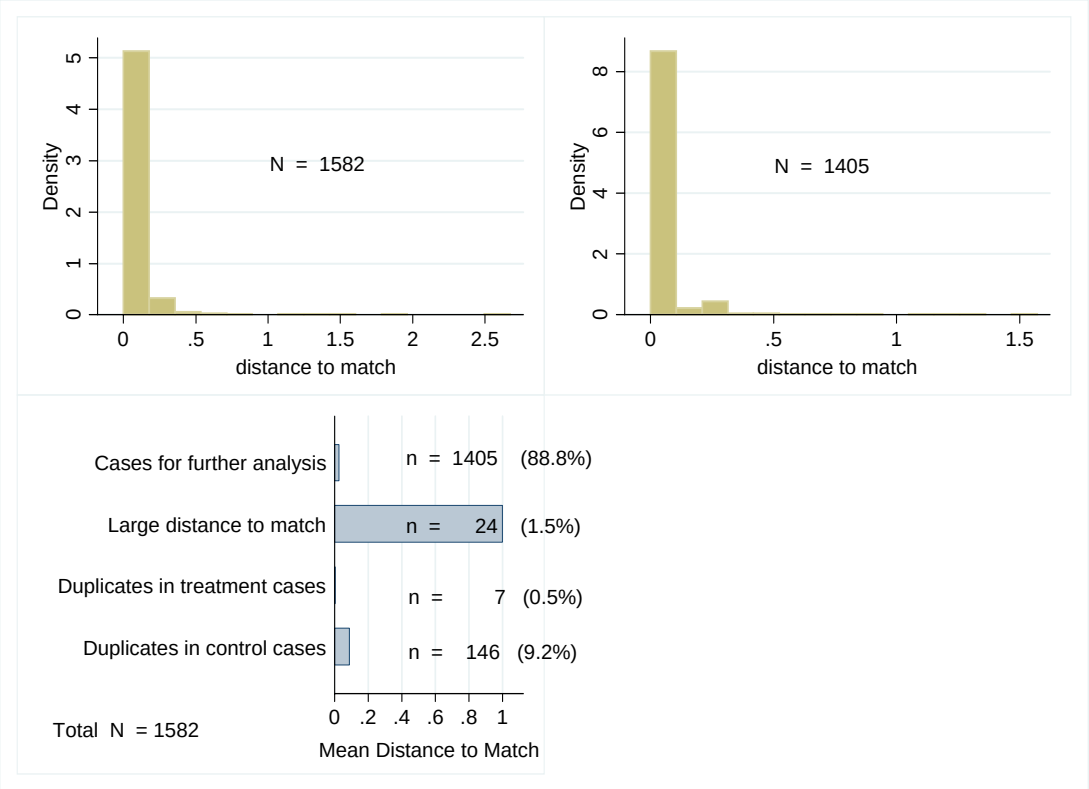


Figure 14 Outlier and duplicate correction for the Network program – Matched pairs

Table 15 Matching statistics for the dataset for the Network program

		Cont. variable Ln Total Income - improved balance					Discrete variables - percent correct			
	Sample	**Obs.**	**Mean**	**Std. Dev.**	**Diff. C-T**	***t*-value**	**Industry**	**Region**	**Newness**	**Labor/ ratio**
Unmatched	Control	44895	8.06	1.546	-1.411	-35.48				
Data	Treatment	**1575**	9.47	1.686						
nnmatch	Control	1582	9.46	1.687	0.050	8.49	99.9 %	97.3 %	100.0 %	95.1 %
results	Treatment	1582	9.41	1.620						
Corrected	Control	1405	9.26	1.585	-0.030	-6.32	99.9 %	97.9 %	100.0 %	95.7 %
Match	Treatment	1405	9.29	1.618						
Percent of cases lost		*10.8 %*								
Percent bias reduction		*87.2 %*								

Figure 15 shows that the similarity of the distribution of Total Income between the treatment group and the control group is substantially improved after matching and that the removal of duplicate cases and outliers further improves the comparability

of these distributions over treatment and control cases. Also, the distribution of differences in Total Income between the two groups appears to follow a symmetrical distribution indicating balance improvements.

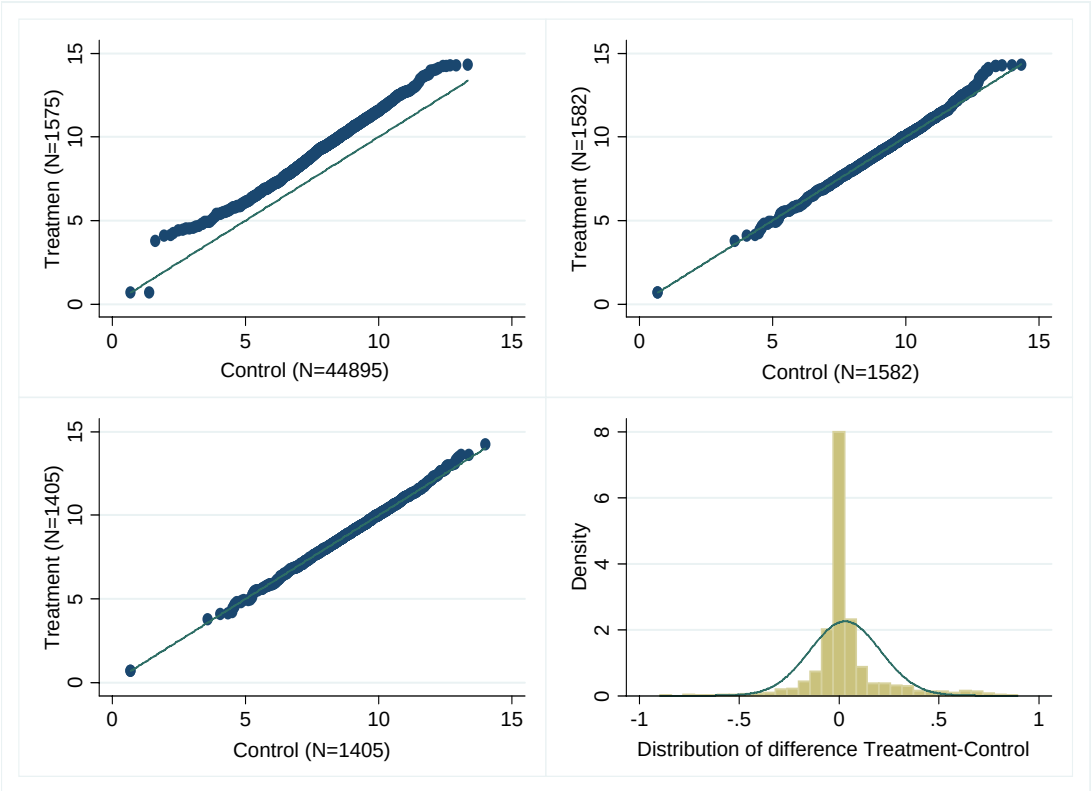


Figure 15 Q-Q plots Total Income (log scale) and the distribution of difference between treatment and control groups – The Network program

8.5 Matching quality – a summing up

With pairs of matches (one-to-one matching) we have created three datasets for further analysis. The corrections for outliers and duplicates cause a slight loss of cases but improve the overall quality of matches. Thus, although covariate tends to be cumbersome compared to propensity score matching it may be well suited as a preprocessing tool, in this case particularly since it facilitates exact matching on a discrete variable, a property so far only found in **nnmatch**.

8.6 Sensitivity analysis for matched pairs

Matching on observables implies that 1) the choice of control variables can always be questioned and 2) we have to recognize that we have limited knowledge about the influence of unobserved variables. Sensitivity analysis may provide some information concerning the potential influence of unobservables. Rosenbaum bounds (Rosenbaum, 1995) provides one way of assessing sensitivity. His arguments go briefly like this: Without randomization, statistical inference is generally not valid and we cannot permute data to form *p*-values. With correctly matched data, however, there should be no differences between the treatment group and the control group, data points should be exchangeable and inference based on permutations should be valid and the *p*-values should be valid, *provided there are no unobserved confounders*. Rosenbaum's method for sensitivity analysis relies on the sensitivity parameter Γ (gamma), the odds of receiving treatment. In a randomized experiment Γ = 1 since randomization ensures that the odds of receiving treatment is equal for all units. If, in an observational study, two units that are identical on matched covariates have a Γ that equals, say, 2 this would mean that one of the units might be twice as likely as the other to receive treatment *because they differ in terms of an unobserved covariate* (Rosenbaum, 1995). If we denote the probability of receiving treatment for case *j* as p_j, then this probability after match should solely be a function of our vector of covariates $\boldsymbol{X}$ and hence, if two units with the same values of $\boldsymbol{X}$ have different *p,* there is hidden bias. Formally, we have hidden bias if $\mathbf{x}j = \mathbf{x}_k$ and $p_j \neq p_k$ for some units *j* and *k*. The odds of receiving treatment for a unit *j* is $p_j/(1 - p_j)$, thus the *odds ratio* of receiving treatment for any two units *j* and *k* with the same values

of **x** is $(p_j/(1-p_j)/\, p_j/(1-p_j))$. If we assume that this odds ratio is bounded between $1/\Gamma$ and Γ, i.e. $\frac{1}{\Gamma} \le \frac{p_j/(1-p_j)}{p_k/(1-p_k)} \le \Gamma$ then this would imply that if Γ equals one, the odds ratios of receiving treatment would be equal for *j* and *k* and there would be no hidden bias. If Γ is greater the 1, say 3, then the odds of receiving treatment for units with the same value of **x** would differ by a factor of 3. If we think of the deviation from Γ equals one (the randomized experiment) as caused by a binary unmeasured covariate ***u*** we can write a logistic regression model that links the covariates to the odds of treatment as $\log(p_j/(1-p_j) = \mu(\mathbf{x}) + \gamma u_j$. where μ is a unknown function and γ is an unobserved parameter and u_i is constrained on $u_j \le 0 \le 1$. Provided that unit *j* and *k* have the same values on **x** i.e. $\mathbf{x}_j = \mathbf{x}_k$, this can be rewritten as $\frac{p_j/(1-p_j)}{p_k/(1-p_k)} = \exp\{\gamma(u_j - u_k)\}$ which implies that two units with the same values of **x** differ in their odds of treatment by a factor of γ and the difference in the unobserved covariates (Rosenbaum, 1995). Hence, we can think of Γ *in terms of the size of the coefficient for the unobserved covariate u* and this line of reasoning can give us an idea of the approximate size of the influence of a hypothetical unobserved variable. The process of sensitivity analysis then goes as follows: We choose values of Γ, say, from 1 to 3 and use the information in our dependent variable, which is based on matched pairs, to assess the results of randomization tests based on the Wilcoxon sign rank test and the Hodges-Lehmann point estimate for the sign rank test for increments in Γ . The program **rbounds** (Gangl, 2004) is used for estimating the Rosenbaum bounds.

Table 16 Rosenbaum bounds - Venture Capital Loans & Investm. Grant – matched pairs

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1,00	0,0000	644,0	644,0	505,5	791,8
1,05	0,0000	585,1	703,5	451,0	856,0
1,10	0,0000	532,4	761,5	401,0	918,5
1,15	0,0000	482,0	819,0	353,5	980,5
1,20	0,0000	435,5	875,0	308,5	1041,0
1,25	0,0000	392,0	930,0	265,5	1100,0
1,30	0,0000	350,5	984,5	225,0	1159,1
1,35	0,0000	310,5	1037,5	186,0	1217,5
1,40	0,0000	273,0	1090,0	148,5	1274,5
1,45	0,0001	237,0	1141,5	113,0	1331,5
1,50	0,0007	202,0	1193,0	78,0	1387,5
1,55	0,0039	169,0	1243,3	45,0	1442,8
1,60	0,0156	137,0	1292,5	12,5	1498,3
1,65	0,0477	106,0	1343,0	-19,5	1552,5
1,70	0,1158	75,5	1391,0	-50,5	1608,5
1,75	0,2300	47,0	1439,5	-80,5	1662,5
1,80	0,3843	18,5	1487,5	-109,0	1716,1
1,85	0,5549	-9,0	1535,0	-138,0	1771,1
1,90	0,7119	-36,5	1582,8	-166,5	1824,5
1,95	0,8337	-62,5	1630,5	-194,0	1878,5
2,00	0,9144	-88,5	1676,5	-221,0	1932,0

Table 16 shows the Rosenbaum bounds for the treatment effects for the unadjusted estimates in Table 10. The first column is the values for gamma which we have chosen to vary between 1 and 2. The table shows the *p* level, maximum and minimum of the Hodges-Lehmann point estimates and the upper and lower confidence interval for the estimates based on the estimated significance level. The first line indicates an approximate medium value of 644 mill. NOK when gamma equals one, the baseline, i.e. as in a randomized experiment. We can see that when gamma equals 1.65 the *p*-level approaches the usual 0.05 threshold and the estimated treatment effect may be as high as 106 or as low as 1343. At the lower bound, this is not significantly different from zero. Hence, the odds of being in the treatment group are 1.65 times higher because of different values on an unobserved covariate *u,* and, despite being matched on the same observed covariates, our inference changes. There are few guidelines for judging whether the odds ratio 1.6 is high or low. Compared to

other social science studies it is relatively high (Rosenbaum, 1995) and indicates that the influence of an unobserved variable has to be substantial to affect estimates.

Table 17 Rosenbaum bounds – The FRAM program – matched pairs

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1,00	0,0072	221,5	221,5	43,0	396,0
1,05	0,0223	179,5	261,5	5,0	433,5
1,10	0,0556	142,0	299,0	-33,0	471,0
1,15	0,1159	106,5	334,0	-68,5	508,0
1,20	0,2071	71,0	367,0	-104,0	543,5
1,25	0,3253	40,0	399,0	-140,0	579,5
1,30	0,4587	9,5	428,5	-171,5	615,0
1,35	0,5918	-20,5	458,5	-205,0	650,0
1,40	0,7109	-49,0	488,5	-237,0	681,0
1,45	0,8075	-77,0	516,0	-267,5	712,5
1,50	0,8792	-105,0	544,5	-299,0	743,0

Table 17 show the Rosenbaum bounds for the estimates for the FRAM program (Table 11) where we found positive but statistically insignificant effects. The table shows that the p-values of the Hodges-Lehman point estimates exceed the usual 0.05 threshold when we change the odds ratio of receiving treatment from 1.0 to 1.1 and that the estimated treatment effect may be as high as -33 or as low as 471. At the lower bound this is not significant different from zero.

Table 18 Rosenbaum bonds – The Network program – matched pairs

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1,00	0,0000	903,7	903,7	567,3	1260,5
1,05	0,0000	756,1	1053,6	429,5	1421,1
1,10	0,0001	621,0	1200,5	300,0	1580,0
1,15	0,0010	494,3	1344,5	178,0	1736,0
1,20	0,0090	376,0	1487,3	64,0	1890,0
1,25	0,0474	263,5	1626,0	-45,0	2040,4
1,30	0,1579	157,5	1764,5	-150,9	2189,5
1,35	0,3590	57,0	1900,0	-251,5	2338,0
1,40	0,6016	-39,9	2032,8	-350,6	2484,5
1,45	0,8035	-133,5	2164,0	-446,3	2630,0
1,50	0,9238	-223,0	2295,6	-541,5	2773,0

Table 18 shows that a 25% change in the odds ration of receiving treatment would render the estimates of treatment effects from the Network program vulnerable to hidden bias.

Our second kind of estimators; the difference-in-differences based upon matched data for consecutive years after the first year after treatment, would, by the line of reasoning outlined above, require sensitivity analysis for a minimum of 48 datasets for matched pairs, provided that we treat Regional Venture Capital Loans and Investment Grant as one dataset. The tables for these analyses would require a considerable number of pages. We therefore present only the minimum gammas necessary for a likely effect of unobserved variables on our initially matched pairs over the observation period. Tables for Rosenbaum Bounds are provided in appendix A.

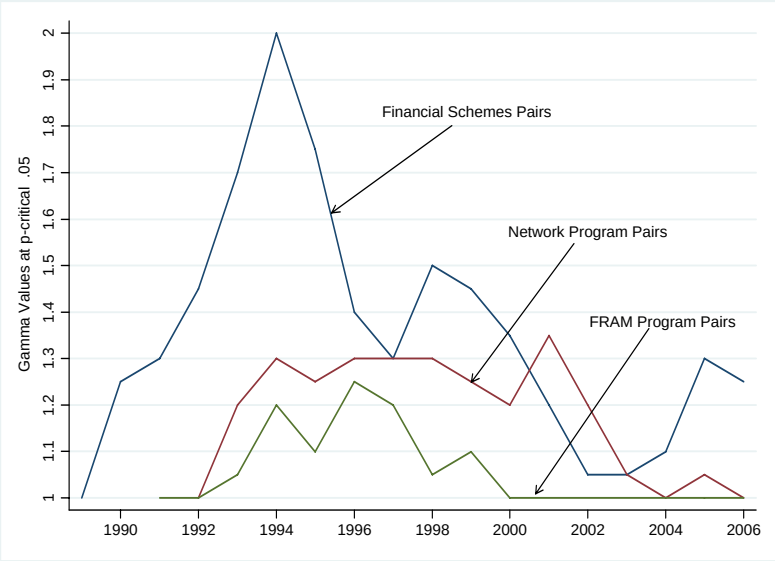


Figure 16 Gamma values at $p \leq .05$ for Matched Pairs

Figure 16 indicates the matched pair datasets for Regional Venture Capital Loans *and* Investment Grant (all three categories) are relatively robust in the sense that for

the period up to the year 2002, with an exception for 1997, the influence of an unobserved variable has to be considerable to affect estimates. Also, the Network program seems fairly robust up to the year 2001, although less healthy than the matched pairs for Regional Venture Capital Loans *and* Investment Grant. The matched pair dataset for the FRAM program appears to be highly vulnerable to unobserved heterogeneity and the estimates could thus easily be altered by variables not included in the matching procedure. Estimates for the years after 2002 are clearly less trustworthy for all three datasets.

8.6.1 Sensitivity analysis – a summing up

The effects for Venture Capital Loans & Investment Grant are fairly insensitive to hidden bias. The estimates for the Network program and for the FRAM program in particular, are highly vulnerable to hidden bias. Thus, for the latter two programs unobserved heterogeneity may be a problem. The sensitivity analysis for these two programs implies that the conditional independence assumption (CIA, see assumption 4, section 6.3, page 61) may be violated. The sensitivity analysis reveals that the insignificant effects of the FRAM program appears to be the most exposed to the influence of hidden bias, while there is little evidence that the stronger the effects the less vulnerable the estimates are to hidden bias. The Rosenbaum bounds is a kind of "worst-case" scenario (DiPrete & Gangl, 2004) which simply shows how large the influence of a confounding variable must be to undermine conclusions from the matching analyses. Or, to quote Arild Aakvik (2001) : "*A sensitivity analysis shows how biases might alter inferences. However, it does not indicate whether biases are present or what magnitudes are plausible.*"

Clearly, the sensitivity analysis[18] hints that we should look for additional evidence that could corroborate our initial estimates. We do so by using more of the information available, namely the possibility of constructing panel data.

[18] Overview of the sensitivity analysis is provided in **Appendix A**.

9. Analysis – two-period estimation

The before-after difference-in-differences two-period estimator is the traditional measure for two-period panel data. We extend this estimator to cover all observation periods where we have a sufficient number of observations. Let *A* be the control group and *B* the treatment group. The basic tool for the two-period estimators is dummy-regression with an equation of the general form:

$y = \beta_0 + \beta_1 * dB + \delta_0 * d2 + \delta_1 d2 * dB + u$, where y is the outcome of interest, here added value. The dummy *dB* captures differences between the treatment group and the control group that may exist prior to the intervention by the government, *d2* catches what could cause changes in y even in the absence of an intervention. The coefficient of interest is δ_1, the interaction between *d2* and *dB* that captures the difference-in-differences. Thus, all variables included in the regression equation are simple zero-one dummies coded 1 for treatment and zero for control, *d2* is a dummy coded 0 for the period before intervention and 1 for period after intervention.

Table 19 Regression setup for difference-in-differences

Difference-in-differences	Before	After	Difference *(After-Before)*
Treatment	$\beta_0 + \delta_0$	$\beta_0 + \beta_1 + \delta_0 + \delta_1$	$\beta_1 + \delta_1$
Control	β_0	$\beta_0 + \beta_1$	β_1
Difference *(Treatment-Control)*	δ_0	$\delta_0 + \delta_1$	δ_1

The two-period estimator has its particular advantages and disadvantages. First and foremost it is vulnerable to effects of unobserved variables. Clearly, the further we move away from the time of intervention, the more likely it is that outcomes in terms of added value are influenced by other factors than the intervention[19]. The main

[19] The sensitivity analyses indicate that this is in particular the case for the FRAM program and the Network program.

advantage of the two-period estimator in this setting is that it provides an opportunity to study how effects evolve over time.

As hinted at in Figure 1, page 39, the patterns of effects over time are not obvious. The progress over time may evolve in a variety of shapes and knowledge of the approximate nature of these patterns may be of crucial importance. In the case of pattern *a)* and *b)* in Figure 1, evaluations that take place at time p_2 or p_3 may provide a positive *(b)* or a negative *(a)* answer to the question of effects of intervention dependent upon the point in time when the evaluation is carried out. It is not the case that effects invariably occur when it is convenient to carry out an evaluation. Moreover, from a governmental point of view, *c* is the preferable pattern. Thus, although we are aware of the weakness of the two-period estimators, they are presented here to provide an understanding of how of the effects of interventions evolve over time.

The two-period estimates follow the setup shown in Table 19 where the interaction term (δ_1) gives the estimates of the DiD which is simply the differences between the difference of the means of added value in the matched treatment group and control group respectively before and after treatment, and thus, simply an estimate of a single mean value $\overline{DiD} = (\overline{Y}_{Treatment,After} - \overline{Y}_{Control,After}) - (\overline{Y}_{Treatment,Before} - \overline{Y}_{Control,Before})$. It is well known that the traditional dummy-regression setup gives unbiased estimates of the means but inflated standard errors. Bootstrapping the means does not solve this problem (Abadie, 2002). In the tables that follow we therefore estimate the means and the standard errors by the rank-based methods as described in Gardner and Altman's (1989) book "Statistics with Confidence", pp. 74-79. A normal approximation with a continuity correction of 0.5 is used, rather than tables of exact

values. The accuracy should be adequate when both groups have at least 5 observations[20]. The tables show estimates and confidence intervals for each year.

Table 20 Regional Venture Capital Loans, ATT-DiD - 1000 NOK

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	302	151	0.00		240.361	-473.01	473.01	*Before*
1990	338	187	134.27		242.180	-342.11	610.65	
1991	338	187	-75.96		226.004	-520.52	368.60	
1992	321	170	139.83		266.082	-383.67	663.32	*Intervention*
1993	311	160	576.53		293.880	-1.73	1154.79	
1994	309	158	959.87	*	344.172	282.64	1637.11	
1995	306	155	479.96		358.942	-226.36	1186.29	
1996	306	155	487.11		447.413	-393.31	1367.52	
1997	295	144	599.34		569.283	-521.06	1719.74	
1998	285	134	910.98		589.863	-250.10	2072.05	
1999	279	128	2037.20	*	520.143	1013.26	3061.13	
2000	275	124	1835.83	*	582.345	689.37	2982.29	*After*
2001	266	115	1359.01	*	648.526	82.07	2635.95	
2002	253	102	1356.54	*	637.585	100.85	2612.24	
2003	250	99	614.49		575.237	-518.48	1747.47	
2004	242	91	1284.53	*	641.128	21.58	2547.49	
2005	235	84	2001.74	*	598.097	823.38	3180.11	
2006	228	77	2183.59	*	725.685	753.62	3613.57	

Average 1989-2006 = 938.05 * *p≤.05*
Average 1995-2006 = 1262.53

Table 21 Investment Grants, ATT – DiD – 1000 NOK

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	424	212	0.00		303.253	-596.08	596.08	*Before*
1990	512	300	183.46		257.301	-322.04	688.96	
1991	549	337	112.18		254.361	-387.46	611.82	
1992	556	344	341.63		274.301	-197.17	880.43	*Intervention*
1993	576	364	733.69	*	305.894	132.88	1334.50	
1994	574	362	1023.92	*	315.773	403.70	1644.14	
1995	560	348	874.00	*	341.334	203.54	1544.45	
1996	564	352	936.95	*	433.616	85.25	1788.66	
1997	546	334	1081.25		575.458	-49.14	2211.65	
1998	523	311	1555.02	*	578.113	419.30	2690.74	
1999	507	295	1811.97	*	734.455	369.00	3254.93	
2000	487	275	2299.33	*	806.736	714.21	3884.46	*After*
2001	465	253	1579.54	*	788.704	29.66	3129.43	
2002	440	228	1015.54		795.883	-548.68	2579.77	
2003	420	208	1571.32	*	792.444	13.65	3128.99	
2004	407	195	1901.06	*	865.542	199.55	3602.58	
2005	394	182	1397.73		789.539	-154.53	2950.00	
2006	376	164	1340.86		919.702	-467.58	3149.29	

Average 1989-2006 = 1097.75 * *p≤.05*
Average 1995-2006 = 1447.05

[20] Note the difference between pairs of observations and total observations. The two would differ since means for the years *prior* to treatment are subtracted from every mean for years *after* treatment and standard deviations are calculated by the method suggested by Gardner and Altman (1989).

Table 22 Capital Loans & Investment Grants – Combined – 1000 NOK

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	598	299	0.00		248.489	-488.02	488.02	*Before*
1990	673	374	438.27	*	220.915	4.50	872.04	
1991	690	391	596.30	*	227.598	149.43	1043.17	
1992	681	382	1049.85	*	247.518	563.86	1535.85	*Intervention*
1993	676	377	1277.25	*	274.103	739.05	1815.45	
1994	671	372	1814.07	*	309.561	1206.24	2421.90	
1995	666	367	2309.14	*	357.225	1607.71	3010.56	
1996	667	368	2013.32	*	424.515	1179.76	2846.87	
1997	641	342	2175.29	*	489.538	1213.99	3136.58	
1998	620	321	2760.49	*	569.185	1642.72	3878.26	
1999	593	294	2621.42	*	582.131	1478.12	3764.72	
2000	579	280	2942.37	*	693.571	1580.14	4304.60	*After*
2001	551	252	1972.08	*	656.114	683.28	3260.88	
2002	537	238	1522.45	*	646.891	251.70	2793.21	
2003	520	221	1279.49	*	641.793	18.66	2540.33	
2004	509	210	1420.75	*	685.671	73.64	2767.86	
2005	493	194	2113.10	*	661.747	812.90	3413.31	
2006	465	166	2444.49	*	766.326	938.58	3950.40	

Average 1989-2006 = 1708.34 * $p \leq .05$
Average 1995-2006 = 2131.20

Gardner and Altman's (1989) method[21] calculates the confidence intervals based on all available data independent of whether data is missing for a pair or not. The estimated means are exactly the same as for the dummy regression model but the confidence intervals are narrower. Since we know that we have inflated standard errors, Gardner and Altman's method compensates this and gives fewer missing cases. For analyses at the program level this may method be an appropriate strategy[22]. Even though we compare balanced pairs of treatment and controls, the central issue here is *yearly* comparisons and thus, the year by year averages are the preferred basis for the two-period difference in differences. The discrepancy between the two estimation methods is shown in table C9 and the figures C1 to C3 in appendix C. Estimates are mostly very similar. The exceptions are Investment Grants and the FRAM program. For the latter the divergence is substantial, in particular after the year 2000.

[21] The enthusiasm for Gardner and Altman's methods among other statisticians is somewhat reserved.
[22] Limiting the differences to true pairs only gives slightly different estimates and more missing cases. The sensitivity analyses are all based on true pairs (no missing data), which by definition is required.

Table 23 Two-period estimates for the FRAM- program – 1000 NOK

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1991	323	646	0.00		222.425	-436.77	436.77	*Before*
1992	349	672	127.77		231.808	-327.39	582.93	
1993	365	688	66.88		198.406	-322.67	456.44	
1994	374	697	358.46		191.654	-17.83	734.75	*Intervention*
1995	380	703	434.07	*	196.238	48.78	819.36	
1996	379	702	642.50	*	209.257	231.65	1053.35	
1997	369	692	675.69	*	254.292	176.40	1174.99	
1998	346	669	489.10		291.506	-83.29	1061.49	
1999	319	642	924.37	*	329.651	277.04	1571.70	
2000	297	620	871.71	*	376.273	132.78	1610.63	
2001	269	592	529.73		559.126	-568.39	1627.85	
2002	250	573	605.69		444.499	-267.36	1478.74	*After*
2003	233	556	518.75		489.980	-443.70	1481.19	
2004	220	543	836.97		534.318	-212.62	1886.56	
2005	206	529	568.49		598.684	-607.61	1744.59	
2006	188	511	443.39		657.031	-847.44	1734.21	

Average 1991-2006 = 505.85 * *p≤.05*
Average 1997-2006 = 643.13

Table 24 Two-period estimates for the Network Program – 1000 NOK

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1992	1364	2728	0.00		558.121	-1094.38	1094.38	*Before*
1993	1360	2724	458.18		582.239	-683.50	1599.86	
1994	1330	2694	830.25		621.703	-388.82	2049.31	*Intervention*
1995	1288	2652	1214.54		656.764	-73.28	2502.36	
1996	1245	2609	1847.26	*	698.991	476.63	3217.90	
1997	1160	2524	1930.82	*	694.840	568.30	3293.33	
1998	1082	2446	2232.43	*	744.199	773.10	3691.75	
1999	1000	2364	4008.54	*	906.567	2230.79	5786.29	
2000	958	2322	3636.47	*	971.249	1731.87	5541.08	
2001	899	2263	5611.40	*	989.927	3670.14	7552.66	*After*
2002	829	2193	4513.24	*	1031.067	2491.27	6535.22	
2003	772	2136	3397.61	*	1009.099	1418.69	5376.53	
2004	727	2091	3722.14	*	1043.948	1674.85	5769.42	
2005	670	2034	4058.36	*	1093.130	1914.59	6202.14	
2006	614	1978	4199.36	*	1150.487	1943.07	6455.66	

Average 1992-2006 = 2777.37 * *p≤.05*
Average 1997-2006 = 3731.04

At a glance the estimates suggest that our hypotheses concerning the ordering of expected magnitudes of the effects of the various financial schemes and programs are supported: It appears as if the combination of Regional Venture Capital Loans and Investment Grants gives the best return. Investments Grants singly appear to produce a somewhat better yield than Regional Venture Capital Loans singly, but the

difference is not considerable. The FRAM program produces as expected the lowest result. The surprise is the Network program which appears to give a substantial return. The result is no less surprising as previous evaluations of the Network program (Econ, 1998) have come to different conclusions.

10. Analysis – panel data estimation

With true panel data we can follow the same units over time. True panel data allows us to address the issue of unobserved fixed effects and hidden bias. Consider the equation: $y_{it} = \beta_0 + \beta_1 dB_{it} + \delta_0 d2_{it} + \delta_1 d2_{it} * dB_{it} + v_{it}$, where i denotes the unit, and t denotes time. As before[23], the dummy *dB* captures differences between the treatment group and the control group that may exist prior to the intervention by the government, *d2* catches what could cause changes in *y* even in the absence of an intervention. The coefficient of interest is δ_1, the interaction between *d2* and *dB* that captures the difference-in-differences. The outcome variable y_{it} is added value *(AV)*. In the panel data setup the only difference from the two-period setup is that $d2_{it}$ denotes the after treatment period in the sense; all observed years after the treatment. Since the $d2_{it}$ may differ for the single record because firms start receiving treatment at different times, we use two versions of $d2_{it}$; one that considers $d2_{it}$ to be the period from the first year after the treatment period and a variable $d^{t+}2_{it}$ that also includes the treatment period. This allows us to separate out the effects that may occur *within* the treatment period e.g. immediate effects of treatment. If we consider the error term of the equation above as $v_{it} = a_i + u_{it}$ i.e. as a composite of the usual error term u_{it} and a time constant component a_i, *a fixed effects*, it is easier to grasp the virtues of the panel model. If this unobserved fixed effect, a_i is correlated with the independent variables, the ordinary least squares (OLS) suffers from omitted variable bias. If, on the other hand, a_i is uncorrelated with the independent variables, OLS gives unbiased estimates. However, if the error terms are correlated, the stan-

[23] Note the similarity between the general equation for the panel data estimation of difference-in-difference and the corresponding equation for the two-period case (see section 9 page 105).

dard errors can be seriously wrong (Wooldridge, 2002). The random effect estimator is a generalized least squares (GLS) method for obtaining more correct standard errors. We estimate[24] both the random effect model (RE) using GLS and the fixed effect model (FE) and compare the estimates of the two models by means of a Hausman test (Hausman, 1978) which compares the fixed effects model and the random effects model under the null hypothesis that the individual effects are uncorrelated with the other regressors in the model. The Hausman test tests the differences between the estimated covariance matrix in FE model and the RE model under the assumption that the difference follows a χ^2 distribution. A *p*-value for the chi-square larger than .05 implies that the random effects model is the most efficient model while a significant *p*-value indicates that the fixed effects model should be preferred. Deciding what model to choose based on the Hausman tests only is, according to Badi H. Baltagi "… not as easy a choice as it might seem" (Baltagi, 2008). Thus our preference for the fixed effects model relies on convenience: The fixed effect model removes the omitted variable bias resulting from the exclusion of unobserved variables that vary over the individual firms but are constant over time. We also rely on traditions (Ashenfelter, 1978; Ashenfelter et al., 1985) since fixed effects can be interpreted as treatment effects and thus are the models we want to exploit. An overview over the contrasts between the fixed effects and the random effects models and the corresponding Hausman tests is provided in Appendix B.

The panel estimators follow a dummy regression setup similar to those presented in section 9, page 105. The explicit inclusion of the time dimension in combination with matched pairs of data as the control group allows us to explore the before-after intervention effect based on the best possible comparison data. As noted earlier, a

[24] The Stata program **xtreg** is used for most panel data estimations.

requirement for difference-in-differences is that the control group is independent of the treatment assignment, that is $Y_{ot} - Y_{ot'} \perp D|X$ and that no comparison unit is beforehand excluded, conditional upon the matching vector, i.e.

$0 < prob(D = 1|X) < 1$, where *t'* means the period before treatment and *t* means the period after treatment. Moreover, the pattern of development over time for the dependent variable y_{it} in the control group should follow a path that does not deviate systematically from that of the treatment group (Imbens et al., 2007). As indicated by Figure 4, page 57 this last assumption appears to be satisfied, at least for the first part of the observation period.

Although the Hausman χ^2 is are insignificant for almost all FE-RE comparisons in Appendix B and all models have very similar estimates for treatment effects we expect that both temporal and spatial correlation may still be a problem. We therefore report both the fixed effect standard errors and the Driscoll-Kraay (1998) corrected standard errors which gives more conservative estimates of the confidence intervals.

Table 25 Fixed Effects Panel Data Estimates for Average Treatment Effects

			Regional Venture Capital Loans			
Treatmen Period	Coef	Std. Err.	*t*	P>|t|	[95% Conf.	Interval]
1995 - 2006	**733.97**	232.111	3.2	0.0020	278.93	1189.01
1991 - 2006	**1125.64**	299.488	3.8	0.0000	538.52	1712.77
With Driscoll-Kraay standard errors						
1995 - 2006	733.97	338.343	2.2	0.0310	68.91	1399.02
1991 - 2006	1125.64	190.197	5.9	0.0000	751.79	1499.50
			Investment Grants			
Treatmen Period	Coef	Std. Err.	*t*	P>|t|	[95% Conf.	Interval]
1995 - 2006	**1615.50**	589.719	2.7	0.0060	459.55	2771.46
1991 - 2006	**888.49**	298.614	3.0	0.0030	303.14	1473.83
With Driscoll-Kraay standard errors						
1995 - 2006	1615.50	812.750	2.0	0.0470	20.28	3210.72
1991 - 2006	888.49	128.879	6.9	0.0000	635.49	1141.48
			Both Investment Grants & Venture Capital Loans			
Treatmen Period	Coef	Std. Err.	*t*	P>|t|	[95% Conf.	Interval]
1995 - 2006	**1815.00**	280.090	6.5	0.0000	1265.97	2364.02
1991 - 2006	**2409.68**	382.680	6.3	0.0000	1659.56	3159.80
With Driscoll-Kraay standard errors						
1995 - 2006	1815.00	581.034	3.1	0.0020	674.69	2955.31
1991 - 2006	2409.68	431.344	5.6	0.0000	1563.15	3256.22
			The FRAM Program			
Treatmen Period	Coef	Std. Err.	*t*	P>|t|	[95% Conf.	Interval]
1997 - 2006	**358.71**	128.631	2.8	0.0050	106.57	610.85
1992 - 2006	**516.79**	169.948	3.0	0.0020	183.66	849.92
With Driscoll-Kraay standard errors						
1997 - 2006	358.71	94.923	3.8	0.0000	172.36	545.05
1992 - 2006	516.79	131.010	3.9	0.0000	259.61	773.98
			The Network Program			
Treatmen Period	Coef	Std. Err.	*t*	P>|t|	[95% Conf.	Interval]
1997 - 2006	**3254.64**	250.065	13.0	0.0000	2764.50	3744.77
1993 - 2006	**2707.81**	274.821	9.9	0.0000	2169.15	3246.47
With Driscoll-Kraay standard errors						
1997 - 2006	3254.64	526.340	6.2	0.0000	2222.57	4286.70
1993 - 2006	2707.81	654.988	4.1	0.0000	1423.48	3992.13

11. Analysis – comparing estimates

11.1 Comparing the various results of the estimations

We have obtained various estimates for all three datasets using different estimation methods. To what extent are these estimates similar or different? Table 26 gives an overview of the results. We are aware that the different estimation techniques should *not* be expected to produce very similar results. Moreover, for the two-period estimates we have nothing that really corresponds to the panel data estimates and have to compare the panel estimates with averages over the observation periods.

Table 26 shows that the estimates are different, but not all that different. Their differences in magnitudes are mainly within a reasonable range. Moreover, the particularity of the intervention may explain some of these differences. For the Regional Venture Capital Loans there is a considerable gap between the panel estimates for post-treatment period and the average over the two-period estimates for the same period. (0.733 mill NOK versus 1.262 mill NOK on the average per firm). Some of this deviance may be explained with the differences within in the treatment period (the entire treatment period minus the first year is included in the panel data analyses) and the increasing number of missing cases towards the end of the estimation period. Moreover, loans should also be expected to have less of an immediate effect upon added value since the payback of loans may affect net operating result. This is also in correspondence with the visual inspection of Table 20 (page 107) and Figure 17 (page 117). Also, the negative estimates within the intervention period may contribute to these contradicting predictions. Comparison of the various esti-

mates and the attempt to explain differences can not go much beyond mere speculations. The most conspicuous differences between the results from the panel estimates and the two-period averages can, however, most likely be explained as differences due to the fact that outcomes *within* the treatment period have substantial influence. The difference between the estimates that include the treatment period and the estimates that consider the post-treatment period only should give a rough estimate of the contribution to ATT that occurs within the treatment period.

Table 26 Comparisons of estimates from various methods – 1000 NOK

	Regional Venture Capital Loans		
Treatmen Period	Panel Data Estimates	Means over Two-Period Estimates	Medians over Two-Periode Estimates
1995 - 2006	**733.97**	**1 262.53**	**1 320.54**
1991 - 2006	**1 125.64**	**938.05**	**762.74**
	Investment Grants		
Treatmen Period	Panel Data Estimates	Means over Two-Period Estimates	Medians over Two-Periode Estimates
1995 - 2006	**1 615.50**	**1 447.05**	**1 476.38**
1991 - 2006	**888.49**	**1 097.75**	**1 052.59**
	Investment Grants & Venture Capital		
Treatmen Period	Panel Data Estimates	Means over Two-Period Estimates	Medians over Two-Periode Estimates
1995 - 2006	**1 815.00**	**2 131.20**	**2 144.19**
1991 - 2006	**2 409.68**	**1 708.34**	**1 893.08**
	The FRAM Program		
Treatmen Period	Panel Data Estimates	Means over Two-Period Estimates	Medians over Two-Periode Estimates
1998 - 2006	**358.71**	**643.13**	**568.49**
1992 - 2006	**516.79**	**505.85**	**524.24**
	The Network Program		
Treatmen Period	Panel Data Estimates	Means over Two-Period Estimates	Medians over Two-Periode Estimates
1997 - 2006	**3 254.64**	**3 731.04**	**3 865.34**
1993 - 2006	**2 707.81**	**2 777.37**	**3 397.61**

Clearly, the various estimates do to a substantial extent corroborate each other. Despite the fact that they are based on very different methods and that the dependent variable, added value is a volatile measure, the estimates are within reasonable

ranges for all datasets. Moreover, the development over time that is revealed by the two-period estimates makes sense. Figure 17 to Figure 19 show the means (the solid lines) and confidence intervals (the short-dotted lines) for the panel estimates and the two-period estimates.

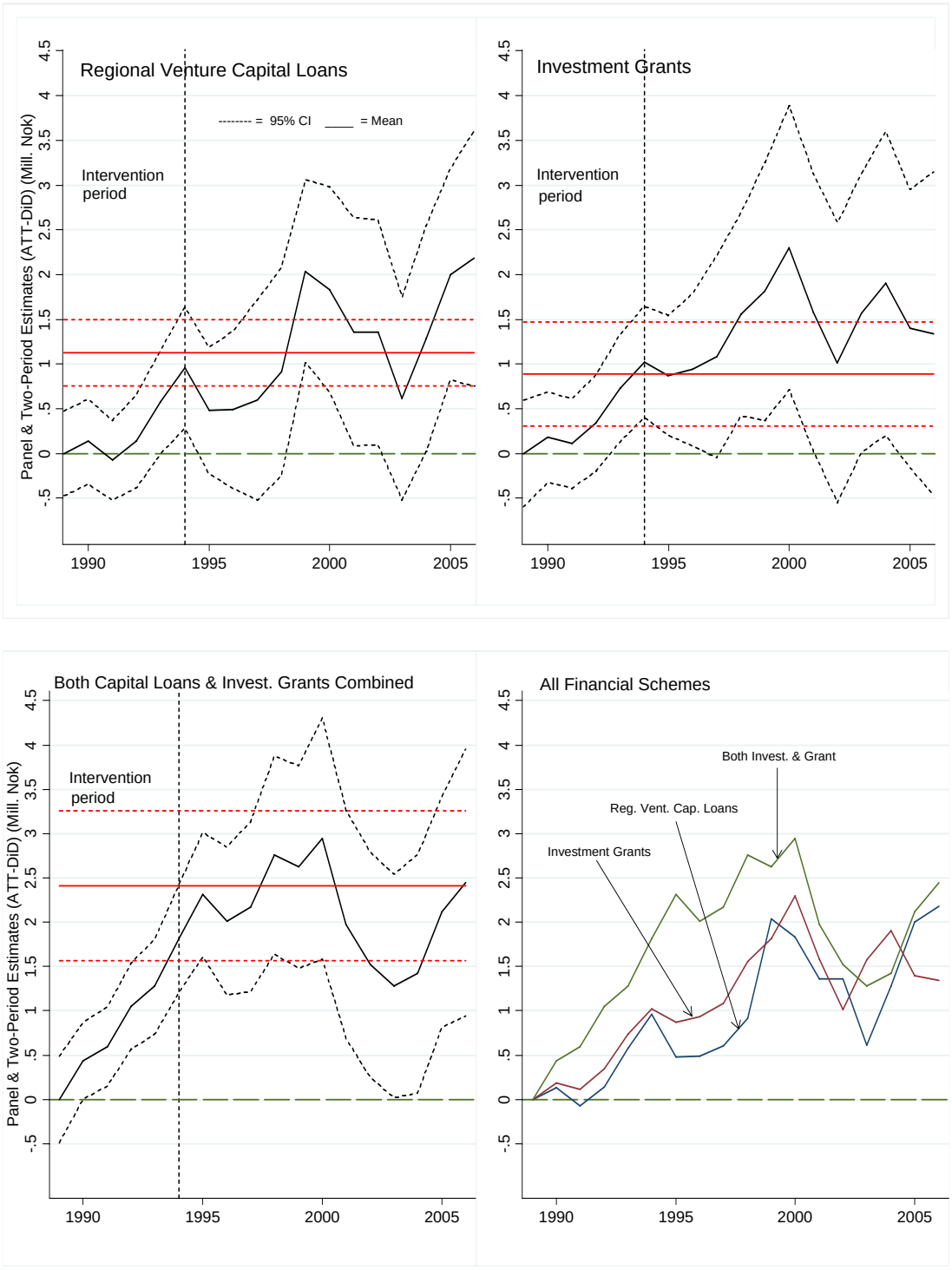


Figure 17 Two-period & Panel data estimates for the Financial Schemes

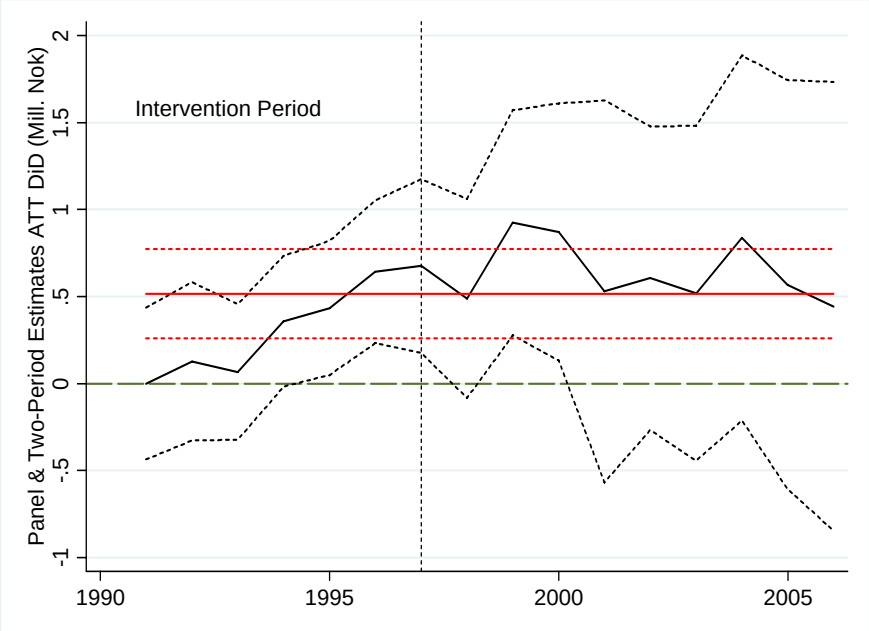


Figure 18 Two-period & Panel data estimates for the FRAM program

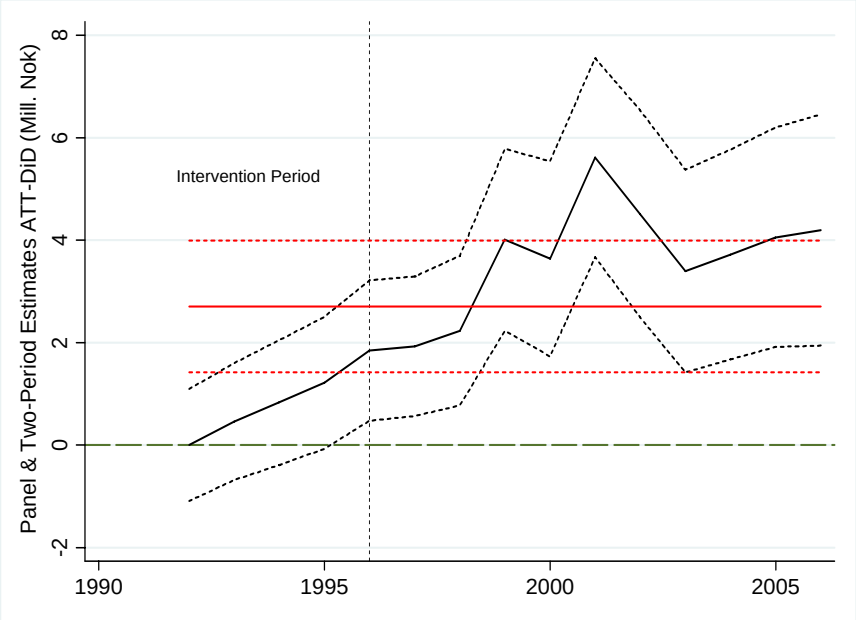


Figure 19 Two-period & Panel data estimates for the Network program

The two-period estimates reveal the patterns of evolvement over time (Figure 17 to Figure 19) while the panel estimates are simply point estimates of means with confidence intervals (the dotted lines), which, for the purpose of illustration are drawn as parallel lines over the entire observation period i.e. we use the panel estimates for the entire observation periods (the shadowed lines in Table 26 and the red lines in the figures). Note that whenever the lower bound for the confidence intervals goes beyond zero, the estimates are not significantly different from zero at the 95% level.

Figure 17 shows that the two-period and panel estimates compares quite well for the financial schemes. The panel estimates appears to be a reasonable compromise. A pattern of a post-treatment increase and-then a decrease in outcomes is evident for all financial schemes. The exception is Regional Venture Capital Loans (upper left) where the two-period estimates appear to have a delayed growth right after the end of the treatment period. Towards the end of the observation period some estimates are not significantly different from zero while others increase, giving the highest estimates in the last year of the observation period. The estimates for Regional Venture Capital Loans and Investment Grants are insignificant towards the end of the observation period. This upward tendency after the year 2003 should be downplayed since the matched pair datasets are increasingly vulnerable to the influence of unobserved variables as we move forward over the observation period. As indicated by the sensitivity analysis (e.g. Figure 16) the estimates for the last couple of years appear to be less robust[25]. For the Regional Venture Capital Loans & Investment Grant combined all estimates are significantly different from zero. The overall pattern for the financial schemes is an upturn from the end of the treatment period in

[25] Note that many authors consider the Rosenbaum bounds test to be too strong and therefore potentially misleading Becker, S. O., & Caliendo, M. 2007. Sensitivity analysis for average treatment effects. ***Stata Journal***, 7(1): 71-83..

1994 until a peak is reached in 2000. From 2000 to 2002 the two-period estimates go down and all three schemes have their lowest estimates in 2003.

The FRAM program shows a more confusing pattern with only a few significant estimates. All estimates are positive, but three of the five significant estimates are within the treatment period. As shown in Figure 18 only estimates within the treatment period or in a few years right after the intervention period are significantly different from zero (above the green dotted line at zero).

The Network program shows an almost steady rise-and fade-off pattern and has relatively large, significant magnitudes of the estimates over the entire period after intervention. The typical low estimate in 2003 is with the subsequent increase in the years to follow is evident in Figure 19.

As shown in Figure 20 and Figure 21 these patterns of evolvements of ATT over the observation period, the rise on the way to 2000 and fall towards 2003, is corroborated by other figures such as the Index of Production and the Business Tendency Survey in manufacturing, from Statistics Norway. Thus, the overall pattern also mirrors general economic tendencies over the period. The similarity of these figures indicates that although the contributions from Innovation Norway have undoubtedly contributed positively to those firms that received funds, firms are nevertheless affected by a variety of other factors.

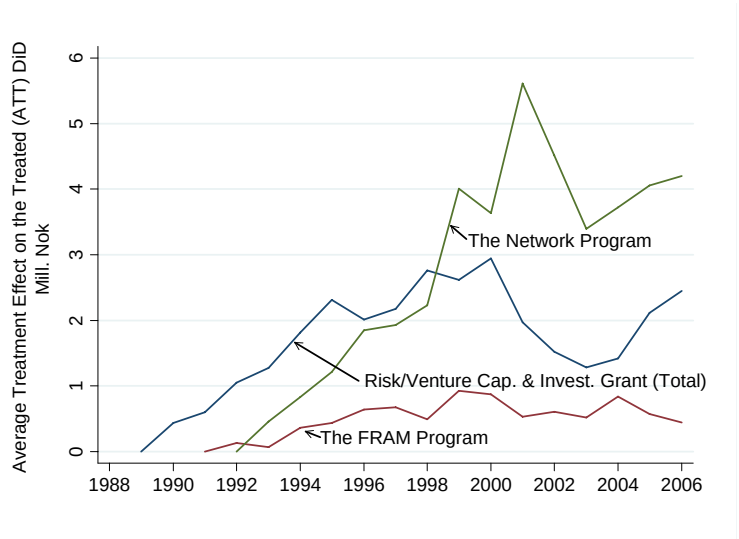


Figure 20 Evolvement of ATT over time for all programs

Figure 21 Business tendencies 1999 to 2008 – Statistics Norway

12. Interpretation of the findings – A scenario approach

12.1 Introduction to the Scenario Approach to Interpretations

This section is different from the other sections in the sense that *interpretations go beyond analysis.* With *go beyond* we mean that interpretations are anchored in a rhetorical context that is different from the analysis perspective. While statistical analyses are based on theories of probability and statistical inference, *interpretations* of results involve a conceptualization of reality i.e. ideas about the situation results are read into. It is well known within the evaluation field that results tend to be construed according to the preferences of the readers and that is hard for the researcher to counter such interpretation with reference to e.g. distributional assumptions. Thus, distorted interpretations of results are likely to occur and they may either overstate or understate the meaning of the results. In the case of causal effects of intervention it is not unlikely that the devotees of market liberalism would have an understanding of the results that deviate from that of, say, the more Keynesian inclined economist.

To avoid the impression that the interpretations outlined here are authoritatively the only possible understandings we use the term *scenario*. A scenario can be defined as a *possible set of future events* or as an imagined or projected sequence of events where any of several detailed plans of possibilities are sketched out. We use the term scenario despite the fact that we are mainly looking at the *past*. If we substitute the term "future events" for the term "likely events" in the definition of scenario, it is easier to see why scenario is also a proper term for descriptions conditional upon

the conjecture that our analyses are correct and a number of assumption that must be made to justify the ceteris paribus clause. Thus, in the following we present various *scenarios* based upon our ATT-estimates. These scenarios assume that our estimates are correct and also rely on untestable assumptions such as no equilibrium effects. Moreover, we assume a constant discount rate over the periods under scrutiny and consider the returns in terms of added value due to positive ATTs as "cash flows" to the society. This "cash flow" concept adds to the versatility of the scenarios by facilitating lines of reasoning known from financial economics. It also makes it easier to make comparisons across the different financial schemes and interventions.

The first part of this section outlines some general principles and discusses the available information on costs associated with the various schemes and programs. The fact that we do not have ATT-estimates for all firms registered in our data as users of the financial schemes, led to the decision to carry out separate comparisons of costs and returns *both* for the proportion analyzed and for all firms known to have received funds. Thus, for the financial schemes we make *two* comparisons; one for all firms registered in the database from Innovation Norway i.e. analyses where we compare costs and gains *as if* the average estimated gains (the ATT) are valid for all firms involved; and one for the proportion of costs associated with those firms actually *in* the analyzed sample. These dual comparisons are carried out for the financial schemes only.

The meaning of loss on loans is discussed in subsection 12.3.1. Since stipulated losses entail considerable ambiguity we sketch two possible interpretations that are used for the scenarios. Subsection 12.3.2 compares the gains from grants and loans

and subsection 12.3.3 uses the "cash flow" perspective in order to calculate internal rates of return. Since the uncertainty concerning cost side is considerable for the FRAM program and the Network program, scenarios for these programs are more briefly sketched in subsections 12.4.1 and 12.5.1 respectively. Finally a comparison of the scenarios for the schemes and programs is introduced in subsection 12.6.

12.2 Average Treatment Effects and the Scenario Approach

Interpretations of the findings should be straightforward in the sense that all we have is estimates of the average *added value* for the firms subject to intervention, compared to what should be expected provided that no intervention had taken place. Expected values above zero are interpreted as a positive contribution to the causal effect of interventions, likewise, below zero constitutes negative effects. Our analyses do *not*, however, justify conclusions at the firm level, only group level figures are warranted, and here, only program level estimates are reported. The sensitivity analysis of pairs (8.6, page 98) provides an idea of the likelihood that our estimates of the average treatment effects (ATT) are affected by other variables *not* included in the analyses. Analogous reasoning at the firm level is not feasible. We have to accept that we have little control over unknown factors and events that may affect outcomes for the individual firm. We can clearly evaluate outcomes at various aggregate levels such as industry levels or regional levels, which may also be of interest. The interpretations presented here do, however, solely concern the program level i.e. the intervention program or financial scheme as a whole.

Any meaningful interpretation of results should relate outcomes to program costs. The presumed benefits from the financial schemes and programs have to be com-

pared to the costs of the initiatives. What we are doing here is, however, not conventional cost-benefit analyses[26]. The ambitions are simply to visualize and demonstrate the meaning of our estimated average differences between the treated and the non-treated firms and speculate over interpretations under various scenarios. When *time* and *interest rate* have to be considered, interpretations may not always be clear-cut. Moreover, the cost side also requires some disputable decisions concerning what to compare.

We do not have accurate measures for all the costs associated with the programs and financial schemes we study. For the Regional Venture Capital Loans and the Investment Grant we have fairly precise figures; we have the exact allotment decisions for each individual firm. The main problem is that Regional Venture Capital Loans are *loans*, and as such, advances that carry substantial risks. We have no exact measures of actual losses, but for the year 1993, 1992 and 1993 a loss account covering 30% of the lending was appropriated. For the FRAM program we have cost estimates from previous evaluation reports (Nesheim, 1997) and for the Network program we have figures from previous evaluations (Econ, 1998).

Interpretations are demonstrated in the form of *scenarios.* That is, we try to exemplify what the estimated average effects of treatment on the treated (ATT) would mean assuming various possible circumstances. Most of these scenarios require additional assumptions which are explicitly stated but not always testable.

[26] We do, however, believe counterfactual reasoning is an implicit albeit mostly ignored assumption in CBA.

12.3 The Financial Schemes

According to available administrative records a total of 4.285 million NOK was allocated for the Regional Venture Capital Loans and Investment Grant programs over the period 1990 to 1994. The firms we were able to identify for our analyses of these programs account for 1014 mill NOK or approximately 24% of total allotments. These figures are based upon the analyses of 5831 registered decisions concerning funding assignments to 3298 firms. Our analysis includes 1117 firms for the post treatment period i.e. approximately a third of the firms involved, and since several decisions concerned the same company, 37.7% of the individual decisions about allotments. Compared to the distribution of total funding across regions, the 24% proportion we use for analysis covers the various regions quite well.

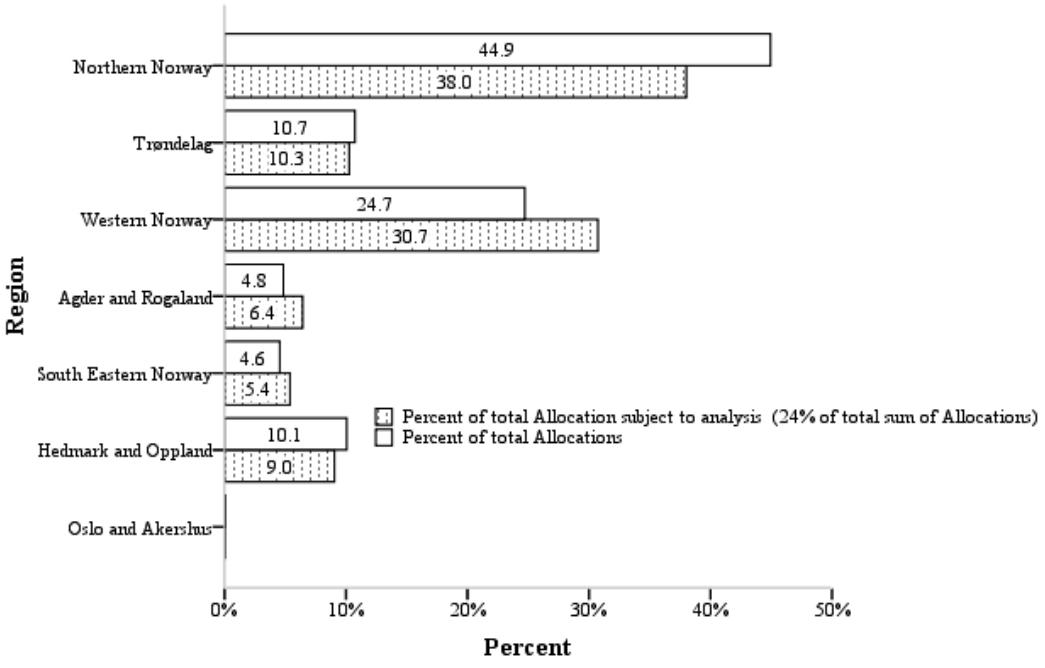


Figure 22 Distribution of Funding across Regions – All Financial Schemes

Only around 10% of the cases are lost due to bankruptcies or other causes of missing data across the observation periods. We are able to identify most of the funding-carrying treatment cases and their corresponding matched pair.

Clearly, the two financial schemes are different; Regional Venture Capital Loans are *loans* that require repayments while Investment Grants are *grants*. Results of the analyses should be interpreted accordingly. For grants returns to society i.e. outcomes in terms of added values above what should be expected in the absence of grants, can be meaningfully compared to the amount of resources allocated for grants, such comparisons may no be so straightforward for loans. A further complication arises from the fact that a substantial proportion of the firms have received a mixture of grants and loans.

The Regional Venture Capital Loans carries a calculated risk of losses, which, evidently is the raison d'être for the initiative. Based on historical data, Innovation Norway estimates these losses to approximately 30%[27]. Thus, the governmental costs associated with the loans and the mixture of loans and grants could be considerably different from what is reported in Table 27 below. Moreover, comparisons of returns and costs are limited by the fact that we analyze only about a forth (24%) of the total allocation of funds and only about 34% of the firms (1117 of the 3298 involved over the five year intervention period). The size distribution of the analyzed sample that represents these 24% of total costs deviates from the size distribution of the population[28] in the sense that bigger firms are not included. The consequence of this dissimilarity is that the mean allotments in the analyzed sample differ

[27] There is no documentation concerning the procedures that produced this 30% estimate.
[28] We simply use the term *population* for the entire set of firms the received allocations from Innovation Norway.

from the mean allotments in the population and thus, direct comparisons of estimated means (ATT) and mean allotments do not seem like a good idea. Hence, comparisons have to be carried out by assuming that the estimates of *average effects of treatments on the treated* (ATT) can be aggregated to conform to the total sums of investments (or expenditures/costs) so that this measure can be used to assess the presumed developments of sums of returns in terms of *added value* under the various scenarios.

Table 27 Statistics for the allotments to all firms – Mill. NOK

	Venture Capital Loans				**Investment Grants**				**Both Regional Venture Capital Loans and Investment Capital Loans Part**				**Investment Grants Part**		
Year	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	St.dev	Sum
1990	0,6815	354	1,507	241,25	1,0567	312	8,534	329,70	0,8691	359	1,503	312,01	0,6385	1,042	229,24
1991	0,7112	238	1,389	169,26	0,6267	320	1,188	200,53	0,8228	328	1,146	269,87	0,5980	0,999	196,14
1992	1,2278	228	3,930	279,94	0,7648	325	1,607	248,55	0,9949	269	2,118	267,63	0,6392	1,034	171,95
1993	0,9993	95	2,517	94,93	0,7341	469	2,021	344,28	0,7253	186	0,881	134,91	0,5256	0,627	97,76
1994	1,1483	54	2,793	62,01	0,6613	604	1,139	399,42	1,1574	112	3,447	129,63	0,9506	3,043	106,46
Mean:	*0,9536*				*0,7687*				*0,9139*				*0,6704*		
Sum:				*847,4*				*1522,5*				*1114,0*			*801,6*
N =		*969*				*2030*				*1254*					

Total N (number of firms=4253
Total sum = **4285,5** Mill. NOK

Table 27 and Table 28 show the differences that cause difficulties to the direct use of means for comparisons. Inspection of the columns for means demonstrate considerable difference between the figures for total governmental expenditure (Table 27) and the figures for the proportion of spending actually included in the analysis. Moreover, as seen by comparing Table 27 and Table 28 (*t*-statistics are shown in table C1 in appendix C) the means for the proportion analyzed are invariably smaller than for the total. Even though only *one* of the differences between the means of total governmental expenditures and the means in the analyzed group is statistically significantly different from zero, (table C1 in appendix C) the consequences of small differences between the means in the analyzed group and the total may carry substantial impact upon comparisons of effect estimates and the actual expenditures.

Table 28 Statistics for the allotments to the firms included in the analysis

	Venture Capital Loans				Investment Grants				Both Regional Venture Capital Loans and Investment Capital Loans Part				Investment Grants Part		
Year	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	St.dev	Sum
1990	0,5517	119	0,921	65,65	0,3153	69	0,320	21,75	0,8184	121	1,605	99,03	0,5076	0,606	212,69
1991	0,3365	82	0,583	27,59	0,3576	72	0,331	25,75	0,7391	111	0,693	82,04	0,5490	0,681	61,42
1992	0,5414	45	0,710	24,36	0,4897	63	0,726	30,85	0,7383	71	1,442	52,42	0,5145	1,020	60,94
1993	0,9460	17	2,045	16,08	0,3977	86	0,359	34,20	0,6292	34	0,581	21,39	0,3844	0,392	36,53
1994	0,3770	13	0,350	4,90	0,5626	178	1,046	100,14	0,6389	36	0,643	23,00	0,6534	0,791	13,07
Mean:	*0,5505*				*0,4246*				*0,7128*				*0,5218*		
Sum:				*138,6*				*212,7*				*277,9*			*384,6*
N =		*276*				*468*				*373*					

Total N (number of firms=1117
Total sum = **1013,8** Mill. NOK

The reason for these differences is rooted both in the matching procedures and correction measures carried out after matching. The combined effect of these procedures is that extreme cases, mostly the bigger allotments, are left out, and thus, the distributions of the expenditures included in the analyses differ from the distribution of total expenditures. Gini-indices are considerably lower for the cases included in the analyses than for the total expenditures in all three categories, Regional Venture Capital Loans, Investment Grants and the two combined.

Thus, the lesson to be learned from the difference between the distributions of the analyzed and the total allocation is that the total sums provide a better basis for comparisons than the means of money allocations per firm. We use a two-step procedure for all comparisons; first we compare the sums of expenditures covered by the sample of firms used in the analyses, then we proceed under the assumption that the estimated average treatment effects on the treated (ATT) *are valid for all cases*, not only for the cases included in the analysis. Clearly, such a procedure encompasses an untestable assumption which is convenient since it permits the hypothetical *what if* that facilitates the comparisons of treatment effects and total expenditures.

Estimated average treatment effects on the treated (ATT) (Table 20 to Table 22, page 107 - 108) are point estimates valid only for the year within which they are estimated. Thus, year-by-year comparisons have to be based upon *the future value* of earlier allocated resources. Assignment of money takes place over five years range with a different number of firms entering every year in the treatment period. Thus, future values $fv = pv(1+i)^t$ where i is the discount rate and t is time, have to be calculated separately for each year and added together as shown in e.g. table C4 and C5 in appendix C. The ATT's are made comparable to the total sums by multiplying the estimated averages by the number of cases involved. This simply implies a linear transformation that brings the figures on an equal footing and facilitates the interpretation of means as proxies for sums, qualified by standard errors or confidence intervals. Clearly, these procedures can not provide exact answers, only approximations that give a fairly good idea about the magnitudes involved.

All calculations are carried out based on a discount rate at 7% over the entire period in time analyzed[29]. The revenue effect due to tax financing is assumed to be 21% for all governmental expenditures.

12.3.1 Projected Costs: The meaning of a 30% loss on loans

It is not all that clear what a 30% loss on loans could mean and how it should be interpreted in money terms. Loans assume expected future income. Thus, one interpretation of a 30% loss on loans is that 30% of the expected income is lost over a five-year period. Another equally reasonable understanding is that 30% of expected

[29] The interest on loans shows considerable variation over the period in question. On average the general interest on loans was just above 8%. Thus, a flat discount rate is a source of errors. Figure xx in appendix D provides an overview over the development of the interest rate.

income is lost over a ten-year period. An expected return rate of 7% would in the first case (five years) mean that the interest is reduced by 30%/5 = 6% and thus that the return rate is reduced from 7% to 1%. In the other case (ten years) the return rate is reduced by 30%/10 = 3% and thus the return rate is 4%. These interpretations yield conservative estimates of losses in the sense that we assume that losses remain constant over the entire observation period. The cumulative sum of the present value of losses over the entire observation period will of course exceed 30% when the calculation period gets above five and ten years respectively. Thus the cost estimates for loans are conservative in the sense that they do not understate potential losses.

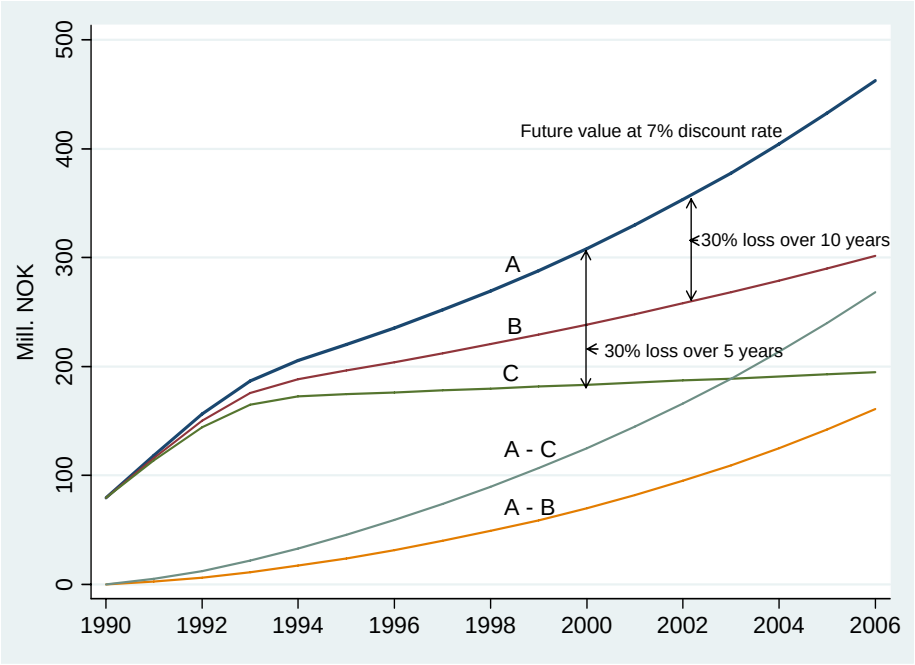


Figure 23 Assumed Losses on Venture Capital Loans, Population

Figure 23 gives a stylized picture of presumed losses over five and ten years based on 1% and 4% interest and the lines A – B and A – C indicate the respective losses. The calculations behind the graphs are shown in table C5 Appendix C. Corre-

sponding interpretations are employed for the analyzed sample of *Regional Capital Loans* and the loan part of the combinations of *Investment Grants* and *Regional Capital Loans.* Calculations are found in Appendix C.

12.3.2 Scenario comparisons of returns from grants and loans

The estimated average effects of treatment on the treated (ATT) are calculated year-by-year as two-period estimates and for the entire period as panel estimates. As such the two-period estimates are considered as future values and the panel estimates as average future values[30]. Thus, the strategy employed for comparisons is to bring costs and returns on an equal footing and compare the presumed results over time. The costs of loans, as shown in Figure 23, are constructed in future value terms, and thus on a scale compatible to the treatment effects on the treated (ATT). The cost of grants is provided in Table 27 and Table 28 as yearly expenditures. To be able to compare the three financial schemes we proceed by emulating a kind of financial analysis, i.e. by finding the most reasonable way compare the present value of governmental expenditures to the present value of future cash flows. By accepting the average effects of treatment on the treated (ATT) as a measure of returns to the firms caused by the financial schemes we can construct an aggregate quantity that can be used as an expression for the "cash flows" to society in terms of added value. These "cash flows" are constructed by simply multiplying the number of firms involved with the figures of average treatment effects (ATT).

[30] Since difference-in-difference estimation is predominantly a cross-sectional technique, discount rate based adjustments *before* estimation is usually not discussed in the literature. It is clear that DiD diminishes but never entirely removes the effects of the nonconformity of the numbers involved in the calculations. The likely effects of using future values for the calculations would be an upward bias while no adjustment would most likely induce a downward bias.

Clearly, such comparisons are *scenarios* in the sense that they go beyond what is actually estimated and involve untestable and dubious assumptions such as the conjecture that every company involved has equal returns from the financial schemes. Scenarios are, by definition, hypothetical descriptions that solely serve the purpose of giving a meaningful interpretation of our findings. Moreover, the "financial analyses" presented here are an integral part of the scenarios and should not be otherwise understood. Clearly, the "cash flows" introduced are not cash flows in the traditional sense; they cannot by reinvested and they do not appear anywhere in the accounts of the individual firm. The "cash flows" are simply the aggregated differences in sums of added value between participants and non-participants where the added value is defined as the sums of the contributions to labor and capital.

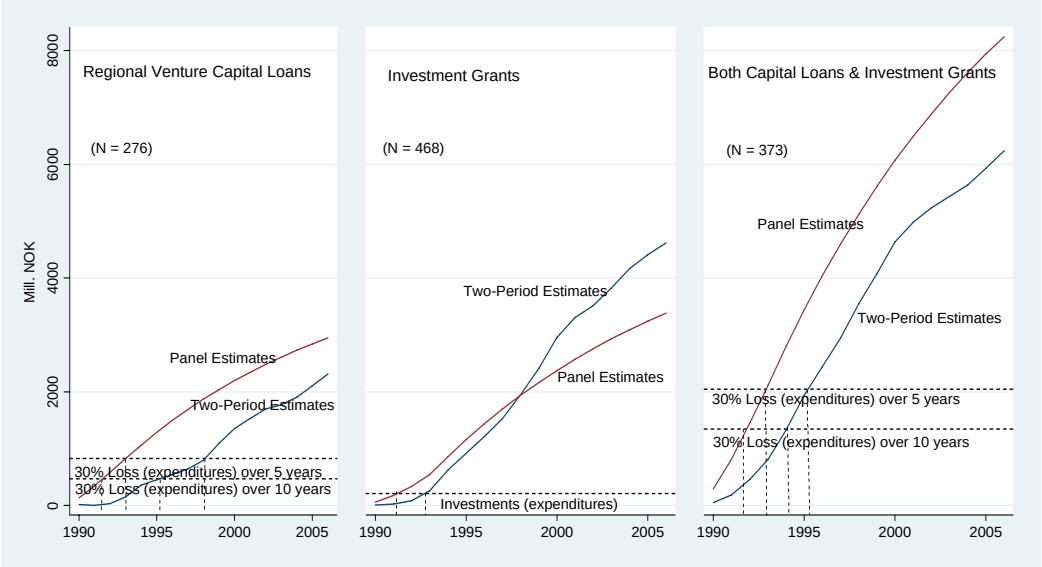


Figure 24 Present value comparisons for financial schemes -Analyzed samples

Figure 24 shows the present values of expenditures (the horizontal dotted lines) and the cumulative present values of the presumed "cash flows" (the solid lines) for the entire period under investigation for Regional Venture Capital Loans, Investment

Grants and the two combined. The three scenarios of Figure 24 indicate a payoff to society well above governmental investments for all three financial schemes. The corresponding scenario for all firms involved in the financial schemes is shown in Figure 25. Note that although Figure 24 and Figure 25 are similar in all respects except for the number of cases used in the multiplication[31] of the ATT, the two figures show some notable differences.

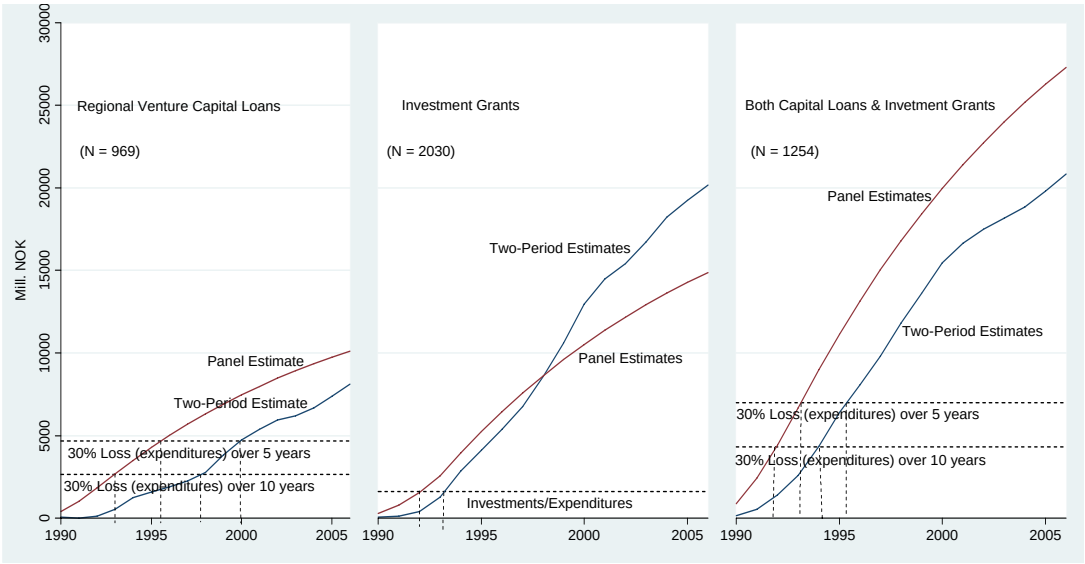


Figure 25 Present value comparisons for financial schemes – Entire Sample

These discrepancies reflect both differences in the proportion of cases that are included in the analyses of the three financial schemes (28.5%, 23% and 29.7% respectively) and the differences in the proportion of expenditures used in the calculations (16.4%, 14% and 25% respectively). Both for the cases included in the analysis (Figure 24, analyzed sample) and for all cases involved (Figure 25, the entire sample) the three scenarios depict overtly optimistic outcomes. The two-period es-

[31] Note the number of cases used for multiplications over the intervention period 1990 to 1994 is less than the N given in the figures, which is the cumulative N over the intervention periods.

timates and the panel estimates produce reasonably similar results and give a fairly good idea about the magnitudes involved.

Although we have quite accurate measures of *costs* since these figures are calculated directly from the allotment decisions for each individual firm, we have to keep in mind that the scenarios reflect the limitations given by the estimates which they are based upon. As shown in Figure 17 page 117 the estimations of the average treatment effects have rather wide confidence intervals. In particular, the 95% confidence intervals for the two-period estimates (the black dotted lines) that are below zero indicate less trustworthy estimates. Since our scenarios are based upon the ATTs (the black solid lines) this has to be kept in mind when judging the reach of the scenarios. On the other hand, it is known that matching procedures tends to produce large and rather inaccurate confidence intervals[32] that cannot easily be re-estimated by bootstrap techniques (Abadie, 2002; Abadie & Imbens, 2006). Furthermore, our scenario calculations based upon two-period estimates do not deviate substantially from those based upon the panel estimates (the red solid lines) where the lower limit of the 95% confidence interval (the red dotted lines) never goes below zero. The close resemblance of the panel estimate based scenarios to the two-period based scenarios is reassuring in the sense that it signals that two entirely different estimation procedures produce comparable scenarios.

[32] Clearly, we could include the confidence intervals when constructing the scenarios and thus make the inexactness of the calculations more explicit. The reasons for not including the confidence intervals are twofold; first and foremost, they are inaccurate, secondly, the confidence intervals for the cumulative figures would add to the inaccuracies since each cumulative step would involve recalculations of the variances involved i.e. the variance in year n plus the variance in year n+1 plus two times the covariance between the measures for year n and n+1 for which we have no estimate. Also, the confidence intervals from such a procedure would become both inaccurate and forbiddingly wide.

12.3.3 Scenario based comparisons of internal rates of return

In line with the interpretative view of the average treatment effects on the treated (ATT) as mean future "cash flows" and governmental expenditures as "investments" we can calculate comparable accounts of the internal rate of return (IRR). The IRR is defined as the discount rate that generates a zero net present value[33]. In finance, IRRs are frequently used for comparing different projects in order to be able to choose the most profitable ones. Here, the sole purpose is to *compare* the projects.

Table 29 Internal Rate of Return for Regional Venture Capital Loans

	All Expenditures (N= 969)				**Analyzed Expenditures (N= 276)**				
	Two-Period Est.		Panel Estimates		Two-Period Est.		Panel Estimates		
	5 years	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	
Analysis									
Number of Cash Flows	17	17	17	17	17	17	17	17	
Period Unit	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	
Undiscounted Sum	11 568.86	13 616.23	12 510.87	14 558.24	3 802.57	4 163.86	4 141.37	4 502.66	
Discount Rate	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	
Net Present Value (NPV)	2 887.47	4 934.84	4 761.61	6 808.98	1 329.36	1 690.65	1 924.99	2 286.28	
Future Value (FV)	9 120.98	15 588.24	15 041.04	21 508.30	4 199.19	5 340.44	6 080.70	7 221.95	*Average:*
Internal Rate of Return (IRR)	12.06 %	19.03 %	17.94 %	30.72 %	17.77 %	25.76 %	29.56 %	48.49 %	25.17 %
Pay-back Period	11.43	9.05	7.01	4.19	9.32	6.56	4.25	2.65	
Present Worth Cost (PWC)	4 725.94	2 678.57	4 686.66	2 639.29	843.92	482.63	830.58	469.29	
Present Worth Revenue (PWR)	7 613.40	7 613.40	9 448.27	9 448.27	2 173.27	2 173.27	2 755.57	2 755.57	
Benefit-Cost Ratio	1.61	2.84	2.02	3.58	2.58	4.5	3.32	5.87	
Present Value Ratio	0.61	1.84	1.02	2.58	1.58	3.5	2.32	4.87	
Reinvestment									
Reinvestment Rate of Return	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	
Modified NPV	2 887.47	4 934.84	4 761.61	6 808.98	1 329.36	1 690.65	1 924.99	2 286.28	
Modified FV	9 120.98	15 588.24	15 041.04	21 508.30	4 199.19	5 340.44	6 080.70	7 221.95	*Average:*
Modified IRR (MIRR)	10.21 %	14.07 %	11.51 %	15.34 %	13.43 %	17.44 %	14.82 %	18.74 %	14.45 %

Table 29 gives an overview of the total sums of presumed "cash flows" and costs over the 17-year period, the net present values (NPV) and the future values (FV) involved for the various scenarios. The benefit-cost ratio reported is simply the ratio of present worth revenue to present worth cost (PWR/PWC) and does not signify anything else, i.e. it does not reflect a cost benefit analysis in the traditional sense. In the same manner the present value ratio is simply NPV/PWC. The expected pay-back period reflects the internal rate of return (IRR) and varies from approximately 12 years for the two-period estimates when all cases for which we have acceptable

[33] Mathematically, IRRs are the roots of an NPV function and in some cases there are multiple roots.

administrative records are accounted for and a 30% loss over a five-year period is assumed, to as little as about 3 years for the panel estimates when only the analyzed cases are considered and we assume that the 30% loss is distributed over a 10-year period. With two different estimation methods, two different ways of understanding the expected losses on loans, and two different samples, the variation across the various scenarios is within a reasonable range. The most misleading feature of IRR calculations is the assumption that all cash flows will be reinvested at the calculated IRR. This is usually not a very realistic assumption and it is particularly unrealistic as applied here where no reinvestment is possible. The modified internal rate of return (MIRR) calculated at 7% reinvestment rate of return may be considered a better candidate as an expression for the interest gain to society due to governmental investment in regional venture capital loans. Thus, regardless of how we understand a 30% loss (whether over 5 or 10 years) or the way we construct the scenario (using all data or the analyzed part only) the dividend exceeds the most likely alternative cost, the foregone interest of 7% (the discount rate used).

Since all scenarios are constructed in a similar manner we can compare any measure provided in the tables showing the financial calculations. The preferred measure for comparisons is, however, the modified internal rate of return (MIRR). A brief glance at the eight scenarios in Table 29 shows that the MIRR varies substantially less across scenarios than does the IRR. The mean of IRR across scenarios is 25.17% with a standard deviation as high as 11.4 while the corresponding number for MIRR is 14.45 and 2.8 respectively.

The scenarios for Investment Grants are simpler in the sense that we compare four situations only; two samples and two estimators. Table 30 shows an average modified rate of return of 24.57% and a moderate variation across the four scenarios.

Table 30 Internal Rate of Return for Investment Grants

	All data (N=2030)		**Analyzed data (N=468)**		
Analysis	*Two-Period Estimate*	*Panel Estimates*	*Two-Period Estimate*	*Panel Estimates*	
Number of Cash Flows	17	17	17	17	
Period Unit	Yearly	Yearly	Yearly	Yearly	
Undiscounted Sum	37 232.82	24 806.12	8 702.49	5 816.83	
Discount Rate	7.00 %	7.00 %	7.00 %	7.00 %	
Net Present Value (NPV)	17 251.74	12 313.05	4 103.90	2 948.23	
Future Value (FV)	54 495.06	38 894.65	12 963.48	9 312.92	*Average:*
Internal Rate of Return (IRR)	44.52 %	51.01 %	57.04 %	70.84 %	55.85 %
Pay-back Period	4.28	3.16	3.81	2.32	
Present Worth Cost (PWC)	1 597.14	1 597.14	214.26	214.24	
Present Worth Revenue (PWR)	18 848.88	13 910.19	4 318.16	3 162.47	
Benefit-Cost Ratio	11.8	8.71	20.15	14.76	
Present Value Ratio	10.8	7.71	19.15	13.76	
Reinvestment					
Reinvestment Rate of Return	7.00 %	7.00 %	7.00 %	7.00 %	
Modified NPV	17 251.74	12 313.05	4 103.90	2 948.23	
Modified FV	54 495.06	38 894.65	12 963.48	9 312.92	*Average:*
Modified IRR (MIRR)	23.72 %	21.53 %	27.68 %	25.36 %	24.57 %

The combination of Regional Venture Capital Loans and Investment Grants involves eight scenarios since the loans part is investigated with two different interpretations of the 30% loss i.e. whether it is spread over 5 or 10 years. As shown in Table 31 below the variation across scenarios for the MIRR is reasonable with an average of 16.18%

Table 31 Internal rate of return for both Loans and Grants combined

	All Expenditures (N= 1254)				Analyzed Expenditures (N= 373)				
	Two-Period Est.		Panel Estimates		Two-Period Est.		Panel Estimates		
	5 years	*10 years*	*5 years*	*10 years*	*5 years*	*10 years*	*5 years*	*10 years*	
Analysis									
Number of Cash Flows	17	17	17	17	17	17	17	17	
Period Unit	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	Yearly	
Undiscounted Sum	30 387.78	33 057.82	39 872.04	42 542.07	9 106.14	9 807.83	12 027.69	12 729.38	
Discount Rate	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	
Net Present Value (NPV)	12 494.82	15 164.86	18 514.30	21 184.34	3 778.64	4 480.33	5 657.48	6 359.17	
Future Value (FV)	39 468.83	47 902.99	58 483.26	66 917.43	11 936.01	14 152.52	17 870.94	20 087.44	*Average:*
Internal Rate of Return (IRR)	21.38 %	30.25 %	30.29 %	44.61 %	21.83 %	29.61 %	31.70 %	44.74 %	31.80 %
Pay-back Period	6.63	5.13	4.35	3.09	6.5	5.19	4.11	2.99	
Present Worth Cost (PWC)	6 986.60	4 316.56	6 986.60	4 316.56	2 049.66	1 347.97	2 049.66	1 347.97	
Present Worth Revenue (PWR)	19 481.42	19 481.42	25 500.90	25 500.90	5 828.30	5 828.30	7 707.14	7 707.14	
Benefit-Cost Ratio	2.79	4.51	3.65	5.91	2.84	4.32	3.76	5.72	
Present Value Ratio	1.79	3.51	2.65	4.91	1.84	3.32	2.76	4.72	
Reinvestment									
Reinvestment Rate of Return	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	7.00 %	
Modified NPV	12 494.82	15 164.86	18 514.30	21 184.34	3 778.64	4 480.33	5 657.48	6 359.17	
Modified FV	39 468.83	47 902.99	58 483.26	66 917.42	11 936.01	14 152.52	17 870.94	20 087.44	*Average:*
Modified IRR (MIRR)	13.65 %	16.92 %	15.47 %	18.78 %	13.78 %	16.62 %	15.67 %	18.56 %	16.18 %

12.4 The FRAM program

The FRAM program consists of several modules that have been evaluated on several occasions in the past. We analyze a smaller part of the program only and our cost estimates are taken from the 1997 evaluation carried out by Nesheim (1997). The FRAM program is a training program aimed at individual competence building for the management of SMEs and no records of governmental expenditures per participating firm are available. Thus, the costs on the supply side are approximated as the Innovation Norway's expenditures associated with this particular segment of the FRAM program. These expenses are estimated by Nesheim (1997) to be approximately NOK 94,000 per participating firm.

Participation is inexpensive, but not free. The program fee is NOK 19,000 per firm. In addition the costs of own efforts and travelling and lodging expenses associated

with participation are incorporated on the demand side. Since the program is mainly organized as meetings at various conference hotels, expenses may vary. A rough estimate of these additional expenses given in Nesheim (1997) is NOK 82,860 before the tax-deductible sum is subtracted.

12.4.1 Scenarios for the FRAM program

A scenario for the FRAM program tries to answer the question "*what would the gain to society be, provided that our estimates of the effect of treatment (ATT) on treated are sound foundations for calculating the "cash flows" from the program?*" and provided that the costs of the program are of an approximately correct magnitude. Figure 26 shows the cumulative present values of aggregate gain, that is, the simple multiplication of 425 times the two-period estimates (Table 23) and the panel estimates (Table 25) assuming that 425 firms participated in the program, where the firm's specific expenses multiplied by number of firms participating per year are subtracted from the "cash flows" in their respective years.

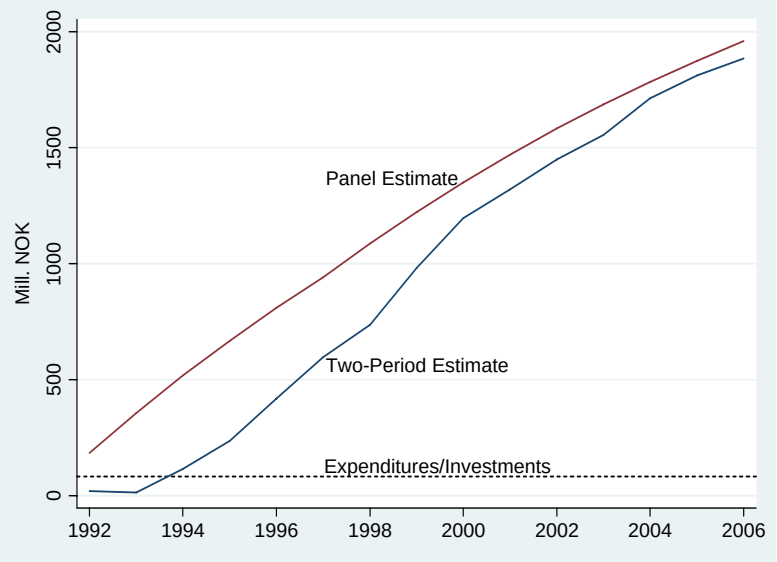


Figure 26 Scenarios of Aggregate Cumulative Present Values – FRAM

Figure 26 indicates that the FRAM program, provided that the scenario assumptions are correct, may produce aggregate outcomes of substantial proportions compared to the inputs in money terms.

Viewed financially, Table 32 indicates a return on investments in the neighborhood of roughly 31% (the modified IRR), which, provided that the scenario assumptions reflect something near the expected, hints that the FRAM program could be an initiative that is easy to defend as a policy for business development.

Table 32 Internal rate of return for the FRAM program - Scenarios

	Two-Period Estimates	Panel Estimates
Analysis	(N=425)	(N=425)
Number of Cash Flows	15	15
Period Unit	Yearly	Yearly
Undiscounted Sum	3 145 154,50	2 999 921,25
Discount Rate	7,00 %	7,00 %
Net Present Value (NPV)	1 679 395,08	1 749 245,45
Future Value (FV)	4 633 504,00	4 826 223,00
Internal Rate of Return (IRR)	74,74 %	221,37 %
Pay-back Period	2,75	0,48
Present Worth Cost (PWC)	89 253,31	83 322,00
Present Worth Revenue (PWR)	1 768 648,39	1 832 567,45
Benefit-Cost Ratio	19,82	21,99
Present Value Ratio	18,82	20,99
Reinvestment		
Reinvestment Rate of Return	7,00 %	7,00 %
Modified NPV	1 679 395,08	1 749 245,45
Modified FV	4 633 503,98	4 826 223,36
Modified IRR (MIRR)	34,12 %	31,48 %

12.5 The Network program

The costs of the Network program is discussed in Econ's evaluation of the program (1998). Admittedly, as stated by the report (Econ, 1998), these figures are not espe-

cially precise. They are, however, the only immediately available source of information on these matters.

Table 33 Costs Associated with the Network Program

Objective *	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*	*1997*	**Total**
Pre-Study	7 500	11 205	7 605	3 869	3 008	773	1 332	35 292
Planning Measures	1 546	6 042	4 937	5 486	4 972	3 292	2 362	28 637
Operating Costs Year 1	4 237	12 523	18 180	16 499	16 071	24 478	13 832	105 820
Operating Costs Year 2		953	2 179	5 467	2 680	8 686	10 134	30 099
Operating Costs Year 3			300		685	990	2 028	4 003
Additional Costs 1)				1 655	5 518	1 517	4 787	13 477
Additional Costs 2)				1 000	1 000	1 000	1 570	4 570
Additional Costs 3)	1 052	1 750	2 180	1 370	1 415	1 925	1 660	11 352
Tourism					3 000	10 000	4 000	17 000
Additional Costs 4)				500	500	500	500	2 000
Additional Costs 5)						420	420	840
Total	14 335	32 473	35 381	35 846	38 849	53 581	42 625	253 090

1) Company specific 2) Competence Development (UNIKE) 3) Counselling
4) Companies without Limits 5) Consultant JFR ****Adapted from Econ (1998)***

The approximate costs of the Network program facilitates a scenario for the Network program that gives us an idea about the magnitudes involved, provided that our two-period estimates (Table 24) and panel estimates (Table 25) offer a justifiable basis for calculations of the presumed "cash flows" from the program.

12.5.1 Scenarios for the Network program

Figure 27 shows the cumulative present values of the estimated effects of treatment on treated (ATT) found in table 24 and table 25 multiplied by 1575, the number of firms involved in the ATT estimations. That is, we try to provide an answer to the question "*what would be the hypothetical gain to society, provided that the estimates of costs and expected "cash flows" are in the neighborhood of what can be expected to be correct*?" provided that exactly 1575 firms participated in the Network program.

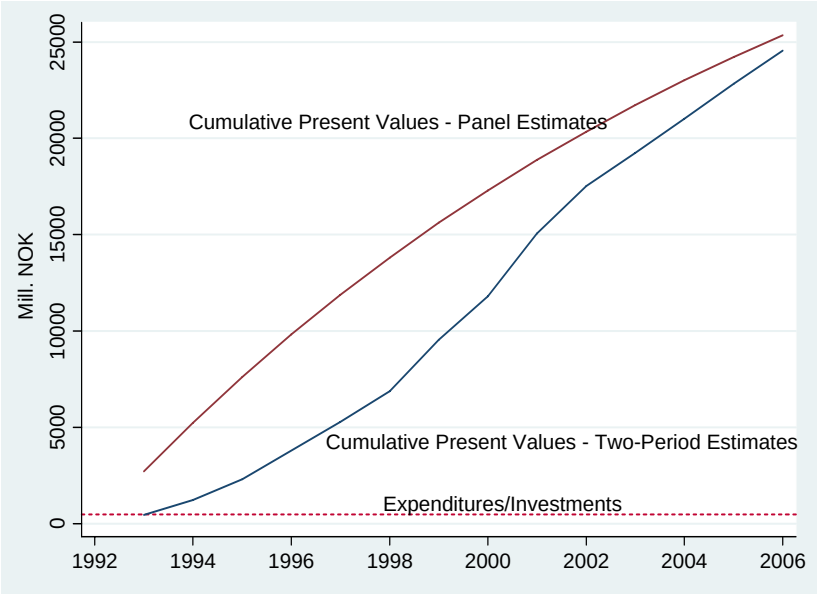


Figure 27 Scenarios of Aggregate Cumulative Present Values – The Network Program

The calculations are based on a cost estimate of 482[34] mill. NOK (Econ, 1998). Figure 27 shows that, viewed on the basis of the premises of the scenario listed above, the Network program could provide substantial gains to society in term of value creation above what could have been expected, provided that the program *not* was carried out.

As shown in Table 34, the above calculations would give an internal rate of return of about 46% over the 14 years considered. Since costs are not very accurately given, this relatively high (modified) internal rate of return should be interpreted with care.

[34] The estimate does in fact concern more firms than the 1575 used in the calculations and we are informed by Econ's report that the estimates are adjusted for revenue effects due to tax financing. It is not a very accurate estimate and it is used here solely for demonstration purposes.

Table 34 Internal rate of return for the Network program - Scenarios

Analysis	Two-Period Estimates (N=1575)	Panel Estimates (N=1575)
Number of Cash Flows	14	14
Period Unit	Yearly	Yearly
Undiscounted Sum	65 133,44	59 225,21
Discount Rate	7,00 %	7,00 %
Net Present Value (NPV)	35 675,43	36 815,68
Future Value (FV)	91 990,31	94 930,48
Internal Rate of Return (IRR)	210,19 %	0,00 %
Pay-back Period	0,71	0,12
Present Worth Cost (PWC)	482	482
Present Worth Revenue (PWR)	36 157,43	37 297,68
Benefit-Cost Ratio	75,02	77,38
Present Value Ratio	74,02	76,38
Reinvestment		
Reinvestment Rate of Return	7,00 %	7,00 %
Modified NPV	35 675,43	36 815,68
Modified FV	91 990,31	94 930,48
Modified IRR (MIRR)	45,65 %	45,98 %

12.6 Comparing the Scenarios – A summing up

The "financial scenario" analysis sets up a link between *inputs* and *outputs;* a relationship between costs and impacts. This relationship is established in terms of governmental expenditures looked upon as *investments* where dividend is assessed by means of its internal rate of return. The investment part is explicit in the sense that expenditures can be traced in public administrative records. The *return part* is more ambiguous since this is payback to the society at large; there is no money that floats back to the investor in the ordinary sense. Thus, the implicit assumption that what is good for business is good for society applies. Moreover, we do not distinguish between the companies that managed to benefit from the various financial schemes and intervention programs and those that did not. Conclusions apply to the overall picture only. That is, we operate at the program level i.e. at the program or scheme as such.

A central purpose of the scenario analyses was the comparisons of the various financial schemes and intervention programs. Table 35 shows the average modified internal rate of return (MIRR) across the different scenarios for all programs and financial schemes investigated. The averages for the financial schemes show a reasonable pattern: Investment Grants has the highest average return, the pure Regional Venture Capital Loan the lowest and the combination of the two is situated in between. Based on the calculations above this is reasonable because our two interpretations of the meaning of a 30% loss (distributed over 5 years/10 years) generate a heavy cost component.

Table 35 Modified Internal Rate of Return (MIRR) – All Interventions

Regional Venture Capital Loans	**Analyzed Data (N= 276)**				**Data for All Expenditures (N= 969)**				***Averages***
	Two-Period Est.		Panel Estimates		Two-Period Est.		Panel Estimates		
	5 years	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	
	13.43 %	17.44 %	14.82 %	18.74 %	10.21 %	14.07 %	11.51 %	15.34 %	*14.45 %*
Investment Grants	**Analyzed data (N=468)**		**All data (N=2030)**						
	Two-Period Estimate	*Panel Estimates*	*Two-Period Estimate*	*Panel Estimates*					
	27.68 %	25.36 %	23.72 %	21.53 %					*24.57 %*
Regional Venture Capital Loans & Investment Grants Combined	**Analyzed Data (N= 373)**				**Data for All Expenditures (N= 1254)**				
	Two-Period Est.		Panel Estimates		Two-Period Est.		Panel Estimates		
	5 years	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	***5 years***	***10 years***	
	13.78 %	16.62 %	15.67 %	18.56 %	13.65 %	16.92 %	15.47 %	18.78 %	*16.18 %*
The FRAM Program	**Analyzed data (N=425)**								
	Two-Period Estimate	*Panel Estimates*							
	34.12 %	31.48 %							*32.80 %*
The Network Program	**Analyzed data (N=1575)**								
	Two-Period Estimate	*Panel Estimates*							
	45.65 %	45.98 %							*45.82 %*

From a substantial point of view viz. based upon the nature of these initiatives these differences make sense: Both loans and grants are based upon thorough judgments of applicants. Although venture capital loans imply a calculated risk, they are loans and a payback is expected. Grants are based upon a different set of considerations

which place a heavier burden upon the applicant in terms of required plans and validation of arguments for the prospects of the investment in question. Thus, we should expect higher thresholds for consent for grants than for loans. Most likely, the differences in the average rate of returns reflect the decision processes behind allotments.

The FRAM- program and the Network program both show very high rates of return. Since the costs that enter the calculations are not very reliable, these rates of return should be interpreted with care.

13. The Balance of Evidence

The use of two different estimation methods, both based upon matched pairs, does not produce unambiguous evidence about the causal effects of financial schemes and intervention programs. It is, however, clear that the *sum* of the bits and pieces as a whole points in the direction of strong causal effects. The purpose of this section is to give a brief assessment of the putative weights of evidence from the various parts of support for the causal claims that are presented in the analyses section. Weight of evidence (WOE) is predominantly practiced within fields like risk assessment and medical testing but is highly relevant for other disciplines as well, especially when a single question is to be answered based on several pieces of evidence. In evaluation research it should be clear that it is advantageous that the author responsible for carrying out the research should also be answerable for the assignments of weight to the various part of evidence presented. It should not be left for the principal to decide what parts of evidence to emphasize.

The term WOE can be defined as "any process used to aggregate information from different lines of evidence to render a conclusion regarding the assessment of causes or/and magnitude of effects". On the road to the overall conclusion a myriad of crucial decisions that affect results are made. Thus, a retrospective assessment and an attempt to weight the importance of the various blocks of evidence provides guidelines about how to read and interpret the findings.

We apply essentially two tools for the assessments of WOE: Analyses of statistical power and sensitivity analyses as discussed in section 8.6. The analysis of statistical

power across the two-period estimates gives us an *approval rate*, defined as the percentage of tests within each category (financial schemes or programs) that have acceptable statistical power. The principles for and use of the matching methods are thoroughly explained in the analysis section. It is the *outcomes* of these principles, choices and methods which are the subjects of scrutiny here. The assessments of *effects sizes* and *power analyses* are explained in the first part of this section. We also revisit the sensitivity analyses since these analyses do not directly affect the outcomes reported in the analyses section. The joint reflections over effect sizes, post-hoc statistical power and sensitivity constitute the WOE.

13.1 Power analyses and Effect Sizes as elements of WOE

All basic results in the analysis section outcomes are simple estimates of means, defined as the outcomes of two kinds of difference-in-differences estimation; two-period estimates of average effect of treatment on the treated (ATT) and the corresponding estimates based upon panel data. In reporting the analyses we have downplayed the use of significance tests[35] and preferred the use of confidence intervals. The logic behind the tables that report one point estimate and its corresponding maximum significance level and confidence interval is, however, a test against the null-hypothesis that the difference between the treatment group and the control group is, in fact, zero. Thus, with one or to stars behind the estimates, or a confidence interval that does not contain zero, we conclude that our alternative hypothesis that the difference between the treatment group and the control group is different from zero. This is the standard test of the difference between two means. Moreover, the standard rhetoric of this procedure usually has a statement that says that the dif-

[35] We have mainly reported the traditional "stars" that signal the results from the statistical tests give probabilities that at least are less than a specified maximum value, usually * = $p \leq .05$ or ** = $p \leq .01$.

ference is *significantly* different from zero. This is deceptive in the sense that it conceals the simple logic of the test of the difference between two means and that the meaning of the term *significant* has strong connotations in ordinary language use and a more specific and less weighty meaning in statistics. The observed value $ATT = \overline{X_T} - \overline{X_C}$ is only one of an infinite number of sample-mean differences that constitutes a population. Thus, the null hypothesis that we state from the outset in order to compare the two sample means is in fact a statement about the mean of the population of sample-mean differences. Moreover, we assume that the transformed score $t = \left(\overline{X}_T - \overline{X}_C\right) / S_{X1,X2}$ follows the bell-shaped *t*-distribution and the H_0, the null-hypothesis is rejected if the observed value of the transformed score is found at the extreme ends of the *t*-distribution, the top or bottom 5% or 1%. Thus, it is more likely to have a score that comes from a population with a non-zero sample mean than it is to have a score that comes from a population with a zero mean. The likelihood of the latter outcome is, by the definition of the test only 5% or 1%. Clearly, the logic of testing against the null-hypothesis reflects that the researcher believes that the H_0 is false and intends to reject it. In the case of the presumed causal effects of the financial schemes and intervention programs we study, it is an almost obvious implicit assumption that our point of departure is that we believe that a causal effect is present and observable in the administrative records which constitute our data. It is important to avoid rejecting a true null-hypothesis (called Type I error) and we have controlled this by setting the probability of committing this error at or below .01 or .05. i.e. the so-called significance level α. However, we cannot rule out the possibility that the null hypothesis is true and thus, we also have to consider the probability of *not* rejecting H_0 (the so-called Type II error). If this probability (called β) is substantial we have made it too easy for ourselves and in fact

constructed a test that gives little chance of concluding that the financial schemes or intervention programs have an effect when they in fact have an effect. Thus, if we want the rejection of the null-hypothesis to serve as evidence for our research hypotheses we have to have a small β or, equivalently to have its complement (1- β), *the power of the test* sufficiently large. The statistical power of a significance test is the long-term probability of rejecting H_0 given the effect size, the α-level and the sample size (Cohen, 1992). Magnitudes close to .5 is the equivalent of tossing a coin[36]. The preferable level of precision suggested by Cohen (1992) is .80 (β=.20) for α = .05. This is simply a convention proposed for general use. We have calculated the observed power by means to the Stata-program sampsi (Mander, 2006). As used here the calculations imply that magnitudes much below .80, provided a sufficiently large effect size, should indicate that the number of administrative records available for estimations are insufficient and the estimate in question should count less in the WOE since it does not qualify as evidence.

The second element of WOE, the effect size, is of importance for a different reason. Clearly, confidence intervals as reported (e.g. in Table 20 to Table 24) offer more information than p-values; it is easy to see that an interval that does contain zero implies that the estimate is significantly different from zero. It is, however, not that obvious how we should compare two estimates of equal magnitude with very different confidence intervals; two equal magnitudes with equal confidence intervals but with a very different number of cases involved in the estimations or various variations in the combinations of these elements: number of cases involved, width of confidence intervals and magnitude of estimates. There are many ways of making estimates comparable across different analyses; the most common one is the *effect size*

[36] According to Cohen (1992) power close .5 to detect medium effect sizes is not unusual in published research.

which basically involves comparing the various estimates in comparable standard deviation units. The effect size use here is Cohen's *d* which is calculated as $d = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{(n_T - 1)S_T^2 + (n_C - 1)S_C^2}{n_T + n_C}}}$ and provides for us a metric that makes the various estimates of ATT comparable across tests. The interpretation of Cohen's *d* is, according to Cohen, as follows (Cohen, 1988): less than .15 is a negligible effect, .15 to .40 is a small effect, .40 to .75 is a medium effect and above .75 is a large effect. Cohen's *d* is coarse measure of effects which provides an accepted standard for comparisons across studies.

13.1.1 The Relationship between Power and Effect Size

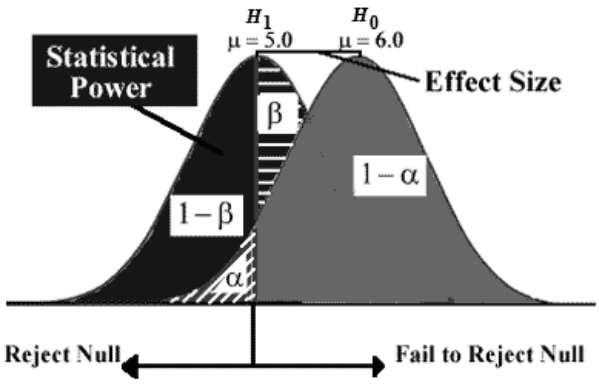
Some researchers may claim that post-hoc power analysis is meaningless and akin to meta-analysis of your own research. Clearly, if a finding is insignificant according to a standard statistical test it also has insufficient power. Statistical power and effect size are for the *planning* of a study and have no meaning after having completed the study. Such contentions imply a simplified view of the research process and assume that you are able to specify the desirable effect size and power in advance and, hence, your sole concern is the size of your random sample.

In this particular study sample size varies substantially across tests and even within the individual test one part of a difference that constitutes the difference-in-differences may vary. Furthermore, sample sizes are fixed by available administrative records over the observation period. It is clear that you need a large sample to be able to detect a small difference between the treatment group and the control group and vice versa that a large difference can be detected in a relatively small sample.

Moreover, with sample size and significance level fixed, the power of a test increases with the magnitude of the effect size.

Thus, as shown in Figure 28 the smaller the effects size (the magnitude in standard deviation units, e.g. Cohen's *d*) given a fixed level of preferred power (e.g., 0.80 as recommended by Cohen (1988) the higher the number of cases necessary for detecting the effect. In observational studies it is not always possible to increase the number of cases. Hence, post-hoc observed power should be reported in order to demonstrate to what extent a causal effect cannot be detected simply because the number of available administrative records is insufficient to achieve statistical power (1-β) that gives a reasonable[37] trade-off between type I error (the probability of rejecting a true null hypothesis) and type II error (probability of rejecting a false null hypothesis).

Figure 28 Statistical Power and Effect Size – One-sided test

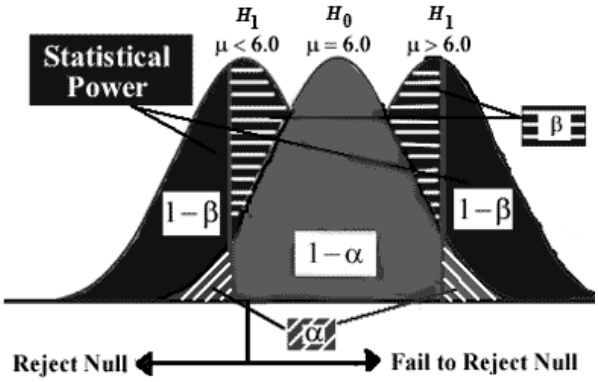


The stylized symmetric distribution above shows that maximizing power (1-β) depends on the α-level and the choice between one-sided or two-sided test. As shown in Figure 29 below, the use of two-sided tests complicates matters and makes it harder to achieve acceptable power of a test. In general a two-sided test decreases

[37] This is not to say that sample size should be increased to make it possible to detect an insubstantial magnitude. What is and what is not a substantial magnitude depends on what you are looking for and is not an issue for statistics.

power, given fixed effect size and sample size. Therefore, for any given hypothesis test, ceteris paribus, there is more possible Type II error and less statistical power in the test. All tests conducted here are one-sided tests and the α-level is the conventional α = .05.

Figure 29 Statistical Power and Effect Size – Two-sided test



With post-hoc observed power we can evaluate the success or failure of our tests and also set some yardsticks with respect to the necessary requirements for future research. We can give a rough calculation of the minimum detectable effect size for a given power κ, a given significance level α and sample size if we know the proportion assigned to treatment and the standard deviation estimates or we can rearrange the equation below to reflect over the number of cases that would be needed in order to estimate the effect sizes that we actually have.

$$MDE = (t_{(1-\kappa)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} * \sqrt{\frac{\delta^2}{N}}$$

Clearly, the role of power and effect sizes in observational studies has to be different from the role as a tool for planning a study. We think that, after all, since we apply parametric statistics the power of tests should be reported as integral part of study. As used here, it is also well-suited as a tool for evaluating the evidence produced.

13.2 The Comparison of Power and Effect Sizes

Figure 30 shows the observed pos-hoc power and effect size for Regional Venture Capital Loans and Investment Grants. The relevant range of the observation period is the post-treatment period, i.e. the years after 1994. It is easy to see from the figures that small effect sizes give low power. Clearly, pos-hoc power analysis is hindsight. With a fixed number of cases, however, it provides important information about the worth of the evidence obtained. The estimates for Regional Venture Capital Loans have small to medium effects over the post-treatment period. The power of the tests is acceptable for 3 of the 13 years of the post-treatment period, yielding an approval rate of approximately 23%. Figure 30 also shows a considerable variation in both effect sizes and the power of the tests over the observation period. Compared to the fluctuations shown in Figure 17, page 117 that exhibits the estimations in money terms, the transformation into effect sizes reflects both the width of the confidence intervals as estimated by the methods suggested by Gardner & Altman (1989) and the fact that the effect sizes are calculated from the two different means that are uses for estimating ATT. These two different means may, due to attrition within years, have different sample sizes. The figures for Investment Grants show only small effects, but 5 of the 13 tests over the post-treatment period have acceptable power thus giving an approval rate of 38%.

Effect sizes and power for both Regional Venture Capital Loans & Investment Grants combined is shown in Figure 31. The left side of the figure shows medium-sized effects in the first years after the end of the treatment period and that the effects level off to a small effect for the remaining of the post-treatment period. The

statistical power of the tests is satisfactory for most of the post-treatment period until 2002 where a sharp decline in power occurs.

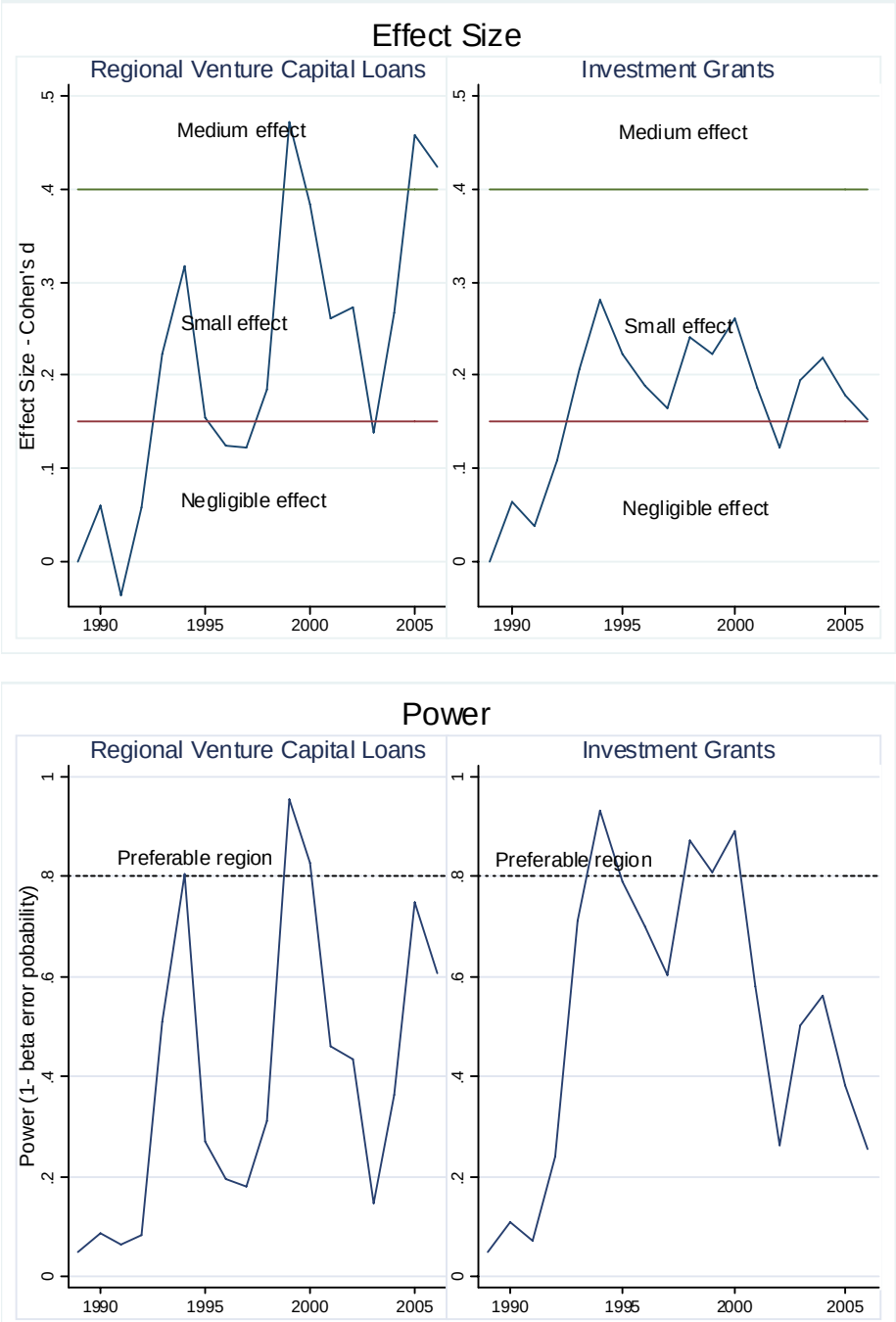


Figure 30 Effect Size and Power for Financial Schemes – part 1

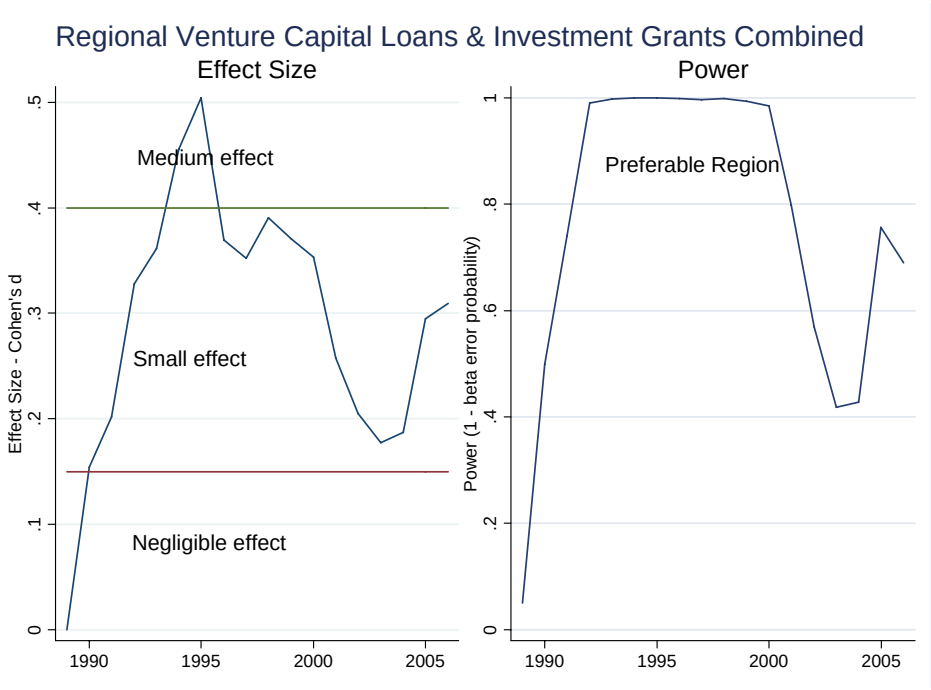


Figure 31 Effect Size and Power for Financial Schemes – part 2

The effect sizes for the FRAM program (Figure 32) are small or negligible over the post-treatment period and essentially of considerable magnitude only in the period right after the treatment period[38]. Only 2 of 10 tests have acceptable statistical power, giving us an approximate approval rate of 20%.

In comparison, the Network (Figure 33) program have small effects and acceptable over the entire post-treatment period, thus yielding an approval rate of 100%. Looking back to Table 24 page 109 it is fairly clear that a part of this outcome can be attributed to the sample size available for the Network program and the surprisingly large magnitude of the estimates.

[38] Note the post-treatment period for the FRAM program starts in 1998 whereas the corresponding years for the financial schemes and the Network program is 1995 and 197 respectively. Note also that we do not take into consideration power and effect sizes within the treatment period.

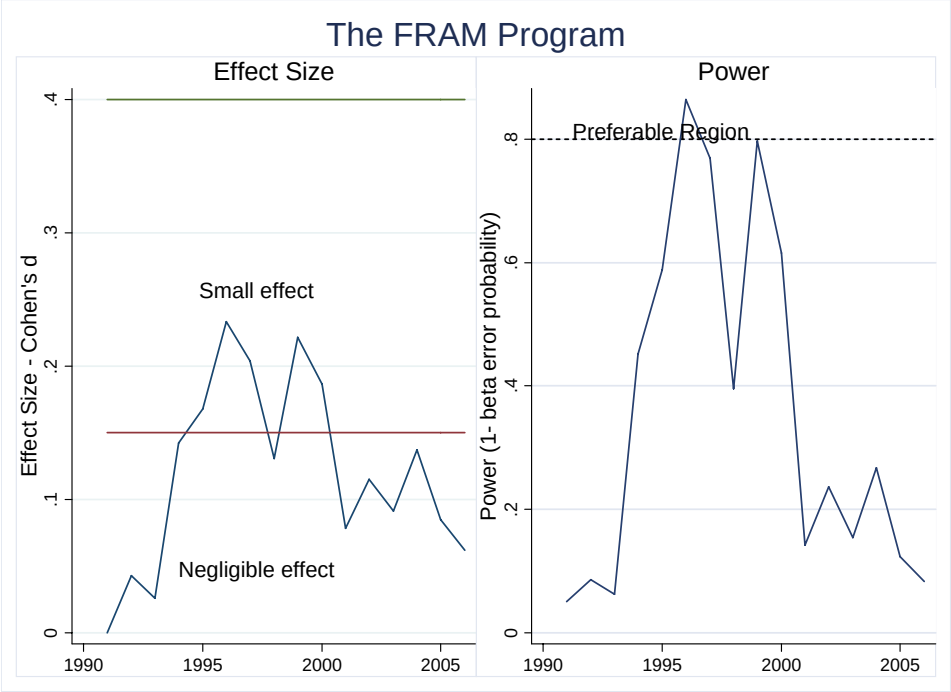


Figure 32 Effect Size and Power for the FRAM program

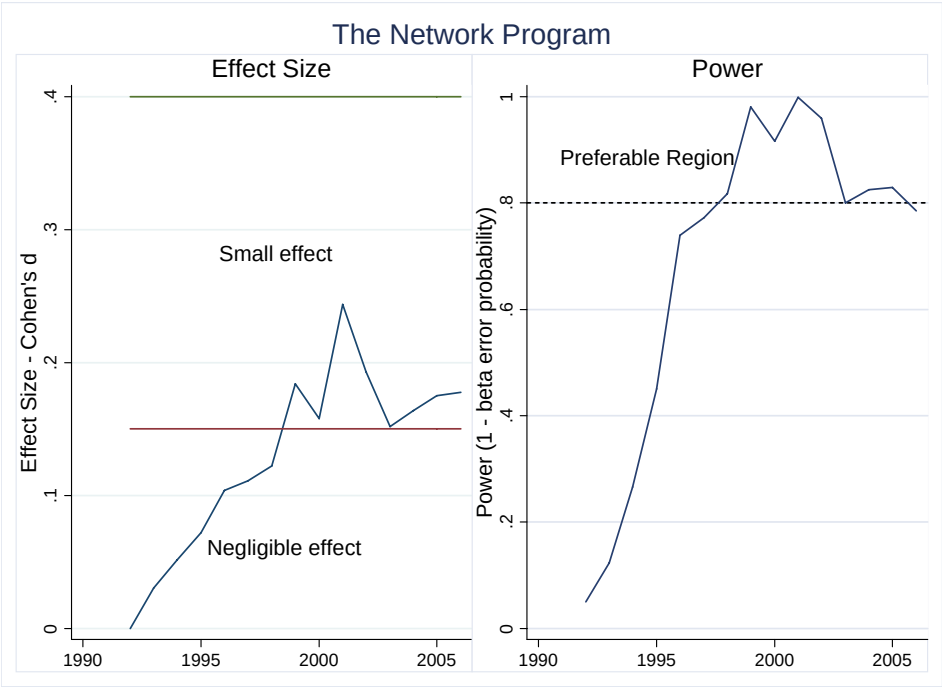


Figure 33 Effect Size and Power for the Network Program

Throughout the study we have compared the two-period estimates against the estimates from the panel data analysis. The goal of the panel data analysis is the same as for the two-period analysis; we want to estimate the average difference-in differences between the treatment groups and the control groups (ATT). Since the panel data estimates use more of the available information than the two-period estimates and are considered to be less vulnerable to variation caused by outside factors, we consider these estimates as generally more reliable. However, in terms of explained variance in the panel regression models, these models are known to have low statistical power[39]. The only parts of the models considered here, however, are the estimates of ATT. These estimates are evaluated as ordinary estimated means and power and effect sizes are compared against the null hypotheses that the difference between the treatment group and the control group is zero. We have demonstrated in the analysis section that the means estimated by panel data methods are within a reasonable range of the two-period estimates. Estimation methods are, however, very different. Thus, comparing the panel data effect sizes to the two-period effect-sizes reveals that the effect sizes from the panel estimates are very small, or what Cohen (1988) terms negligible. Due to the large number of cases when the entire samples are used for estimations, the statistical power of the tests is satisfactory for all categories. The problem can be demonstrated by using the panel estimate for the effects of receiving Regional Venture Capital Loans for which we estimated to be 1.125 mill NOK with a confidence interval ranging from the lower limit of .538 mill NOK to 1.712 mill NOK at a $p \leq .05$. As shown in Figure 34, to be able to detect an ATT of this magnitude (i.e. compared against $H_0 = 0$) takes a considerable sample size. At a sample size of 1000 statistical power for a mean of this magnitude would

[39] Details from the panel data estimations are reported in appendix B.

be around .5, the equivalent of tossing a coin. As seen from the figure a sample size around 3500 would be the minimum required to get a power near or above .80. Since all effect sizes for the panel estimates are in the neighborhood of the effect size for Regional Venture Capital Loans we do not find it necessary to show the graphs, which would look almost the same.

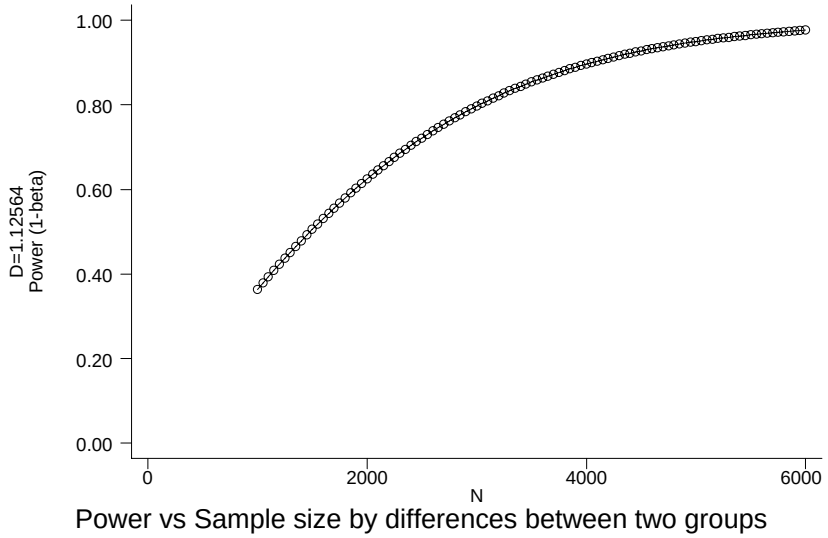


Figure 34 Power and necessary sample sizes (Reg. Vent. Capital Loans)

The actual effect sizes and post-hoc power for the panel estimates are shown in Table 36. As can be seen from the table; despite the small effect sizes all tests have acceptable statistical power due to the large sample sizes.

Table 36 Effect size and Statistical Power for the Panel data estimates

	Effect Size	**Post-hoc Power**	**N of obs.**
Regional Venture Capital Loans	0.051	0.98	5442
Investment Grants	0.028	0.91	11911
Both Loans & Grants	0.057	0.99	12413
The FRAM program	0.030	0.92	10963
The Network Program	0.050	1.00	38057

Thus, we accept the panel estimates as reliable estimates of ATT

13.3 WOE – Principles and summing up

The principles for WOE used here are straightforward: we compare to what extent we have succeeded in detecting the effects that show up in the data, given the fixed sample sizes that we have collected from the available administrative records. All effects we consider are ATTs, effects of treatment on the treatment, and all effects are based upon matching on *observables.* Thus, we also have to take into consideration to what extent it is likely that influential unobserved variables have sizeable influence. The sum of the evaluation of these two pieces of information decides the weight we will assign to the various estimates. Clearly, the weights of evidence (WOE) do not provide any clear-cut answers. As an assessment of the procedures we have carried out, it does, however, give us a way to range the three categories of financial schemes and the two intervention programs according to the degree of trustworthiness. That is, for those categories where we have relatively high confidence in the estimates we believe that causal effects and the approximate size of the effects in money terms, are substantiated. Where we cannot trust the estimates we have to conclude that results are inconclusive.

Table 37 Elements of WOE – Statistical Power and Sensitivities

	Percentage of estimates above the Power of .80 in the Post-treatment period	Γ- values at p –critical ≤ 0.05 for initially matched pairs
Regional Venture Capital Loans	23%	1.25
Investment Grants	38%	1.60
Capital Loans & Grants Combined	62%	1.65
The FRAM Program	20%	1.10
The Network Program	100%	1.20

As shown in Table 37, for the two-period estimates the lowest ranking category has the FRAM program. The combination of columns 1 and 2 tells us that only 20% of the tests have sufficient power and that the effect of a hidden variable does not have

to be large to render our conclusions from the comparisons of matched pairs unsteady. The fact that matched pairs might exceed *p*-critical at the conventional .05 level for the Hodges-Lehman point estimates when the assigned odds of receiving treatment (Γ) is changed from 1 to 1.10 indicates that if matched pairs differ with as little as 10% in unobservable characteristic (see Table 17 page 101) the confidence interval for the difference between the two means that constitute the ATT may include zero. Note that this does not necessarily mean that the CIA assumption is violated; it simply implies that if unobserved heterogeneity exists it *has the potential* of affecting our estimates. Thus, the estimates for the FRAM program are vulnerable and easily affected by unobserved factors. The result that *p*-critical is exceeded at Γ set to 1.10 is, however, no proof of the actual existence of influential hidden bias.

For the Network program the situation is less clear. We have the odd combination of relatively low gammas and a 100% approval rate on the statistical tests. Matched pairs ought not to differ by more than 20% due to unobservable characteristics if the estimated ATTs should be taken as robust to the influence of hidden factors. Hence the magnitude of a relevant unobserved variable does not have to be large in order to affect estimates. It is also worth noting that one of the reasons for the large sample size is that we have incomplete information regarding the timing of events; we do not know what year firms participated in the Network program, we simply know that firms joined the program in one year or more between 1992 and 1997. This lack of information gives us the larger sample size which in turn accounts for the high statistical power. The sum of these shortcomings, the relatively high vulnerability to unobserved heterogeneity and imprecise information concerning the timing of

treatment has to be emphasized as factors that make the estimates for the Network program less trustworthy.

The analysis of *the financial schemes* demonstrates evidence that seems more reliable. First and foremost, we have better data. While we analyze only a small fraction of the firms that have participated in the FRAM program and have incomplete timing of participation for the Network program, we have exact timing of participation and fairly complete records for the number of participants for the financial schemes. Furthermore, we have information about the size of allotments which makes it possible to remove obvious outliers such as loans or investment grants that stand out because they are extremely large compared to the common participants.

The weakest part of the analyses is for the Regional Venture Capital Loans where only 23% of the tests have a post-hoc power near or above .80. Moreover, the sensitivity analyses indicates that the estimates for ATT may be unstable if an influential hidden factor that may cause matched pairs to differ with more than 25% is present.

The Investment Grants have a relatively low approval rate, 38% of the tests have post-hoc power above the recommended .80, whereas matched pairs might differ up to 65% in unobservable characteristics and still have significance values below .05 for the ATT estimates.

The combination of Regional Venture Capital Loans & Investment Grant has a high approval rate; 62% of the tests have a statistical power of .80 or above. Moreover,

the assigned gamma value at *p*-critical is as high as 1.65. Thus, we consider the two-period estimates for this category to be highly trustworthy in comparisons with the other four categories.

Table 38 sums up the rankings of the estimation results for the financial schemes and intervention programs based on the above reflection over to what extent we have succeeded with respect to simultaneously having both acceptable power of the tests and estimates based upon matched pairs that are robust towards the effects of potential unobserved heterogeneity.

Table 38 WOE – Rankings of the Financial Schemes and Programs

	WOE Ranks of Trustworthiness
Regional Venture Capital Loans	3
Investment Grants	4
Capital Loans & Grants Combined	5
The FRAM Program	1
The Network Program	2

1 is the lowest score, 5 the highest

The WOE ranks the FRAM program at the bottom and the combination of Capital Loans and Investment Grants at the top[40]. The implications of these ranking is *not* that we consider the top ranking estimates as near perfect and the bottom ranking estimates as worthless. All estimates assume the unconfoundedness assumption (CIA) holds and the sensitivity analyses give us an indicator for to what extent this is true. If the data suggest that if it would take the presence of unobserved covariate that changes the odds of participation by a considerable factor, say 10, in order to increase to influence the *p*-value we could conclude that our results are robust. For e.g. the FRAM program this factor is as low as 1.10 and we have to conclude that

[40] Note that the results from the panel data analyses do not enter the rankings in Table 38

the two-period estimates are not very robust since all estimates, including the panel data estimates rely on the quality of the initially matched pairs. We have gammas ranging from 1.10 to 1.65. It is clear that 1.10 is a relatively small value, but it is less clear to what extent 1.65 is considerable. It is rare, however, that values above 1.6 are found in social science studies (Becker et al., 2007). Note also that Rosenbaum bounds are worst-case scenarios. Thus, a small value of Γ as the 1.10 for FRAM, does not mean that unobserved is present and that there is no effect of treatment on the outcome variable, it simply means that we have to interpreted results with care since they are not very robust to overlooked confounders.

Thus, drawing conclusions from the analyses should be done with the rankings in Table 38 in mind. This means that all estimates should be considered as numbers that indicate the approximate magnitudes of the causal effects of treatment on the treated, but some estimates are more robust than others. For say, the FRAM program, we can conclude that we have some estimates of causal effects of magnitudes that, conditional on an agreement that we have chosen relevant observables for the matching procedure, most likely would be within this range, provided that no considerable confounder is present. By the same token, we can conclude that it is likely that we have found positive measurable causal effects of the Network program. For these particular estimates we have some doubts concerning the magnitudes; both because we believe that the inexact timing of treatment have caused to many records to be included and thus that we believe that in this case, the power of the tests is in fact inflated and the estimates are relatively vulnerable to unobserved heterogeneity.

For the financial schemes, and, in particular for the combination of both Regional Venture Capital Loans and Investment Grants we believe we have substantiated causal effects of treatment and that the estimated magnitudes are the best estimates, provided that we have chosen relevant observables for the matching procedures.

In terms of the hypotheses put forward on page 18 we have found support for the suggested rankings of yields from the five initiatives. The WOE analysis has, however, demonstrated that hypothesis testing against the null is seductive in the sense that it implies that a hypothesis that is supported by mere coincidence is accepted as evidence. Moreover, we have also demonstrated that for analyses based on matched pairs, it is of importance to evaluate the likelihood that the estimates are affected by unmeasured factors.

14. Discussion

14.1 Reply to invited commentators

We have received a number of comments Dr. Arvid Raknerud and Dr. Morten Henningsen (2009) (hereafter R&H), Statistics Norway. The first section of the discussion attempts to provide an answer to their questions and suggestions.

14.1.1 Propensity score procedures and covariate matching

Their first remark concerns the use of covariate matching (CVM) viz. we have followed a step-by-step procedure based on propensity score matching (PSM) while the actual estimations are covariate matching[41], or matching on $\boldsymbol{X}$ using Heckman's (1998a) terminology. Thus, the question is whether covariate matching and propensity score matching are equivalent with respect to assumptions and implications. In particular, what are the consequences of the fact that there is no region of common support within CVM?

Rosenbaum and Rubin (1983) have shown that matching on the propensity score $p(X) = prob(D = 1 \mid X)$ implies that $D \perp X \mid p(X)$ i.e., that assignment to treatment is independent of *X*, conditional on *p(X)* and that $Y_1, Y_0 \perp D \mid (X)$ i.e., the outcomes on the effect-variables are independent of the assignment to treatment given the covariate vector *X* and that $Y_1, Y_0 \perp D \mid p(X)$ the outcome on the effect variables are independent of the assignment to treatment, given equal propensity scores. Propensity score matching has become popular and is widely used since it reduces the

[41] Part of the incitement to switch from propensity matching to covariate matching came from the good advice from Dr. Raknerud who generously commented a PCM-based report in 2007.

matching problem from many dimensions to one and thus eliminates the curse of dimensionality. Intuitively CVM is more reliable than PSM since PSM implies a reduction of the matching problem, nearly a shortcut. A closer inspection reveals that major differences are situated in the matching logic itself and that these differences have consequences. The matching logic will be discussed first.

Provided that the *unconfoundedness assumption* $Y_1, Y_0 \perp D \mid p(X)$ and the *common support assumption* $0 < prob(D = 1 \mid p(X) < 1$ are satisfied, selection bias is only due to the observables *X*. In Rosenbaum and Rubin's (1983) terminology, treatment assignment is *strongly ignorable* given these assumptions. This is the CIA, the "conditional independence assumption". The proof of that this assumption, when satisfied for *X* is also fulfilled for *p(X)* is the so-called balancing property that:

Prob(X|D=1, p(X) =p) = prob(X|D=0, p(X) =p) = prob(X|p).

The crucial point here is the common support assumption, which is distinct within the PSM framework. As shown in Figure 5 page 62, it is straightforward to decide whether a propensity score is inside or outside the range of the probability of being assigned to the treatment group. A corresponding region or any other condition that logically restricts the acceptable distance between the treated sample and the comparison sample, $\| X_1 - X_0 \|$, is not available for CVM. The distance is expressed as a metric that is the outcome of matrices of the form $(\boldsymbol{X}_{D=1} - \boldsymbol{X}_{D=0})\boldsymbol{W}(\boldsymbol{X}_{D=1} - \boldsymbol{X}_{D=0})^T$ where $\boldsymbol{W}$ is a weight matrix and $\boldsymbol{X}$ is the matrices of the covariates[42]. This metric depends on the chosen weight matrix but is scale-invariant and facilitates the inspections of outliers. The region of common support in PSM represents a minimum criterion. Within the region, further checking of the balancing properties of *X* and

[42] The program **nnmatch** has several options for weight matrices; the Mahalalobnis weight matrix is the inverse of the covariance matrix of $\boldsymbol{X}$ whereas the one we have used, the Abadie-Imbens weight matrix is a diagonal matrix with the inverse of the variances as its elements.

the re-evaluation of the chosen matching algorithm is usually necessary. The balancing properties of the CVM have to be evaluated in the same manner. For PSM better balancing properties should mean better match between the propensity score of the treatment sample and the comparison sample, for CVM improved balancing properties should decrease the magnitude of the distance metric and all covariates should balance over the treatment and comparisons samples. Analyses of the balancing properties are outlined in section 8 page 88.

Noticeably, perfect match on *X* is better than perfect match on *p(X)* since with a lengthy *X* more than one single unit in both the treatment sample and the comparison sample may have exactly the same propensity score while the scores on the individual variables in *X* may differ within the same unit. Thus, in terms of the potential for exactness CVM is superior to PSM. Collapsing the data into one variable is not free of costs. Clearly, PSM is more convenient than CVM and what is shown by Rosenbaum and Rubin (1983) is in fact that matching on the empirical propensity score (which may differ from the true propensity score) may work (almost) just as well as matching on each single variable. The widespread use of PSM is not an argument in favor of the method. In fact the more technical part of the literature tends to point out CVM as preferable for a number of reasons, especially because recent research on PSM has found deficiencies in term of its capacity to fully account for the information[43] in *X*. Also, the prevalent reliance upon statistical tests such as the *t-test* for the assessment of the balancing properties (King & Zeng, 2006; King et al., 2007) may be flawed. When the number of cases in the treatment sam-

[43] First and foremost, PSM introduces *model dependency* since the propensity score is modeled by means of one or other regression type model, usually a logistic regression model. Also, the variable that is closely related to the treatment-control dichotomy accounts for a higher proportion of the propensity score than the other variables in the model thus creating the false impression that two units are equal along many dimensions when in fact, the success of the balancing tests may be purely accidental.

ple is gradually reduced during the matching procedure in order to improve the balancing properties it is easy to forget that the process that reduces the number of cases is not a random process and that the number of cases in the comparison sample does not constitute a population from which cases are randomly drawn. Balancing tests based upon the means may also ignore differences in the shapes of the distributions for the treatment sample and the comparison sample. In this paper we have predominantly used *qq*-plots to evaluate the full empirical distributions of the continuous variable and percentage hit (i.e. identical scores) for the discrete variable to avoid biased conclusions concerning the balancing properties. The decision to use CVM was partly practical; **nnmatch** has some advantageous properties that make it suitable; in particular the option for exact match[44] which made it possible to analyze programs that go over more than one year. The alternative would be to analyze each year separately and manually check for overlaps in the year that follow. With respect to differences in terms of statistical properties it is recently shown analytically that the semi-parametric variance bounds for matching on X is lower than for propensity score matching (Frölich, 2007b). Thus we consider CVM to be a good solution in this particular setting.

14.1.2 Inclusion of pre-treatment level of Y in X

Raknerud and Henningsen (2009) suggest that the lagged pre-treatment level of the endogenous variable Y should be included in X the vector of covariates. The reason for their suggestion can be found in Figure 4 page 57 which shows the common trend in the development of added value over time. Their interpretation of the figure

[44] The exact match is not part of the matching algorithm. It is simply an option for the simultaneous operation on several subsets of the files involved.

is that $Cov(D_i, Y_{i,t-1}) > 0$ and $Cov(\Delta Y_t, Y_{i,t-1}) > 0$ i.e. that treatment assignment is *not* independent of the pre-treatment level of the dependent variable and that growth rate in *Y* over the period figure depends on the level of Y_{t-1} so that the larger companies have a higher growth rate. The alleged consequence is an overestimation of effects.

Firstly, Figure 4 is included to demonstrate the common trend assumption, which has to be satisfied for the use of difference-in-differences analyses. The figure shows data matched on the values of the five variables chosen for the *X*-vector *in the year prior* to treatment. The figure shows that the difference between the mean values of *Y*, added value is zero in the first year prior to treatment for the Financial Schemes and the FRAM program, but deviates from zero for the Network program data. Thus, the figure indicates that the matching procedure has produced an equal starting point for the dependent variable even when the level of the dependent variable is not included. This is not to the same extent the case for the Network program.

Second, our main argument for *not* including the dependent variable among the matching covariates is that it would violate the conditional independence assumption, $Y_1, Y_0 \perp D \mid X$. We will not, however, completely reject the suggestion of including the level of Y in the covariate vector e.g. by means of coarse blocking on levels of *Y*. This is an issue that is discussed in the literature e.g. (Black & Smith, 2004). Recent research (Glynn & Quinn, 2008) also implies that there is a need for rethinking the principles and underlying causal mechanisms when choosing the conditioning covariates. In this study the decisions concerning *X* are clearly limited by what is obtainable from the sets of administrative records at hand. We are open for further discussions on the proper selection of appropriate conditioning variables.

The problem suggested by R&H that may arise for the reason that $Cov(\Delta Y_t, Y_{i,t-1}) > 0$ i.e. because there is a positive level dependent growth rate is of less concern for our two-period estimates since the ATT is not measured as the direct difference between the added value in the treatment sample and the comparison sample. ATT is throughout this study always measured as the difference-in-differences, which implies that the difference in levels between the two groups cancels out. This property of the DiD estimator also reduces the problems implied for the Network program because of the difference in *Y* in the pre-treatment year. The problem that may occur in the panel analyses because $Cov(\Delta Y_t, Y_{i,t-1}) > 0$ is comprehensively discussed by Bertrand, Duflo and Mullainathan (2004). Their main concern is that the nominal growth phenomenon invariably produces positive results when difference-in-differences is employed for natural experiments. We believe that the matching procedures employed in this study to a substantial extent reduce this problem[45].

14.1.3 Unobserved changes in the dependent variable over time

The dependent variable Added Value is defined as the sum of labor costs and net operating result. The proper question from Raknerud and Henningsen (2009) concerns the *stability* of this variable in terms of the labor cost to operating result ratio over the period we study. As they point out, changes in labor costs and/or return on

[45] None of the 92 DiD-paper reviewed by Marianne Bertrand and her co-authors was based upon matching procedures. However, their criticism may still apply.

capital may affect the dependent variable and cause a spurious effect to be credited the governmental initiatives[46].

The central question to be answered in this paper concerns the value creation that is caused by the initiatives we study, two financial schemes and two intervention schemes. Although the notion of *value creation* has a central role in the public debate about industrial policies, it is not a well-defined concept. We constructed a proxy for value creation, a composite of the contribution to labor and capital, labeled it *Added Value* and defined it as the sum of labor costs and net operating result. The concept is less than perfect but corresponds closely with the popular conception of value creation.

In order to avoid that changes in the dependent variable are misinterpreted as causal effects we included a variable which we labeled *technology*. Technology is defined as labor cost as a percentage of total income. We called this variable semi-continuous since we collapse it into 11 categories where 0 means that labor costs is between 0 and 10% and 11 means that labor costs is between 90 to 100% of total income. The percentage correct match for this variable varies from 97% to 100% (Table 13 to Table 15 page 91 to page 96). Clearly, the close match on this variable reduces the problem, but does not eliminate it. If a significant change in the ratio of labor costs to net operating result should occur over the period we study, it would most likely show up the further we move away from the initial match i.e. the more we

[46] Since nothing is said about the direction of this spurious effect, we believe they assumed this effect to be positive and thus being a contribution to overestimation.

move towards 2006. The direction of a presumed impact is not, however, very clear[47]. It is however; straightforward to check if such a change has occurred.

As pointed out by R&H, it is very clear that a close match in the pre-treatment period cannot be expected to be a good match over the entire observation period. That is one of the main reasons for demonstrating both the two-period estimates and the panel estimates. Note, however, that three of the matching variables, industry, region and newness should be expected to be fairly, if not completely stable, over the observation period. Thus, along these dimensions we should expect a close match.

R&H also suggest using alternative measures for value creation, e.g. a dependent variable where labor costs are dropped. In fact, the pre-processing perspective and the procedures employed facilitate the replacement of any dependent variable as long as the new variable relates to the conditioning variables in a way that satisfied the conditional independence assumptions (CIA).

[47] Moreover, a major change in the ratio of labor costs to net operating result sufficiently large to influence the difference between averages at the national level would be sensational and probably publicly known in advance.

14.2 Appropriateness of the choosen estimation strategies

The evaluation of impact of the two financial schemes and the two intervention programs are carried out by means of a matching strategy aimed at the construction of a control group that facilitated pairwise comparisons of similar firms, where one of the two firms received treatment and the other did not receive treatment. We called this first phase *a nonparametric preprocessing method*, thus indicating that we look at this first step, not as an integral part of the analyses, but as a procedure that left several options for analyses open. The two estimation strategies, two-period difference-in-differences estimation and panel data analysis are two distinct different statistical methods. The two estimation methods have two things in common; they both use the difference-in-differences estimator and they both use the same data. Furthermore, we use *the same* five covariates for matching for the two financial schemes and the two programs namely *company size*, *labor costs relative to company size*, *industry classification*, *newness* and *location*.

It is clear that the chosen estimation strategies and the preprocessing view introduced cause our analyses to deviate from the more common procedures based on propensity score matching where both matching and estimation are seen as one coherent estimation method. Traditionally, propensity score matching is predominantly a cross-sectional method. The *model* that is used for creating the propensity scores is usually central to discussions. It is known the results from various algorithms such as nearest neighbors matching, radius matching, kernel matching and local linear regression matching may yield slightly different results (Imbens, 2000); (Rosenbaum et al., 1985). Consequently, lengthy discussions concerning the selection of matching algorithms are frequent. Moreover, one-to-many matching,

i.e. using more than one control case per treatment case is known to reduce bias (Abadie et al., 2001), which is a useful property in situations with a limited pool of potential control cases. Also, with few control cases available matching *with replacements* is frequently in use so that any control case can be used for more than one treatment case. However, whereas one-to-many matching with replacement is easily implemented in a cross-sectional setting, it is for obvious reasons very awkward to use with longitudinal data where we want to be able to trace the same unit over time.

Thus, we have chosen one-to-one[48] matching without replacement for our construction of control groups. It is known that replacement of cases in general will lower the bias and increase the variance (Abadie et al., 2002). The decision to use one-to-one matching is based on the fact that we have a large pool of potential control cases to choose from.

The decisions concerning preprocessing of the data, the choice of two different estimation methods and the selection of *the same five variables* to be used as the basis for matching for both financial schemes and programs together with the use of covariate matching (instead of propensity score matching) are choices that make this paper different from many other applications of matching procedures. We believe that these choices make the results more robust. Compared to the well-structured discussions following Caliendo's (2005) proposal for propensity score matching (Figure 7, page 70), however, the discussion should be slightly different: Firstly, the choice of covariate matching (CVM) renders the discussions of models for

[48] Not to be confused with exact matching although. Because **nnmatch** produce only a single distance metric one-to-one matching appears to be exact matching even though such a thing does not exist as long as at least one continuous variable is included among the matching variables.

propensity score estimation and the choice of matching algorithm obsolete. Second, there is little need for discussing regression adjustment since such bias reducing procedures may not make much difference in one-to-one (near exact) matching. There is, however, a need for discussing the common support assumption even though this cannot be carried out in the same manner as for propensity score matching since separate casewise estimates of the probabilities of being in the treatment group (the propensity score) does nor exits. The only comparable numeral for CVM is the metric that measures the distance between two vectors of covariates, which by definition implies that a single number only exists for each matched pair. Thus, the term "outside the region of common support" has no meaning. The degree of closeness, which is also discussed within the PSM literature (Wilde & Hollister, 2007) does, however make good sense. Lastly, judgments concerning matching quality and the potential influence of unmeasured variables (potential confounders) should be equally relevant for both CVM and PSM.

15. Concuding remarks

The report states that it is fully possible to evaluate the long-term impact of governmental initiatives and that evaluations can be carried out based upon available administrative records. The lines of reasoning put forward are based upon state of the art algorithms and procedures. Provided that the administrative records used are free of errors and that estimation procedures are carried out in manners that preclude considerable mistakes, we believe that the causal effects of the financial schemes and intervention programs are substantiated and that the estimated sums in money terms are within reasonable range of their true magnitudes. The routines used are in line with the bulk of the literature produced by the most prominent scholars within this field. Thus, despite a number of necessary compromises due to the inherent limitation of available information, we believe the results stated in this report reflect the actual causal effects of the governmental initiatives analyzed. The applicability of the methods suggests that a range of programs and other initiatives can safely be evaluated by utilizing the information stored in administrative records.

References

Abadie, A. 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. ***Journal of the American Statistical Association***, 97(457): 284-292.

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. 2001. Implementing Matching Estimators for Average Treatment Effects in Stata. ***The Stata Journal***, 1(1): 1-18.

Abadie, A., & Imbens, G. 2002. Simple and Bias-Corrected Matching Estimators for Average Treatment Effects: 1-57. Cambridge, MA: National Bureau of Economic Research.

Abadie, A., & Imbens, G. 2006. On the Failure of the Bootstrap for Matching Estimators. San Francisco: Department of Economics, and Department of Agricultural and Resource Economics, University of California at Berkeley.

Arokiasamy, C. V., Robertson, J. E., & Guice, S. E. 1993. Effective strategies for directing and managing change in the rehabilitation setting. In N. D. S. L. J. F. Charles J. Durgin (Ed.), ***Staff development and clinical intervention in brain injury rehabilitation***: 335-372: Aspen Publishers, Inc, Gaithersburg, MD, US.

Ashenfelter, O. 1978. Estimating the effects of training programs on earnings. ***Review of Economics and Statistics***, 60: 47-57.

Ashenfelter, O., Ashmore, D., & Deschenes, O. 2005. Do unemployment insurance recipients actively seek work? Evidence from randomized trials in four US States. ***Journal of Econometrics***, 125(1-2): 53-75.

Ashenfelter, O., & Card, D. 1985. Using the Longitudinal Structure of Earnings to Estimate the Effect of Training-Programs. ***Review of Economics and Statistics***, 67(4): 648-660.

Baltagi, B. H. 2008. ***Econometric Analysis of Panel Data***. Hoboken, NJ: John Wiley & Sons Inc.

Becker, S. O., & Caliendo, M. 2007. Sensitivity analysis for average treatment effects. ***Stata Journal***, 7(1): 71-83.

Becker, S. O., & Ichino, A. 2002. Estimating of average treatment effects based on propensity scores. ***The Stata Journal***, 2(4): 358-377.

Bertrand, M., Duflo, E., & Mullainathan, S. 2004. How much should we trust differences-in-differences estimates? ***Quarterly Journal of Economics***, 119(1): 249-275.

Bjørklund, A., & Moffitt, R. 1987. The estimation of wage gains and welfare gains in self-selection models. ***Economics and Statistics***, 69: 42-49.

Black, D. A., & Smith, J. A. 2004. How robust is the evidence on the effects of college quality? Evidence from matching. ***Journal of Econometrics***, 121(1-2): 99-124.

Blossfeld, H. P., & Rohwer, G. 1997. Causal inference, time and observation plans in the social sciences. ***Quality & Quantity***, 31(4): 361-384.

Bollen, K. A. 1989. ***Structural equations with latent variables***. New York: Wiley.

Britton, J. N. H. 2003. Network structure of an industrial cluster: electronics in Toronto. ***Environment and Planning A***, 35(6): 983-1006.

Bryson, A., Dorsett, R., & Purdon, S. 2002. The use of propensity score matching in the evaluation of active labour market policies. London: Department for Work and Pensions.

Caliendo, M., & Kopeinig, S. 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. Bonn: Forschunginstitut zur Zukunft der Arbeit.

Chang, W., & Herrmann, P. 2007. What we imagine versus how we imagine, and a problem for explaining counterfactual thoughts with causal ones. ***Behavioral and Brain Sciences***, 30(5-6): 455-+.

Choi, S. 2007. Causation and Counterfactual Dependence. ***Erkenntnis***(67): 1-16.

Cohen, J. 1988. ***Statistical power analysis for the behavioral sciences*** (2nd ed.). Hilsdal, NJ.: Erlbaum.

Cohen, J. 1992. A Power Primer. ***Psychological Bulletin***, 112(1): 155-159.

Cook, T. D., & Campbell, D. T. 1979. ***Quasi-experimentation Design & analysis issues for field settings***. Boston: Houghton Mifflin Co.

Cox, D. R. 1958. ***Planning of Experiments***. New York: Wiley.

DiPrete, T. A., & Gangl, M. 2004. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. ***Sociological Methodology, Vol 34***, 34: 271-310.

Dorsett, R. 2005. Unemployed couples: the labour market effects of making both partners search for work. ***Journal of the Royal Statistical Society Series a-Statistics in Society***, 168: 365-385.

Driscoll, J. C., & Kraay, A. C. 1998. Consistent covariance matrix estimation with spatially dependent panel data. ***Review of Economics and Statistics***, 80(4): 549-560.

Econ. 1998. ***Evaluering av SNDs Nettverksprogram ; utarbeidet for SND***. Oslo: ECON Senter for økonomisk analyse.

Eichler, M., & Lechner, M. 2002. An evaluation of public employment programmes in the East German State of Sachsen-Anhalt. ***Labour Economics***, 9(2): 143-186.

Frölich, M. 2007a. On the inefficiency of propensity score matching. ***Advances in Statistical Analysis AStA***, 91(3): 279-290.

Frölich, M. 2007b. On the inefficiency of propensity score matching. ***AStA Advances in Statistical Analysis***, 91(3): 279-290.

Gangl, M. 2004. RBOUNDS: Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated, Social Science Centre ed. Berlin.

Gardner, M. J., & Altman, D. G. 1989. ***Statistics with confidence : confidence intervals and statistical guidelines***. London: British Medical Journal.

Glynn, A., & Quinn, K. 2008. Choosing an Identifying Set of Matching or Conditioning Variables, ***Department of Government and The Institute for Quantitative Social Sciences***: 38. Cambridge, MA: Harvard University.

Hadi, A. S. 1992. Identifying Multiple Outliers in Multivariate Data. ***Journal of the Royal Statistical Society, Series (B)***, 54(761-771).

Hadi, A. S. 1994. A Modification of a Method for the Detection of Outliers in Multivariate Samples. ***Journal of the Royal Statistical Society, Series (B)***, 56(393-396).

Hausman, J. A. 1978. Specification Tests in Econometrics. ***Econometrica***, 46(6).

Heckman, J., J. , Ichimura, H., & Todd, P. 1998a. Matching As An Econometric Evaluation Estimator. ***Review of Economic Studies***(65): 261-294.

Heckman, J., J., & Smith, F., A. 1999. The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies. ***The Economics Journal***, July(109): 313-348.

Heckman, J., Tobias, J. L., & Vytlacil, E. 2001. Four parameters of interest in the evaluation of social programs. ***Southern Economic Journal***, 68(2): 211-223.

Heckman, J. J. 1976. Common Structure of Statistical-Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. ***Annals of Economic and Social Measurement***, 5(4): 475-492.

Heckman, J. J. 1977. ***Sample selection bias as a specification error: with an application to the estimation of labor supply functions*** (Rev. ed.). Stanford, CA: Center for Economic Analysis of Human Behavior and Social Institutions National Bureau of Economic Research.

Heckman, J. J. 1979. Sample Selection Bias as a Specification Error. ***Econometrica***, 47(1): 153-161.

Heckman, J. J. 1992. Randomization and Social Program Evaluation. In C. Manski, & I. Garfinkel (Eds.), ***Evaluating Welfare and Training Programs***: 201-230. Boston: Harvard University Press.

Heckman, J. J. 1999. ***Casual parameters and policy analysis in economics : a twentieth century retrospective***. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., Ichimura, H., & Todd, P. E. 1997. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. ***Review of Economic Studies***, 64: 605-654.

Heckman, J. J., & Smith, J. A. 1995. Assessing the Case for Social Experiment. ***Journal of Economic Perspectives***, 9(2): 85-110.

Heckman, J. J., & Smith, J. A. 1998b. ***Evaluating the welfare state***. Cambridge, MA.: National Bureau of Economic Research.

Heckman, J. J., & Smith, J. A. 1998c. Evaluating the Welfare State. Cambridge, MA: National Bureau of Economic Research.

Himmelberg, C. P., & Petersen, B. C. 1994. Research-and-Development and Internal Finance - a Panel Study of Small Firms in High-Tech Industries. ***Review of Economics and Statistics***, 76(1): 38-51.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. ***Political Analysis***, 15(3): 199-236.

Holland, P. W. 1986. Statistics and Causal Inference. ***Journal of the American Statistical***

Association(81): 945-970.

Holland, P. W., & Rubin, D. B. 1988. Causal Inference in Retrospective Studies. ***Evaluation Review***, 12(3): 203-231.

Hubbard, R. G. 1998. Capital-market imperfections and investment. ***Journal of Economic Literature***, 36(1): 193-225.

Imbens, G. 2000. The Role of Propensity scores in Estimating Dose-Response Functions. ***Biometrika***, 87(3): 706-710.

Imbens, G. W., & Angrist, J. D. 1994. Identification and Estimation of Local Average Treatment Effects. ***Econometrica***, 62(2): 467-475.

Imbens, G. W., & Wooldridge, J. M. 2007. What's New in Econometrics? Difference-in-difference estimation, ***NBER Lecture Notes 10, Summer '07***. Cambridge, MA.

Johansson, B., Stough, R. R., & Karlsson, C. 2005. ***Industrial clusters and inter-firm networks***. Cheltenham, UK: Edward Elgar.

Jöreskog, K. G., & Sörbom, D. 1989. ***LISREL 7 : a guide to the program and applications*** (2nd ed.). Chicago, Ill.: SPSS Inc.

Kacirkova, M. 2009. Industrial Cluster and Inter-Firm Networks. ***Ekonomicky Casopis***, 57(2): 199-202.

Kay, J. 1993. ***Foundations of corporate success : how business strategies add value***. Oxford: Oxford University Press.

Kay, J. A. 1995. ***Why firms succeed***. New York: Oxford University Press.

King, G., & Zeng, L. C. 2006. The dangers of extreme counterfactuals. ***Political Analysis***, 14(2): 131-159.

King, G., & Zeng, L. C. 2007. When can history be our guide? The pitfalls of counterfactual inference. ***International Studies Quarterly***, 51(1): 183-210.

Klette, T. J., Moen, J., & Griliches, Z. 2000. Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies. ***Research Policy***, 29(4-5): 471-495.

Krugman, P. R. 1991. ***Geography and trade***. Leuven: Leuven University Press.

Kvart, I. 1994. Counterfactuals - Ambiguities, True Premises, and Knowledge. ***Synthese***, 100(1): 133-164.

Kvitastein, O. A. 1995. Do Industrial Clusters Matter? A Simple Empirical Model of the Impact of Industrial Clusters Upon Firms' Survival Expectancy. In T. B. Holmesland, Ims, Knut J., Pedersen, Ansgar. (Ed.), ***Essays in Marketing and Management A Festschrift in Honor of Kjell Grønhaug***. Bergen: Fagbokforlaget.

LaLonde, R. J. 1986. Evaluating the econometric evaluation of training programs with experimental data. ***American Economic Review***, 76: 604-620.

Lechner, M. 2001. Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption. In M. Lechner, & F. Pfeiffer (Eds.), ***Econometric Evaluations of Active Labor Market Policies in Europe***. Heidelberg: Physica/Springer.

Lee, J. C. 1986. Causal Condition, Causal Asymmetry, and the Counterfactual Analysis of Causation. ***Synthese***, 67(2): 213-223.

Lerner, J. 2002. When bureaucrats meet entrepreneurs: The design of effective 'public venture capital' programmes. ***Economic Journal***, 112(477): F73-F84.

Leuven, E., & Sianesi, B. 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score

matching, common support graphing, and covariate imbalance testing: Boston College Department of Economics, Statistical Software Components.

Lewis, D. 1973. ***Counterfactuals***. Cambridge, MA: Harvard University Press.

Lewis, D. 1986. Postscript to 'Causation'. In D. Lewis (Ed.), ***Philosophical Papers: Volume II***. Oxford: Oxford University Press.

Little, R. J. A., & Rubin, D. B. 1987. ***Statistical analysis with missing data***. New York: Wiley.

Little, R. J. A., & Rubin, D. B. 1989. The Analysis of Social-Science Data with Missing Values. ***Sociological Methods & Research***, 18(2-3): 292-326.

Long, J. S., & Bollen, K. A. 1993. ***Testing structural equation models***. Newbury Park, Calif.: Sage.

Mandel, D. R. 2007. Differential focus in causal and counterfactual thinking: Different possibilities or different functions?' ***Behavioral and Brain Sciences***, 30(5-6): 460-+.

Mander, A. 2006. Sampsi - Software for Power Analysis, ***Stata Archieves***. UK.

March, J. G. 2005. Mundane Organizations and Heroic Leadership. In J. G. March, & T. Weil (Eds.), ***On Leadership***: Appendix A. Malden, Mass.: Blackwell.

March, J. G., & Sutton, R. I. 1997. Organizational performance as a dependent variable. ***Organization Science***, 8(6): 698-706.

Nesheim, T. Kvitastein, Olav A.;Lines, Rune,;Grønhaug, Kjell;Espedal, Bjarne. 1997. Evaluering av FRAM-programmet i SND. Bergen: Foundation for Research in Economics and Business Administration (SNF).

Neyman, J. S. 1923 [1990]. On the Application of Probability to Agriculture Experiments. Essays on Principles, Section 9. D.M. Sabrowska and T.P. Speed (trans. and eds.) from the Polish original, which appeared in Roczniki Nauk Rolnicych Tom X (1923), 1-51 (Annals of Agriculture). ***Statistical Science***, 5.

Orvedal, L. 2005. Prinsipper for næringspolitikken: Næringsnøytralitet versus konkurransenøytralitet. ***Økonomisk Forum***(9).

Pack, H. 2000. Industrial policy: Growth elixir or poison? ***World Bank Research Observer***, 15(1): 47-67.

Pack, H., & Saggi, K. 2006. Is there a case for industrial policy? A critical survey. ***World Bank Research Observer***, 21(2): 267-297.

Pearl, J. 2000a. ***Causality***. Cambridge: Cambridge University Press.

Pearl, J. 2000b. ***Causality : models, reasoning, and inference***. Cambridge: Cambridge University Press.

Pearl, J. 2000c. The logic of counterfactuals in causal inference. ***Journal of American Statistical Association***, 95(450): 428--435.

Pollard, P. 1983. Confirming Confirmation Bias. ***Behavioral and Brain Sciences***, 6(2): 258-259.

Quandt, R. 1972. A New Approach to Estimating Switching Regression. ***Journal of the American Statistical Association***, 67: 306-310.

Raknerud, A., & Henningsen, M. 2009. Kommentarer til Kvitastein, ***Statistics Norway***: 1-3. Oslo.

Reve, T. 1994. ***Toward an integrative model of strategy development : from dynamic clusters to core competencies***. Bergen: Center for Research in Economics and Business Administration Norwegian School of Economics and Business Administration.

Roed, K., & Raaum, O. 2003. Administrative registers - Unexplored reservoirs of scientific knowledge? ***Economic Journal***, 113(488): F258-F281.

Romanelli, E., & Khessina, O. M. 2005. Regional industrial identity: Cluster configurations and economic development. ***Organization Science***, 16(4): 344-358.

Romer, P. M. 1986. Increasing Returns and Long Run Growth. ***Journal of Political Economy***(94): 1002-1036.

Romer, P. M. 1991. A New View of Economic-Growth - Scott,Mf. ***Journal of Economic Literature***, 29(1): 127-129.

Romer, P. M. 1994. The Origins of Endogenous Growth. ***Journal of Economic Perspectives***, 8(1): 3-22.

Rosenbaum, P., & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. ***American Statistician***, 39(33-38).

Rosenbaum, P. R. 1995. ***Observational studies***. New York: Springer-Verlag.

Rosenbaum, P. R. 2006. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. In D. B. Rubin (Ed.), ***Matched Sampling for Causal Effects***: 193-206. New York: Cambridge University Press.

Rosenbaum, P. R., & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. ***Biometrika***, 70: 41-55.

Roterud, O. 2005. ***Industrial clusters & new venture performance***. [Oslo]: [Forfatterne].

Roy, A. D. 1951. Some Thoughts on the Distribution of Earnings. ***Oxford Economic Paper***(3): 135-146.

Rubin, D. B. 1973a. Matching to Remove Bias in Observational Studies. ***Biometrics***, 29(1): 159-183.

Rubin, D. B. 1973b. Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. ***Biometrics***, 29(1): 185-203.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. ***Journal of Educational Psychology***, 66: 688-701.

Rubin, D. B. 1991. Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism. ***Biometrics***, 47(4): 1213-1234.

Rubin, D. B. 1997. Estimating Causal Effects from Large Data Sets Using Propensity Scores. ***Annals of Internal Medicine***, 127(8): 757-763.

Schaffer, J. 2007. Cause and chance: Causation in an indeterministic world. ***British Journal for the Philosophy of Science***, 58(4): 869-874.

Schweder, R. 1999. Causal explanation and explanatory selection. ***Synthese***, 120(1): 115-124.

Scriven, M. 1993. ***Hard-won lessons in program evaluation***. San Francisco: Jossey-Bass.

Sekhon, J. S., & Diamond, A. 2005. ***Genetic Matching for Estimating Causal Effects***. Paper presented at the Annual Meeting of the Political Methodology, Talahassee, FL.

Sen, A. 1990. Justice - Means Versus Freedoms. ***Philosophy & Public Affairs***, 19(2): 111-121.

Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. ***Experimental and quasi-experimental designs for generalized causal inference***. Boston: Houghton Mifflin.

Shalkowski, S. A. 1992. Supervenience and Causal Necessity. ***Synthese***, 90(1): 55-87.

Stefano M. Iacus, Gary King, & Porro, G. 2008. Matching for Causal Inference Without Balance Checking. Cambridge MA: 3Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University,.

Stinchcombe, A. L. 1965. Social Structure and Organizations. In J. G. March (Ed.), 1 ed.: 142-193. Chicago: Rand McNally & Company.

Takeda, Y., Kajlkawa, Y., Sakata, I., & Matsushima, K. 2008. An analysis of geographical agglomeration and modularized industrial networks in a regional cluster: A case study at Yamagata prefecture in Japan. ***Technovation***, 28(8): 531-539.

Tinbergen, J. 1975. ***On the Theory of Economic Policy***. New York: North-Holland Publishing Company.

Weiss, J. A. 1982. Coping with Complexity - an Experimental-Study of Public-Policy Decision-Making. ***Journal of Policy Analysis and Management***, 2(1): 66-87.

Wilde, E. T., & Hollister, R. 2007. How close is close enough? Evaluating propensity score matching using data from a class size reduction experiment. ***Journal of Policy Analysis and Management***, 26(3): 455-477.

Williamson, O. E. 1985. ***Economic Institutions of Capitalism***. New York: Free Press.

Winship, C., & Morgan, S. L. 1999. The Estimation of Causal Effects from Observational Data. ***Annual Review of Sociology***, 25: 659-706.

Wold, H., & Jöreskog, K. G. 1982. ***Systems under indirect observation : causality, structure, prediction***. Amsterdam: North-Holland.

Wooldridge, J. M. 2002. ***Econometric analysis of cross section and panel data***. Cambridge, Mass.: MIT Press.

Yablo, S. 1992. Cause and Essence. ***Synthese***, 93(3): 403-449.

Aakvik, A. 2001. Bounding a matching estimator: the case of a Norwegian training program. ***Oxford Bulletin of Economics and Statistics***, 63(1): 115-+.

Long-term Effects Evaluations of Governmental Industrial Policies: the Survival value of a Small Business Enterprise (SME) Training Program

Olav A. Kvitastein[1]

Norwegian School of Economics and Business Administration

January 2010

ABSTRACT

A leadership training program is studied by means of accounting records from participating firms. Over a period of eight to fourteen years after key managers completed the program we have traced the effects of the program on the performance of participating firms. Observational study methods are applied for the construction of the causal counterfactual which serves as the baseline for the estimates of impact. Transition rate methods are applied in order to assess the survival of participating firms, i.e., that bankruptcies are avoided. The key findings are that the overall impact of the training program with respect to survival is negligible and that this overall effect can be decomposed into a positive impact upon the mature participating firms and a catastrophic negative effect on the newcomers; the newly established firms that participated.

Keywords

Program evaluation, transition rate models, observational studies, matching models

JEL Classification Codes

H43, C31, C41, C81

[1] Norwegian School of Economics and Business Administration, Department of Strategy and Management, Breiviksveien 40, NO-5045 Bergen, Norway, olav.kvitastein@nhh.no

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Leadership training is big business. Considerable amounts of money are spent worldwide by organizations in both public and private sectors for leadership development. Thousands of journal and magazine articles are written and numerous books on leadership theory can be found on the bookshelves at airports. Some of these books appear regularly on the best-seller lists and are acceptable items for the business executive's attaché case. Others are self-help books that are better left unexposed to customers and business associates.

Leadership theory is a central part of the core curriculum at most business school where the ambition is to cultivate the future leader. Leadership theory is however, not an integrated body of knowledge. It is more a set of separate theories that have evolved into a number of sub-disciplines aimed at various aspects of the leadership challenges such as how to relate to subordinates, e.g. human relations management theories, or how to act in a competitive world, e.g. strategic management theories. Strict definitions of the exact scope of leadership are futile. Most leadership training programs generally reflect that the consultants are free to preach a variety of gospels.

The money and time invested in leadership training programs, the knowledge amassing in written texts, and the efforts by scholars reflect the perception that leaders play an essential role in the operations of organizations and that leadership skills are more abstract and difficult to learn than the skills required for the rank and file members of the organization. Moreover, in order to justify the time and money spent there has to be an underlying premise or belief that the improved leadership skills enhance organizational performance. The prime

goal of this paper is to examine whether these premises and beliefs can be substantiated for one particular category of leadership programs; those that explicitly target the small and medium-sized enterprises.

1.1 SME - targeting leadership programs

The European Commission defines medium-sized enterprises as firms which employ fewer than 250 people and whose annual turnover does not exceed EUR 50 million or whose annual balance-sheet total does not exceed EUR 43 million. Small enterprises are companies below these boundaries. A second part of the definition concerns the ownership structure. An autonomous SME is not owned by more than 25% by another enterprise or public body and does not itself hold more the 25% in another enterprise. The Norwegian enterprise population is different from the corresponding population of firms in central Europe in the sense that Norway has fewer large firms and a relatively higher proportion of small firms. Thus, leaving the ownership definition equal, Norwegian SMEs are generally smaller than what is implied by the European Commission's definition.

Nearly all industrialized countries utilize taxpayers' money to offer consultancy support for small and medium-sized enterprises. SMEs are targeted by several EU programs, by public agencies, such as ALMI in Sweden, Innovation Norway in Norway, and indirectly, through subsidized private-sector consultancies, such as Law 44 in Italy and the regional Funds in France. In Sweden, advisory services to small firms was estimated to absorb seven to eight percent of net industrial costs more than ten years ago (Lundstrøm, Boter, Kjellberg, & Ohman, 1998) . In the UK, the corresponding costs are estimated to be around £650 million per year (Gavron, Cowling, & Westall, 1998). Although we do not have exact figures for

Norway, we believe that Norwegian public capital input to advisory services is at least of a comparable magnitude.

The basic rationale for the creation of SME programs and interventions aimed at stimulating the performance of SMEs in general, is, for most of the literature, the market failure paradigm (Bator, 1958), i.e. the notion that there are some "production factors" that the market "fails" to offer the SMEs. That is, it is commonly assumed that SMEs have problems with both formulating and acquiring new knowledge and skills necessary for efficient operation. It is also assumed that SMEs are of vital importance to a country's economic growth and employment (Commission, 1995). Taken together, the importance of SMEs in the national economy and their presumed problems with attaining the competence necessary for growth and performance constitutes the reasons for programs like the FRAM program.

Theories that provide guidelines for the substantial content of programs like FRAM are usually theories concerning issues of strategic management and the firm's potential benefits from improved skills and competence in the workforce (Grønhaug & Nordhaug, 1992). We can label this kind of theories "internal" because they concern specific suggestions about how performance can be enhanced by means of the implementation of goal-setting processes and procedures. Theories that seek to explain *why* such "internal" measures may work, seem, however, to be of an "external" kind. Most prominent among them are agglomeration theories or the industrial cluster hypotheses (Porter, 1990) derived from newer theories of endogenous economic growth (Romer, 1986). The "internal" theories prescribe what should be done in the firm, while the "external" theories explain why the prescribed actions would produce desirable outcomes, provided favorable external circumstances.

Provided that some loss with respect to detail is acceptable, it is our ambition to demonstrate that program impact can be demonstrated. The loss we have to accept concerns the processes that mediate the effects. That is, we have to accept conclusions at the level decided by our units of analysis. In our case the lowest level is the individual firm. We cannot draw conclusions about the *micromediation* of the putative causes and effects. E.g. outcomes may depend upon specific person-to-person interactions within the firm that "caused" the implementation of say, an assumed successful strategy. Hence, when we attribute impact to programs, we establish a *molar* causal relation that is fallible and hence, probabilistic, (Cook & Campbell, 1979) because it is contingent on many other conditions and causal relations at lower levels where we have no empirical information. We can e.g. only speculate about to what extent specific theories of leadership and strategic management that might have been applied in the programs receive any support by our findings.

1.2 The structure of the paper

Section two introduces the problem that is studied. Section three presents the case that is examined and gives a brief description of the data. Section four outlines the preprocessing methods that lead to the construction of the dataset of matched pairs of firms which are as equal as feasible along a set of chosen dimensions such as size, technology, financial standing, industry and localization. Two different matching strategies are applied and compared. Section five introduces the hazard rate framework and decides whether we can defend to assume time as continuously measured or whether we should apply discrete time modeling methods. Section six presents the analyses and sums up the findings. Section seven discusses the implications of the findings.

2. The problem adressed

Clearly, knowledge that has the power to release or instigate enhanced performance (I) in the organization must exist and be in the possession of those in charge of the leadership training program. The individual consultant(s) (II) who run the program should have the teaching skills necessary for getting the message across (III) to the leaders. The leaders should have the managerial discretion essential for generating actions based on their newly acquired wisdom (IV) and the improved organizational performance (V) should be measurable in order to demonstrate that participation in the program was useful.

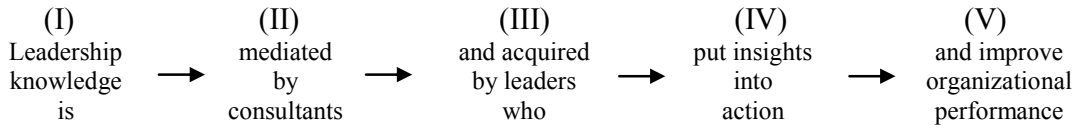


Figure 1 The presumed leadership program knowledge transmittance chain

The stylized chain of presumed knowledge transmittance reveals a number of problems. If we take no notice of the predicament that the knowledge base may be unsatisfactory and assume it so powerful that, provided seamless dissemination and implementation of the message, improved performance would of necessity occur, some problems may still remain. If everyone involved in the in the knowledge transmission chain were clever enough to send on eighty percent of the original wisdom, only about forty percent of the message would remain intact for the implementation phase. Moreover, managerial discretion rarely amounts to dictatorship; unruly markets and other external factors may dominate outcomes and thus making it hard to decide whether changes in organizational performance are due to good

management or simply a matter of luck or coincidence. Thus, even though the two basic premises that justify leadership training are valid, namely that; (i) the participants in the program are leaders who play an essential role in the operations of their organizations and, (ii) improved leadership skills enhance organizational performance, the potential impact of any leadership program can be evaluated only in retrospect. Exactly how much time that should elapse before results can be expected to show is not easy to decide on in advance. In this particular study we have an observation window of 8 to 14 years, that is, we have observations from a leadership program from 1992 to 1997 and the study ends in 2005.

Establishing the existence of a link between leadership training programs and organizational performance empirically is still not straightforward: Leadership skills are abilities at the individual level while organizational performance has to be measured at one or other aggregate level i.e. at division or subdivision level or for the smaller firms, for the entire company. Thus, this study concerns only one aspect of the results of the leadership training program; the firm's ability to stay in business. This perspective assumes that bankruptcy is an undesirable outcome; an outcome to be avoided. The declared goal of the FRAM program is to stimulate growth and new hiring of employees. Hence, it might be just to declare bankruptcy as a failure. We expect the FRAM program to inspire enthusiasm and disseminate insights that translate into action assumed by the knowledge transmittance chain in Figure 1. Thus, our working hypothesis is that *the probability of bankruptcy is decreased among firms that have participated in the FRAM program*, compared to other firms.

Patterns of the bankruptcies may shed considerable light on the functioning of the FRAM program and may be useful for making adjustments to the program.

3. The Case and the Data

3.1 The FRAM program

The FRAM program aims at developing leadership skills and strategic competence for SMEs. The program's ambition is to contribute to lasting transfer of competence, improved competitive power and profitability. FRAM was developed in 1992-1993 by The Norwegian Industrial and Regional Development Fund (SND) as a follow-up program of a technology transfer program (BUNT). The program was offered to the SMEs as an individually tailored development program, and companies are invited to participate, provided that thy meet the following criteria:

- they should not be engaged in competition with other participants
- they should have supplier or customer relationships to other participants

Found eligible for participation, companies are "screened" by undisclosed criteria that have to with judgments concerning expectations about the company's ability to benefit from the program. The program is organized as separate projects for groups of similar firms, e.g. firms within the tourism industry, construction, and so on, usually groups of 8 to 12 persons, who in most cases are the CEO of their respective companies. The program is run by experienced process consultants, and concentrates on strategic development processes, very much in line with the standard textbook theory of normative strategic management.

3.2 The FRAM data sets

The FRAM program data set consists of 469 firms that have participated in the program over the period 1992 to 1997. For the 469 firms, survey data was collected as part of an evaluation study carried out by the *Foundation for Research in Economics and Business Admini-*

stration in 1997 (Nesheim, Kvitastein, Lines, Grønhaug, & Espedal, 1997). The survey data was supplied with information from administrative records from Dun & Bradstreet and data on bankruptcies from the Brønnøysund Register Centre, the official registry for company information in Norway. Since our primary concern in this paper is the hypothesized differences in bankruptcy rates between FRAM program participants and other, comparable firms, the public registries should be considered to be our prime source of information. To be able to carry out the matching procedure we need information about the companies in question for the year *prior* to the first time it entered the program. As we can see from Table 1 there is a slight difference between the column for the original FRAM participants and the column showing the number of FRAM participants actually used as treatment cases indicating that for a few firm data could not be identified in the registry of administrative records. The loss of 59 cases is not substantial. Noteworthy, the dataset we have, includes only a minor portion of all companies that have participated in the many variants of the FRAM program. Thus, the conclusion from this study concerns outcomes from the group of included companies only.

Table 1 FRAM-participants, treatment cases and potential controls

Year of inter-vention	Origninal FRAM participants	Treatment cases used	Cases lost	Potential control cases	Total number of cases
1992	19	18	1	25622	25640
1993	21	14	7	5699	5713
1994	110	104	6	31631	31735
1995	131	112	19	32237	32349
1996	142	124	18	38154	38278
1997	46	38	8	41439	41477
N =	469	410	59	174782	175192

Table 1 shows that we have a considerable pool of potential control companies for the six years treatment period we look at. For the years 1992, 1993 and 1997, however, the treatment groups are small and may not be well suited for stand-alone analyses.

Looking back, a decade later, in 2005, Table 2 shows that 304 of the 410 firms we were able to identify are still active. 54 firms are registered as bankrupt and 52 firms have left the market for other reasons. We have information that shows that 24 of these latter 52 firms have either merged with other firms or are in the process of negotiating acquisitions. In total, about 13% of the firms involved in the FRAM program have as far as 2005 experienced bankruptcy and an approximately equal percentage are no longer registered as active for other reasons, i.e. close to a third of the participating firms have disappeared after 2005.

Table 2 The FRAM - participants in 2005

Year of Intervention	*Firms in the FRAM-survey*	*Firms Identified*	*Firms leaving until 2005*				*Percent Total*
			Bankruptcies		*Other reasons*		
1992	19	18	2	11.1 %	2	11.1 %	22.2 %
1993	21	14	0	0.0 %	4	28.6 %	28.6 %
1994	110	104	15	14.4 %	10	9.6 %	24.0 %
1995	131	112	10	8.9 %	23	20.5 %	29.5 %
1996	142	124	20	16.1 %	9	7.3 %	23.4 %
1997	46	38	7	18.4 %	4	10.5 %	28.9 %
N =	469	410	54		52		

3.3 Measures

Apart from the hazard rate, our dependent variable that will be explained later, we measure an independent variable *concern ownership* as a dummy variable code 1 if a company is owned by more than 25% by another company and zero otherwise. We measure *short-time debt* as the by means of short-time liabilities as a percentage of total assets based upon accounting records, and, due to the liability of newness hypothesis, (Stinchcombe, 1965) we define a variable *frail* coded 1 if a company is established after 1991 and coded zero otherwise. Of these three variables, only short-time debt has to be used as a time-dependent variable, i.e., a variable that may take a different value every time it is observed. Variables used for preprocessing purposes will be reported in the sections on preprocessing methods.

4. Preprocessing methods

4.1 Selection bias, non-randomness and the causal counterfactual

Participants for the FRAM program are not picked at random. Applicants have to qualify for the programs and from the firm's side the decision to hand in the application depends heavily upon the extent to which the busy SME leaders believe the program is of any worth in their daily work. The obstacles, motivations and other more or less rational consideration that led to the decision to participate or not to participate are all part of a process that constitutes a process of non-random selection into the program. Hence, we are left with a particular problem that may be the source of considerable bias, the so-called *selection problem*. If, say, only SMEs with a solid profit margin and sufficient supply of stable long-term contracts ready at hand found participation justifiable, life expectancy would, from the outset on be higher among participating firms than among any randomly picked sample of comparable SMEs. Or, vice versa, if the FRAM program was designed to target firms on the edge of bankruptcy, we should not be surprised to find that participating firms had a lower life expectancy than comparable non-participating firms. When trying to estimate the impact of the FRAM program upon survival expectancy, the selection mechanism into to the program clearly is of importance. Hence, we have to find means of limiting selection effects as a potential threat to the validity of conclusions.

The key to estimating the impact of a program is to consider what would have happened in the absence of the program and compare this counterfactual situation to the factual situation. The problem is that the counterfactual, by definition, is something that does not exist. Thus, it has to be constructed in a way that most closely represents what alternatively would have happened. The impact caused by the program is then represented as one or other kind of distance between the factual and the counterfactual situation. This way of thinking about

causality and impact of intervention has become commonplace in the econometric evaluation literature and is looked upon as development that has evolved over time as an amalgamation between the logical foundation outlined in David Lewis's (1973) work from the early seventies, the seminal work in statistics by Donald Rubin and Paul Rosenbaum (1985a; 1983; 1984; 1985b) in the eighties and the contribution by econometricians (1989; 1998; 1992; 1989; 1997; 1991; 1998) and many other scholars throughout the eighties and the nineties[2]. A comprehensive summing up of the causal counterfactual way of reasoning is summed up in Judea Pearl's Lakatos prize winning book "Causality" in 2000 (Pearl, 2000).

A crucial feature necessary for feasible detection of the effects of a program is that we are able to distinguish between the group of firms that are *exposed to treatment* (to use the experiment jargon) and those which are *not exposed to treatment*, that is, between participants and non-participants. This is not a problem in the FRAM case. It is however important that we are able to check whether participating firms are engaged in *other* programs. To be able attribute effects as caused by the program in question, involvements in other programs could contaminate findings.

Note also that no unit (firm) can be observed in the two states we intend to use for our comparisons, as both receiving the treatment and not. If this was possible, the effect could be calculated by comparing the two states for the same unit. A major problem is that the effect of treatment has to be calculated by comparing a unit that received the treatment with *another* unit that did not receive the treatment.

[2] The list of authors could easily be extended. The burgeoning literature in the field that we loosely can label "econometric evaluation methods" makes it difficult to pay justice to all significant contributions.

Formally, say a unit can be in either a treated state, denoted state "1" or an untreated state, denoted state "0" and say outcomes Y_1 and Y_0 are associated with each state. The gain from treatment could then be calculated as the difference $\Delta = Y_1 - Y_0$. Because we cannot determine impact of treatment for the individual, we have to rely on the distribution of impact across units, call it $F(\Delta)$ or on certain features of this distribution. The expected gain to a randomly selected unit in the population, denoted $E(\Delta)=E(Y_1 - Y_0)$ refers to the expected value or population average. Given that the entire population was included, this parameter provides information necessary to carry out benefit-cost analysis when combined with information about average cost. This situation is, however, rare in program evaluations. The FRAM program targets certain firms, and we have to focus on what happens to those who actually did participate. Denoting participation $d=1$ and non-participation $d=0$ we can write the distribution of gains for participants as $F(\Delta|d=1)$ and the impact for participants as $E(\Delta|d=1)=E(Y_1 - Y_0|d=1)$. The problem is that we do not know $E(Y_0|d=1)$, it has to be estimated, and this is not straightforward. We cannot directly use the mean outcome among non-participants as a proxy for what would have happened to participants, given that they had not participated. This is easily seen by subtracting the mean outcome among non-participants from the mean outcome of participants, $E(Y_1|d=1) - E(Y_0|d=0)$, *yields*

(1) $\{E(Y_1|d=1)-E(Y_0|d=1)\}+\{E(Y_0|d=1)-E(Y_0|d=0)\}$

The expression for the counterfactual, $E(Y_0|d=1)$, that is, the outcome for the participants, provided that they had *not* participated, appears twice in the equation. The first term in the curly brackets gives the mean impact of participation, and the second term represents the *selection bias* caused by the fact that non-participants differ from participants in the non-participating state. This selection bias may be different from zero if, as mentioned above, the

SMEs' selection into the program is regulated by, say, the perceived good economic prospect of the firms. In such a case non-participants could be expected to have outcomes higher than participants, and hence, a negative selection bias could yield incorrect estimates.

Randomization solves this problem, provided that randomization does not alter the pool of participants or their behavior, and that close substitutes for the treatment are not available. Given that randomization is used both for the treatment (participant) group and the control group (non-participants) and that the control group is denied access to the treatment group, the outcomes of both groups in the zero state would be equal. That is, $E(Y_0|d=1) = E(Y_0|d=0)$ and the right hand side term in the curly brackets in the equation above would cancel out, implying no selection bias. Clearly, as is usually the case in retrospective evaluations of public intervention programs, randomization is not an option in this case of the FRAM program.

Following the recommendations of Rosenbaum and Rubin (Rosenbaum et al., 1983) we suggest *matching procedures based on propensity scores* as a way to emulate en experimental setting. *Matching*, in this context, means that for each individual firm participating in the FRAM program we seek for a "comparable" non-participating firm that "matches" the participating firm. *Propensity scores* are probabilities, i.e. scores that tell us how likely it is that a particular firm could be found in the participant group. Denoting the participant group d=1 and the non-participating group d=0, the probability of membership in the participating group can be expressed as $e(\mathbf{X})=prob(d=1|\mathbf{X})$, where **X** for each individual SME represents a vector of variables, i.e. a range of attributes of the firm that may discriminate between participating and non-participating firms. Hence, if $e(\mathbf{X}) = prob(d=1|\mathbf{X})$, i.e. if the estimated, predicted probability of membership in the participant group, conditional upon a number of

firm characteristics, **X**, is approximately equal for two different firms, and one of them is known to be *in* the participating group and the other is known *not* to be in the participant group, they are considered as matches.

This procedure has remarkable advantages compared to former matching techniques. The traditional matching techniques assume matching on each individual variable. Thus, even with a few, discrete variables, the number of possible combinations quickly becomes large and requires a very large pool of potential candidates for match. By means of propensity scores, that can easily be estimated using an ordinary logistic regression model, it is possible to match on one single variable only, namely the propensity score. Simulation studies also show that propensity scores, when compared to older matching techniques like the Mahalanobis metric matching, often removed more than twice as much bias (Gu & Rosenbaum, 1993).

The justification for the propensity score procedure is straightforward: In the simplest randomised experiment one may just toss a fair coin and, say, assign head to the treatment group and tail to the control group and in this way ensure that $e(\mathbf{X}) = prob(d=1|\mathbf{X}) = ½$ for each unit involved. With propensity score, this would clearly not be the case. Some units may have probabilities less than ½ and other may have probabilities close to 1. The point is that as long as two units have approximately *the same probabilities of membership* in the treatment group, they can be compared.

It is clearly of importance that the outcomes for the treatment and comparison group, conditional on $\boldsymbol{X}$ are independent of the (0/1) grouping itself. This is the so-called conditional independence assumption (CIA). If we assume the propensity score procedure produces this result, that is:

(2) $(Y_1, Y_0) \perp d \mid p(X)$ and thus $(Y_1, Y_0) \perp d \mid X$ where $\perp$ means "independent of",

Then, in the terminology of Rosenbaum and Rubin (1983) treatment assignment is *strongly ignorable* given the observed covariate vector. This implies that, conditional on X, outcomes in the non-participant group (comparison group) have approximately the same distribution as outcomes in the participant group, *provided that the participant had not participated.* Returning to equation (1), this implies that the equation reduces to:

(3) $E(Y_0|\boldsymbol{X},d=1) = E(Y_0|\boldsymbol{X},d=0) = E(Y_0|\boldsymbol{X})$

That is, the missing counterfactual can be estimated from the outcome among non-participants. This is an important property of the propensity score matching procedure. It facilitates approximate emulation of an experiment because it balances observed covariates, i.e. the $\boldsymbol{X}$. Clearly, the propensity score procedure does not take into account *unobserved* relevant covariates. In this respect the propensity score procedure is inferior compared to a procedure that allowed for random assignment of treatment. This is a limitation that may impose problems. Hence, even though bias in estimates may be considerably reduced, we still have to be aware of possible sources of bias that may inflict on the validity of conclusions.

4.2 Transition data analysis and the propensity score model

Propensity score matching is predominantly a cross-sectional technique. We intend to analyze longitudinal data. Our hypothesis is that participating in the FRAM program makes a difference for the participants, i.e. that bankruptcy is an event that is less likely to occur for the firms that participated than for those comparable firms that did *not* participate. Akin to what is common in the medical literature, (e.g., (Foster, 2003; McIntosh & Rubin, 1999;

Rubin, 1993; Stone et al., 1995) we construct a control group based upon matched pairs and inspect the differences in survival rates between treated and non-treated units; the only difference is that we compare firms, not people. This difference may be considerable in the sense that the dissimilarity among firms is thought of as more prevalent than the variation that is usually present among the patients that are the units in, say, the analyses of the effects of a particular drug. Clearly, this presumption can be questioned. Most likely, both firms and people show considerable variation along arrays of relevant traits that may affect the validity of conclusions. Firms and people change over time and thus, any two chosen units that are fairly equal at the beginning of the observation period may be very different by the end of the period. This is a problem that is easily overlooked. Moreover, since survival analysis implies that the data-generating processes evolve over time, the problem of the potential influence of hidden or unmeasured variables parallels central problems in the Neyman-Rubin model (Holland, 1986). Also, other near untestable problems such as the stable unit-treatment assumption (SUTVA) which demands that the effect of treatment on unit *i* should be independent of the effects of treatments on any other unit will also apply to transition rate models.

4.3 The practical implementation of propensity score matching

According to Caliendo and Kopeinig (2005) the implementation of propensity score matching can be thought of as a five-step procedure:

(1) Propensity score estimation
(2) The choice of matching algorithm
(3) Checking of overlap/common support
(4) The evaluation of matching quality
(5) Sensitivity analysis

The first step involves the choice of conditioning variables for the matching procedure. This clearly involves decisions with wide-ranging consequences.

4.3.1 Propensity score matching and the choice of conditioning variables

Propensity score is estimated by means of a logistic regression model, which is by far the most commonly applied. The covariates used for estimating the propensity scores are shown in Table 3 below. The chosen covariates reflect that we want the treatment group and the control group to balance with respect to size, (measured as the logarithm of total turnover) the financial situation (where the equity ratio is used as a proxy for the relative wealth of the company) two measures for capital/labor intensiveness (where factor costs as a percentage of total turnover are used as proxies). In addition, we use a number of dummy variables to account for *localization* and *business sectors* broadly defined in order to increase the probability of finding matches that roughly operate within the same markets[3] within the same regions. Even though most of the firms involved are SMEs we have a dummy variable for those firms that are big enough to be characterized as medium-sized enterprises. The last variable, newness is of particular importance since new firms are more vulnerable to external forces than older firms according to the *liability of newness* hypothesis (Stinchcombe, 1965).

The decisions concerning the selection of covariates for the estimation of the propensity score is of crucial importance in the construction of the counterfactual. Since causal reasoning in observational studies usually assumes that treatment and control cases are matched at

[3] The term business sector is used simply because we deviate somewhat from the Standard Industrial Classification (SIC). We are well aware that two firms categorized within the same business sector do not necessarily operate under equal competitive conditions. Even with many more dummy variables added this has to be a coarse approximation to comparable competitive conditions.

one or other point in time *prior* to treatment the conditioning variables have to be measurable in the pretreatment period. The chosen covariates are decisive in the sense that they give the dimensions along which we can consider the treatment group and control group to be comparable. Unfortunately, the literature provides no clear advice with respect to the choice of conditioning covariates. The classical econometric advice is that when one is unsure whether a measured pre-treatment variable should be included or not it is best to err on the side of including potentially irrelevant covariates. The famous quotation often used to justify the almost unrestrained inclusion of hopefully relevant matching variables is:

"The conclusion, then, is that if the specification error consists of including some irrelevant explanatory variables in the regression equation, the least squares estimators of the regression coefficients are unbiased but not efficient. The estimators of the variances are also unbiased, so that, in the absence of other complications, the usual tests of significance and confidence intervals for the regression coefficients are valid" (Kmenta, 1986).

The question under which circumstances conditioning covariates effectively reduce selection bias in observational studies is investigated in a series of more recent articles by Thomas D. Cook, William R. Shadish et al. (2008; 2009; 2009; Diaz & Handa, 2006; Pohl, Steiner, Eisermann, Soellner, & Cook, 2009). Their findings, mainly based upon social science research in the context of the regression-discontinuity design, are mixed, but indicate that a substantial part of bias can be eliminated by means of various ways of conditioning on covariates. Review of the labor economics literature shows, however, that in practical application matching models almost always fail to remove all selection bias (Glazerman, Levy, & Myers, 2003; Michalopoulos, Bloom, & Hill, 2004). In theory, the covariates that support the strong ignorability assumption are those that are closely related to both the real selection process and study outcomes and still leaves outcomes independent of the treatment assignment.

Since our primary source of data is administrative records the choice of covariates is constrained by what is available in the records. The chosen covariates displayed in Table 3 reflect this limitation.

Table 3 Logistic regression – treatment/control with selected covariates

Covariates	**Coef.**	**Std. Err**	**z**	**P>z**	**[95% Conf. Interval]**	
Total turnover (log scale)	0.63	0.053	11.86	0.000	0.53	0.74
Equity ratio (percent)	0.00	0.002	-2.37	0.018	-0.01	0.00
Labor costs	0.03	0.004	7.53	0.000	0.02	0.04
Other factor costs	0.00	0.003	-0.31	0.753	-0.01	0.01
Localization						
Hedmark and Oppland	1.36	0.224	6.07	0.000	0.92	1.80
South Eastern Norway	1.08	0.195	5.54	0.000	0.70	1.46
Agder and Rogaland	1.11	0.209	5.31	0.000	0.70	1.52
Western Norway	0.85	0.200	4.25	0.000	0.46	1.24
Trøndelag	1.03	0.235	4.36	0.000	0.56	1.49
Northern Norway	1.35	0.215	6.27	0.000	0.93	1.77
Business sectors						
Other Industries	0.69	0.224	3.09	0.002	0.25	1.13
Construction	-0.30	0.171	-1.74	0.081	-0.63	0.04
Property Management	-0.11	0.231	-0.49	0.627	-0.57	0.34
Information Technology	0.28	0.239	1.16	0.247	-0.19	0.74
Expert-knowledge services	-1.18	0.297	-3.98	0.000	-1.76	-0.60
Furniture Industry	0.80	0.321	2.49	0.013	0.17	1.43
Food industry	-0.08	0.334	-0.23	0.817	-0.73	0.58
Plastics industry	0.98	0.297	3.30	0.001	0.40	1.56
Tourism industry	-0.18	0.224	-0.80	0.421	-0.62	0.26
Wood products	0.87	0.211	4.13	0.000	0.46	1.28
Commodity trade	-1.39	0.196	-7.08	0.000	-1.77	-1.00
Shipbuilding industry	0.36	0.371	0.96	0.338	-0.37	1.08
Size						
Medium-sized enterprises	-1.65	0.392	-4.20	0.000	-2.42	-0.88
Firm maturity						
Newness	0.01	0.004	2.55	0.011	0.00	0.02
Constant	-13.08	0.569	-22.98	0.000	-14.20	-11.97

Number of obs =	175192	*410*	*Treatment cases*
LR chi2(25) =	521.95	*174782*	*Potential control cases*
Prob > chi2 =	0		
Pseudo R2 =	0.0902		

The table shows the estimates for the chosen conditioning variables measured in the year prior to the participation in the FRAM program. This is the propensity score producing variables used for constructing datasets of matched pairs, i.e. to each and every unit we have chosen to proceed by using only one control unit for each treatment unit.

4.3.2 Two different Matching Strategies

Table 1 (page 196) shows that the year of participation (intervention time) spans from 1992 to 1997, a period that covers six years. The two first years and the last year have relatively few participants. The logistic regression shown in Table 3 above is based on all six years jointly used as the treatment category and coded 1 on the binary dependent variable, and the controls coded as zeros. All covariates are measured one year prior to intervention. Thus, the binary dependent variable for intervention in year *t* is moved to year *t-1* and the estimated predicted (the propensity score) value for intervention in year *t* is estimated in year *t-1*. This is the conventional procedure in the cross-sectional usage of propensity scores. In the longitudinal setting the logic of the procedure is that of constructing datasets of matched pairs which facilitate following the development of each comparable pair over time. The problem with this procedure is of course that pairs that are relatively equal at *t-1* measured from the time of intervention may become increasingly unequal over time. Note also that since our intention is to compare the probability of bankruptcies between firms that have participated in the FRAM program and comparable non-participants, we find it convenient to restrict the number of comparison cases to one for each firm, i.e. use matched pairs[4].

The criticism of the year-by-year deterioration of matching quality as we move further away from the matching year is relevant for longitudinal studies based upon matching methods. We therefore introduce a second matching regime where we use the best matches within each year. A brief look at Figure 1 can explain the difference.

[4] This is strictly not necessary. We could use more than one comparison case and weight each control case according its distance from the best match. Many comparison cases and weighting may, however, become very cumbersome in the survival analyses.

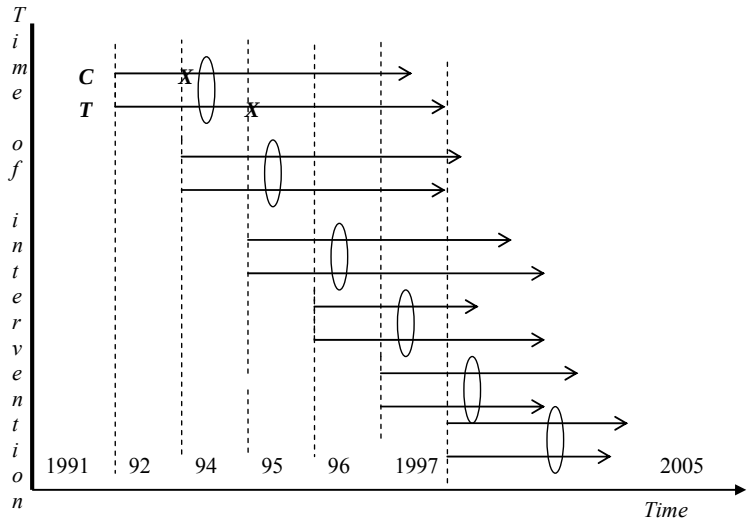


Figure 2 The conventional matching regime

The conventional matching regime for program evaluations usually employs a before-after strategy where matching is based upon information gathered prior to the intervention, and thus, the counterfactual is searched established and utilized for comparisons after the intervention event of interest. This is also what we have done here in what we can call the *first regime* which is based on the logistic regression model in Table 3. As illustrated in Figure 2 we intend to construct pairs of firms that are as equal as possible, conditional upon our chosen covariate vector at each year of intervention 1992 to 1997, as marked with ellipsoids embracing each pair. Thus we will produce a dataset with exactly twice as many units as we have treatment cases, provided that we are able to find an exact match for every treatment case. This procedure makes it easy to measure duration time i.e. time from intervention to event, either bankruptcy or further activity, which is the input for further analysis.

What we have called the *second matching regime* is radically different in the sense that we now estimate for each year i.e. in the observation period 1992 to 2005 we estimate 14 models using the same conditioning variables as in Table 3. Thus, instead of constructing

pairs that are as equal as possible in the year prior to intervention we construct pairs that are as equal as possible within each year that we observe. This implies that we use 14 different estimates of the propensity score instead of just one. Thus, we eliminate the problem of increasing differences within matched pairs over time and create another one, namely that we now do not follow the same control unit over time. The logic of this second matching regime is different from the first regime in the sense that whereas in the first regime we follow matched pairs of treatment units and control unites over time, we now ask the question; what is the duration time in the state as active (or until event) for those (control unit) firms that are maximally equal to the firms in the treatment group within each year they are observed? Clearly, this implies that while we follow all treatment cases over the entire observation period any control case may be substituted for a different one within each year. The advantage of this procedure is each matched pair is more equal in the sense that the absolute distance between the propensity score for the control cases and the treatment cases are more effectively minimized and thus we should achieve an improved bias reduction. Moreover, we also minimize the potential impact of hidden (unmeasured) variables and thus reduce the influence of unobserved heterogeneity.

4.3.3 The choice of matching algorithm

A number of user supplied matching algorithms available in Stata© and ***R***+ can be applied for finding the closest possible match between the propensity score in the treatment group and the control group. The many programs and algorithms offer an extensive collection of the many ways two numbers can be matched. Our two matching regimes do, however, make it most convenient to use algorithms that facilitate exact matching (for each year) in addition to the matching on the propensity score. We therefore use a modified version of the Stata ado-program **nnmatch** (Abadie, Drukker, Herr, & Imbens, 2001) where we use the option for exact matching for the variable years and otherwise use Mahalanobis matching on

the logit of the propensity score. The one modification we do is that we do not allow a control case that is selected as match in one year to reappear as a match in any subsequent year. This modification is done in a sequence of matches *outside* the ado-file **nnmatch**.

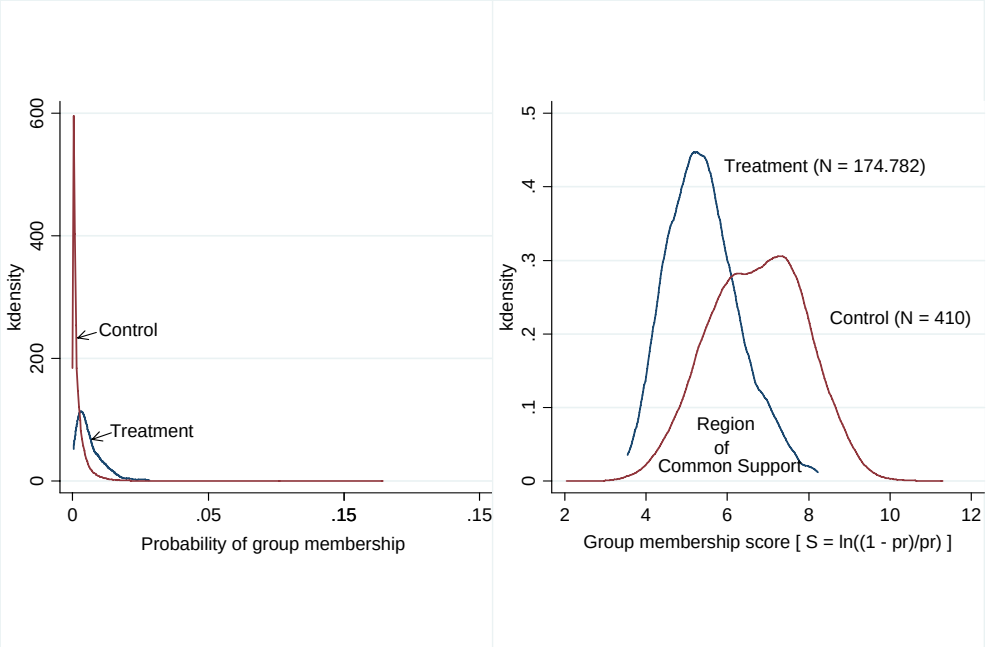


Figure 3 The effect of a logit transformation on the propensity score

Taking the logit of the propensity score makes no difference when we apply the algorithm for Mahalanobis distance[5]. Figure 3 is based on the predicted values from the logistic regression model for what we call regime 1 matching (Table 3 above). The figure reveals a very common pattern with more values near zero in the control group than in the treatment group. The right hand side of the figure demonstrates how it is easier to see the region of common support when we use that logit of the propensity score. Thus, we can decide the boundaries of the region of common support in advance and use both the value of the Mahalanobis distance metric and the requirement that we have to be inside the region of

[5] The logit transformation makes a difference for other matching algorithms such as caliper matching and other algorithms.

common support in our decisions concerning what cases to accept and what cases to disregard. In both matching regime 1 and matching regime 2, of the control cases 408 are inside the region of common support.

4.4 The balancing properties of matching regime 1

There are many ways to present the achieved balancing properties of the matched samples, from the least meaningful but most frequently applied, the *t*-tests for differences in means[6] to the quantile-quantile plot, which is excellent for this kind of comparison since it also shows the equality of the distribution of the covariate in the treatment and control groups. Because of the considerable number of comparisons due to our two matching regimes we apply the suggestion by Rosenbaum & Rubin (Haviland, Nagin, & Rosenbaum, 2007; Rosenbaum et al., 1985a) for the continuous variables since it gives a compact presentation of the balance achieved. Table 4 reports two absolute standardized measures similar to Cohen's *d* (Cohen, 1988) where, $d_x = \frac{|M_{xt} - M_{xp}|}{s_x}$, $d_{xm} = \frac{|M_{xt} - M_{xc}|}{s_x}$ and $s_x = \sqrt{(s_{xt}^2 + s_{xc}^2)/2}$ where M_{xp} is the mean of all potential control cases (N = 174782), M_{xt} is the mean of the treatment cases (N= 410) and M_{xc} is the mean of the matched control cases and s_x is the overall standard deviation in the calculations of absolute differences.

For the dummy-variables included we simply report the proportions since with exact matching the proportion of any dummy-variable should be close to 50% in both the control group and the treatment group[7]. As shown in Table 4 above, matching substantially improves the balance of the covariates. Apart from the percent bias reduction, which

[6] Simple tests for differences in means are deceptive because the *t*-statistics depends on the number of cases involved and the number of cases before match will normally outnumber the number of cases after match and cause an inevitable decrease in the *t*-values after match.

[7] Effect sizes would balance and both d_x and d_{xm} would be 0 at 50%.

depends heavily on the relative imbalance before matching, what we want to see is values near zero for the d_{xm} – column. Note that the propensity score is included in the table even though it is not a covariate. It is included to demonstrate the effect of matching in terms of d_x and d_{xm}.

Table 4 Covariate Imbalance before and after Matching - Matching Regime 1

Covariate	*Year*	d_x	d_{xm}	*Percent bias Reduction*
Estimated propensity score	1991	*1.26*	0.13	89.7
"	1992	0.98	0.01	99.0
"	1993	0.93	0.08	91.4
"	1994	0.80	0.03	96.3
"	1995	0.94	0.00	100.0
"	1996	0.86	0.11	87.2
Total turnover (log scale)	1991	0.31	0.12	61.3
"	1992	0.41	0.15	63.4
"	1993	0.46	0.19	58.7
"	1994	0.44	0.06	86.4
"	1995	0.59	0.02	96.6
"	1996	0.51	0.37	27.5
Equity ratio (percent)	1991	0.00	0.52	na
"	1992	0.36	0.35	2.8
"	1993	0.09	0.03	66.7
"	1994	0.05	0.10	-100.0
"	1995	0.00	0.01	na
"	1996	0.01	0.13	-1200.0
Labor costs	1991	0.40	*0.60*	-50.0
(as percentage of turnover)	1992	0.35	0.30	14.3
"	1993	0.52	0.04	92.3
"	1994	0.48	0.00	100.0
"	1995	0.46	0.03	93.5
"	1996	0.73	0.03	95.9
Other factor costs	1991	0.45	0.39	13.3
(as percentage of turnover)	1992	0.48	0.12	75.0
"	1993	0.39	0.03	92.3
"	1994	0.33	0.02	93.9
"	1995	0.26	0.01	96.2
"	1996	0.46	0.02	95.7
Maximum		1.26	0.60	

Table 4 also shows that the years with the lowest number of cases (1991, 1992 and 1996 with respectively 18, 14 and 38 cases) do not achieve the same level of balance as the years with a higher number of cases. The same pattern is also observable in Table 5 where most dummy-variables balance around 50% for the years 1993 to 1995.

The quantile-quantile plot in Figure 4 gives a better visualization of the distribution of the propensity scores after matching. Any two equal distributions will follow the 45% degree

line in the figures while the differences between the two distributions will appear outside the 45% degree line.

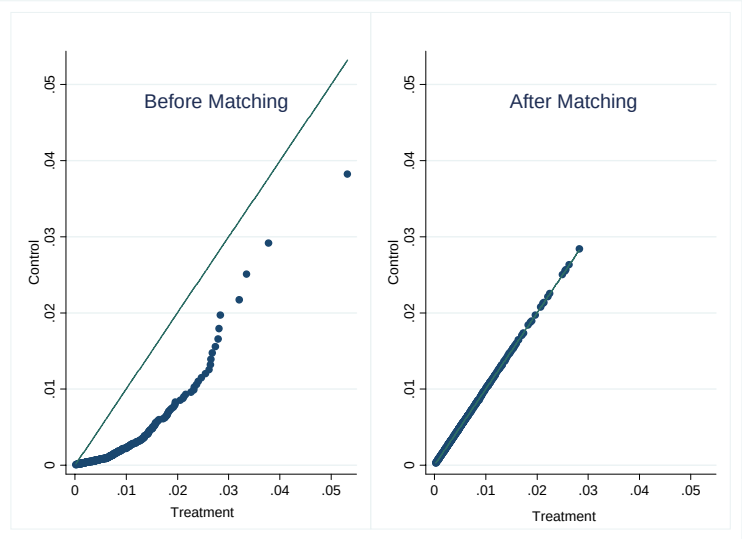


Figure 4 QQ- plot of the propensity score before and after matching – Regime 1

Table 5 Dummy covariate Imbalance after Matching - Matching Regime 1

Dummy variable	*One Year prior to Intervention*						*Average*
Covariates	1991	1992	1993	1994	1995	1996	*over*
Localization			*(percent)*				*years*
Hedmark and Oppland	0.0	0.0	57.9	57.1	46.4	65.0	*54.2*
South Eastern Norway	50.0	57.1	48.9	52.2	48.4	50.0	*50.0*
Agder and Rogaland	62.5	60.0	43.8	43.3	47.5	63.6	*48.4*
Western Norway	66.7	37.5	63.2	55.8	50.0	50.0	*55.4*
Trøndelag	50.0	0.0	45.5	50.0	53.9	25.0	*46.1*
Northern Norway	33.3	50.0	40.0	47.2	52.8	18.2	*43.9*
Business sectors							
Other Industries	66.7	50.0	53.8	58.3	28.6	100.0	*51.0*
Construction	50.0	0.0	50.0	54.8	56.0	37.5	*49.5*
Property Management	100.0	50.0	53.3	41.7	42.9	33.3	*46.8*
Information Technology	0.0	50.0	28.6	66.7	64.3	80.0	*57.9*
Expert-knowledge services	0.0	0.0	50.0	40.0	71.4	100.0	*65.0*
Furniture Industry	100.0	0.0	45.5	33.3	20.0	0.0	*38.5*
Food industry	50.0	50.0	50.0	60.0	20.0	33.3	*43.5*
Plastics industry	50.0	100.0	66.7	33.3	60.0	100.0	*65.0*
Tourism industry	0.0	0.0	25.0	50.0	45.8	0.0	*38.5*
Wood products	33.3	66.7	66.7	45.5	58.8	33.3	*56.4*
Commodity trade	0.0	0.0	57.1	46.4	52.9	50.0	*51.1*
Shipbuilding industry	0.0	0.0	0.0	55.7	100.0	0.0	*61.5*
Size							
Medium-sized enterprises	0.0	0.0	0.0	33.3	20.0	45.0	*50.0*
Average over dummies	*37.5*	*30.1*	*44.5*	*48.7*	*49.5*	*46.5*	

4.5 The balancing properties of matching regime 2

Following the same line of reasoning as for matching regime 1, Table 6 shows the covariate imbalance before and after matching when treatment and control cases are matched, not the year prior to treatment as in matching regime 1, but in the year they are observed. Clearly, with respect to matching quality, matching regime 2 shows a considerable improvement over matching regime 1.

Table 6 Covariate Imbalance before and after Matching - Matching Regime 2

	Covariates									
	Estimated propensity score		Total turnover (log scale)		Equity ratio (percent)		Labor costs (as percentage of turnover)		Other factor costs (as percentage of turnover)	
Year	d_x	d_{xm}	d_x	d_{xm}	d_x	d_{xm}	d_x	d_{xm}	d_x	d_{xm}
1992	1.21	0.02	0.50	0.07	0.04	0.08	0.39	0.00	0.24	0.01
1993	1.23	0.02	0.56	0.09	0.08	0.01	0.41	0.08	0.25	0.07
1994	1.29	0.02	0.63	0.04	0.08	0.01	0.40	0.03	0.23	0.05
1995	1.37	0.02	0.68	0.06	0.18	0.07	0.41	0.06	0.22	0.08
1996	1.42	0.01	0.72	0.06	0.22	0.02	0.42	0.10	0.28	0.07
1997	1.32	0.02	0.63	0.01	0.16	0.04	0.39	0.00	0.24	0.00
1998	1.20	0.02	0.51	0.09	0.23	0.03	0.32	0.03	0.21	0.01
1999	1.18	0.02	0.46	0.00	0.24	0.07	0.29	0.07	0.22	0.03
2000	1.18	0.02	0.44	0.00	0.23	0.03	0.26	0.06	0.19	0.06
2001	1.17	0.01	0.36	0.09	0.22	0.07	0.27	0.03	0.26	0.05
2002	1.12	0.02	0.33	0.08	0.17	0.02	0.27	0.00	0.24	0.02
2003	1.13	0.01	0.32	0.05	0.25	0.01	0.21	0.05	0.23	0.07
2004	1.12	0.01	0.33	0.00	0.21	0.01	0.16	0.05	0.23	0.00
2005	1.12	0.02	0.28	0.01	0.16	0.06	0.13	0.01	0.21	0.00
Max	*1.42*	*0.02*	*0.72*	*0.09*	*0.25*	*0.08*	*0.42*	*0.10*	*0.28*	*0.08*

Table 7 Percent bias reduction due to matching

Year	Est. Prop. Score	Total turnover	Equity ratio	Labor costs	Other costs
1992	98.3	86.0	-100.0	100.0	95.8
1993	98.4	83.9	87.5	80.5	72.0
1994	98.4	93.7	87.5	92.5	78.3
1995	98.5	91.2	61.1	85.4	63.6
1996	99.3	91.7	90.9	76.2	75.0
1997	98.5	98.4	75.0	100.0	100.0
1998	98.3	82.4	87.0	90.6	95.2
1999	98.3	100.0	70.8	75.9	86.4
2000	98.3	100.0	87.0	76.9	68.4
2001	99.1	75.0	68.2	88.9	80.8
2002	98.2	75.8	88.2	100.0	91.7
2003	99.1	84.4	96.0	76.2	69.6
2004	99.1	100.0	95.2	68.8	100.0
2005	98.2	96.4	62.5	92.3	100.0
Average	98.6	89.9	81.3	86.0	84.1

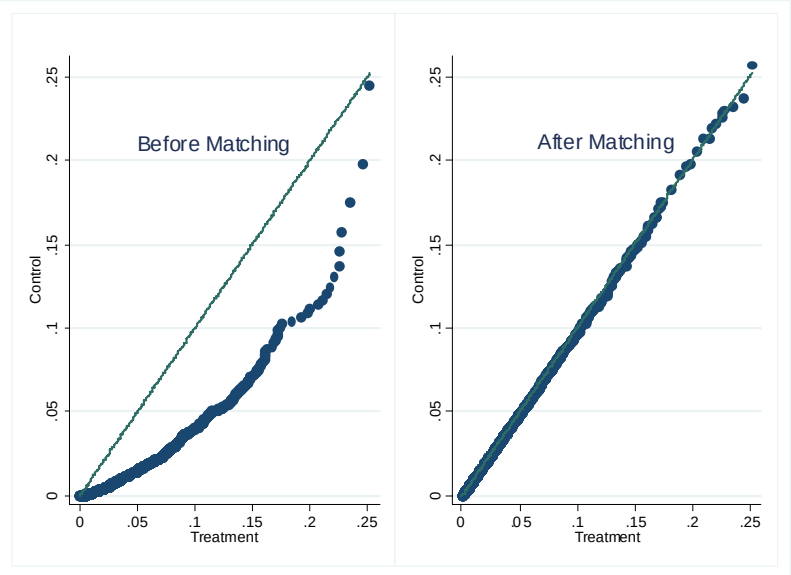


Figure 5 QQ- plot of the propensity score before and after matching – Regime 2

Table 8 Dummy covariate Imbalance after Matching - Matching Regime 2

Dummy variable Covariates	*Year of Observation (percent)*													Average over years
	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	
Localization															
Hedmark and Oppland	51.9	51.3	51.9	49.4	50.0	52.3	56.0	55.4	46.5	47.6	51.4	49.3	48.0	50.0	*50.8*
South Eastern Norway	51.9	53.3	52.3	49.7	51.1	50.8	50.9	50.0	50.3	46.1	51.9	57.9	51.5	46.1	*51.0*
Agder and Rogaland	56.3	51.4	51.3	55.6	48.4	49.6	50.9	52.3	50.5	57.3	52.3	47.3	53.3	48.3	*51.8*
Western Norway	51.4	51.3	50.6	50.3	48.5	51.9	50.3	53.1	54.6	48.3	45.4	50.0	51.5	56.0	*50.9*
Trøndelag	51.8	59.3	50.7	43.2	46.1	55.9	55.0	42.3	47.8	57.7	54.9	47.4	46.4	49.0	*50.5*
Northern Norway	43.3	46.5	44.6	51.0	54.7	46.8	47.0	46.3	48.9	54.5	51.3	48.1	48.1	52.1	*48.8*
Business sectors															
Other Industries	61.1	46.2	58.5	48.1	50.0	51.0	55.0	46.8	46.8	48.7	51.3	50.0	66.7	45.7	*51.8*
Construction	45.7	46.2	50.0	50.5	52.0	46.4	51.0	53.4	50.6	49.4	51.4	47.4	42.9	50.0	*49.1*
Property Management	51.2	51.1	59.0	46.0	46.9	51.1	60.5	41.8	48.9	53.7	51.2	42.3	46.8	54.1	*50.3*
Information Technology	45.9	47.4	56.8	44.0	47.8	57.6	43.9	63.0	53.3	53.3	52.0	66.7	44.8	44.8	*51.5*
Expert-knowledge services	61.5	55.0	40.0	60.0	43.3	48.1	50.0	52.0	50.0	55.0	55.6	52.6	45.5	50.0	*51.3*
Furniture Industry	58.8	64.7	45.8	61.1	61.1	68.8	57.9	50.0	52.6	40.0	47.1	50.0	53.8	57.1	*54.9*
Food industry	52.9	50.0	50.0	50.0	47.4	52.9	50.0	47.4	50.0	42.9	50.0	53.3	44.4	50.0	*49.4*
Plastics industry	54.5	56.5	59.1	54.2	48.1	56.5	43.3	54.5	57.1	48.0	52.2	52.4	44.0	50.0	*52.2*
Tourism industry	51.2	55.0	46.2	51.0	52.1	56.1	58.3	51.3	44.2	47.5	52.9	41.5	48.6	63.0	*51.3*
Wood products	60.4	51.7	46.9	51.7	58.5	54.4	58.8	52.8	52.8	56.0	52.9	50.0	57.1	51.9	*54.0*
Commodity trade	45.9	47.4	56.8	44.0	47.8	57.6	43.9	63.0	53.3	53.3	52.0	66.7	44.8	44.8	*51.5*
Shipbuilding industry	43.8	36.8	58.3	53.8	58.3	50.0	57.1	40.0	80.0	53.8	44.4	57.1	61.5	47.1	*53.0*
Size															
Medium-sized enterprises	53.3	48.6	50.0	47.9	54.8	53.1	47.1	54.4	52.5	58.3	57.1	44.3	55.1	51.0	
Average over dummies	*52.3*	*51.0*	*51.5*	*50.6*	*50.9*	*53.2*	*51.9*	*51.0*	*52.2*	*51.1*	*51.4*	*51.3*	*50.3*	*50.6*	

5. Analyses – the choice of models

The nature of the data confronts us with two basic choices; (1) the choice between competing risk modeling *or* the acceptance of bankruptcies as a single absorbing state, and (2) the choice between assuming time is measured discrete or continuous. Also, in the preprocessing of the data we have used two different matching regimes in the construction of our matched pair datasets. Thus we also have to assess the difference between the two matching regimes and decide which to use for further analysis.

5.1 Competing risk or single event

It is clear from Table 2 (page 197) that about half of the firms left the market for other reasons than bankruptcies. The official Norwegian Company Registry in Brønnøysund report 52 of the 410 firms as simply "no longer in the registry". The registry does, however, have information that 24 of the firms we cannot find in the registry were involved in mergers or acquisitions in the year prior to their last registration. Thus, it is relatively likely that the category "no longer in the registry" may imply that they are still active and even that this is a positive outcome e.g. that the company is sold at a profit and thus that in some cases exit may be a success story. In other cases "no longer in the registry" may mean that the company does not make sufficient money and, thus, phasing out is a rational action although not necessarily a success story. Bankruptcy, on the other hand, is usually not to be regarded as a desirable outcome. Thus, based on these lines of reasoning we find it reasonable to use bankruptcy as the single absorbing event of interest in the further analyses.

5.2 Discrete or continuous time

The available data from Dun & Bradstreet are yearly registered accounting records including information on industry, localization, year of establishment and of number of employees.

Thus, we do not have *event histories*, i.e. we have not have knowledge of the exact dates of events. The exception is the records for bankruptcies from the official Norwegian Company Registry in Brønnøysund where we have the exact dates for the declaration of bankruptcies. There is usually a gap in time between first warnings, declarations and final closure. In this period of negotiations and frantic efforts to avoid bankruptcy, some deliver complete accounting reports as required, most companies do not. This leaves us with two problems. First, we do not have complete records for the last years prior to bankruptcies; second, we have a kind of hybrid data which have exact dates for the ending time of spells but inexact dates for the starts of spells. The latter problem implies that a model that uses discrete time is a more appropriate choice than models that assume that time is measured continuously. If we choose to use a discrete time model, however, we throw away information and run into problems with the differences between the exact dates for bankruptcies and the clear-cut dates needed for discrete modeling (Allison, 1982). In the following we will report results from single spell episodes for both discrete and continuous time.

5.3 Matching in the year prior to intervention versus yearly matching

The results in sections 4.4 and 4.5 indicate that matching regime 2 has better balancing properties than matching regime 1. Matching regime 1 is in line with the standard procedures for propensity score matching for cross-sectional data (Caliendo et al., 2005) where the propensity scores are estimated on covariates observed one year prior to the intervention. Matching regime 2 is, however, very unusual. Matching regime 2 facilitates *time varying covariates* and creates a different propensity score for each observation year. The advantage of this kind of preprocessing is that we are able to create datasets of pairs that are maximally equal to each year, conditional on our vector of covariates. With one-to-one matching this means that we may assign a different comparison case to the treatment case in

question every year over the observation period. Thus we construct two comparable single spells where we ignore the history of the control units. Since time varying external forces may be the most decisive factors for survival this comparison strategy may be more robust to the effects of unobserved variables (unobserved heterogeneity).

Table 9 Number of events for matching regime 1 and 2

	Matching Regime 1 (Conditioning on covariates in one year prior to intervention)		Matching Regime 2 (Conditioning on covariates in every year over the observation period)	
	Treatment	Control	Treatment	Control
Censored Cases	356	363	354	349
Events (bankruptcies)	52	45	54	59
Total	408	408	408	408

As shown in Table 9, the results of the two matching regimes differ with respect to the number of events in the treatment and control groups. The difference between the numbers of events in the treatment groups occurs because two cases outside the region of common support matches two firms that later on went bankrupt, and thus, two treatment cases are discarded. Otherwise, the most striking difference is the unequal number of cases with events in the control groups. While matching regime 1 results in a lower number of bankruptcies in the control groups than matching regime 2, Table 9 does not tell us much with respect to the choice between the two matching regimes. Their differences are not surprising. Since matching regime 1 implies that we follow each firm in the control group over its entire life span from the year prior to intervention (the year prior to the treatment group's participation in the FRAM program) until event eventually occurs or we leave the

observation window in 2005. We also know that at any point in time, provided that a firm has managed to stay in business so far, the chances of surviving until the end of the following year is improved. This is a truism in the sense that this is always the case for any episode. The point to be made here is that matching regime 1 forces this phenomenon to be present while we have no knowledge of the prevalence of a corresponding effect in matching regime 2 where any comparison case is void of history and thus, the survival probability in the next year has a random distribution.

In the further analyses, based on the presumption of better properties, we base our models on the preprocessing resulting from matching regime 2. However, we present some rudimentary demonstrations of the differences between the two matching regimes.

6. Analyses – The Hazard-rate Framework

6.1 The continuous time model

The fundamental dependent variable within the hazard-rate framework (Petersen, 1993) is *the hazard*, in the continuous case defined as $r(t) = \lim_{t' \to t} \frac{\Pr(t \le T < t' \mid T \ge t)}{t' - t}$ where T is a random variable that represents the duration from the beginning of an episode, t_0 (for simplicity and in accordance with conventions we assume $t_0 = 0$) until a change in the dependent variable i.e., a transition from the origin state (here, the firm is active) to the destination state (here, the firm is bankrupt) occurs (Blossfeld, 1995). Thus, the basic for the single episode data we intend to model is of the form $r(t) = \lambda(t, x)$ where x represents our covariates in this regression type model and λ represents the functional forms. Since the hazard rate is an abstraction that is not directly observable we also demonstrate outcomes in terms of the directly observable survivor function $G(t) = 1 - (F(t) = \Pr(T > t)$ where $F(t) = \Pr(T \le t)$ is the probability distribution of T. The specific *shape* of the time dependence for the hazard rate may take many forms and have specific interpretations such as in the parameterization known as the *Hernes distribution* (Hernes, 1972) where the hazard of getting married is a decreasing function of time. Other common forms are an increase which levels off, such as the risk of fatal events after childbirths (e.g. a log-logistic model) or a rate that does not change over time (the exponential model). We have no knowledge about the shape of the time dependence in the case of the survival of firms. Our basic hypothesis is that participation in a leadership-training program should in general *reduce* the probability of failures. Hence, e.g. an increase in the risk of failure immediately after participation should not be expected. However, we use various parameterizations of time dependence in order to investigate model misspecification e.g. potential effects of unobserved heterogeneity, i.e.

effects of variables *not* included in our models. We also use the Cox proportional hazards model which leaves the function of time unspecified. In log- form the Cox model can be written in the regression form as $\log\lambda(t,x) = \alpha(t) + \beta' x$ where β is the familiar effects of the independent x-variables and $\alpha(t)$ is the unspecified function of time. Note that for all analyses here we assume no left censoring, that is, all censored cases are censored to the right i.e. is observed at *t*, but not *t+1*. Our observation window runs until 2005 and firms still active are considered as right censored.

6.2 The discrete time model

The discrete-time model (Allison, 1982) is quite similar to the continuous time models but assumes a rearrangement of our matched pair datasets so that each individual episode constitutes one case, thus an individual firm that is active for say five years and then goes bankrupt is repeated five times, if it is observed six times it is repeated six times and so on. This process of expanding the datasets go on until an event occurs or the case is censored. Using the rearranged dataset the discrete model can be estimated with a logistic regression model which in its logit form can be written as $\log\left[p/(1-p)\right] = \alpha_t + \beta' x$ where p is the proportion of cases with events, i.e. events are coded as ones and censored cases are coded as zeros. This model is very similar to the Cox-model in the sense that the α_t - part does not carry essential information, it simply picks up the effects of the dummy variables which is coded as ones for each time-specific line of data, zeros otherwise.

When comparing the continuous and the discrete model some discrepancies may arise that are due not to the true difference between the two models but the fact that the discrete model

cannot be estimated with gaps within the time span over which it is estimated. In our dataset there are no events in the first year after the intervention in the treatment group. Thus, we have to manually change the data to make sure that time starts at 1 and not at 2. The consequences of these changes should be minor.

6.3 Results – graphical displays

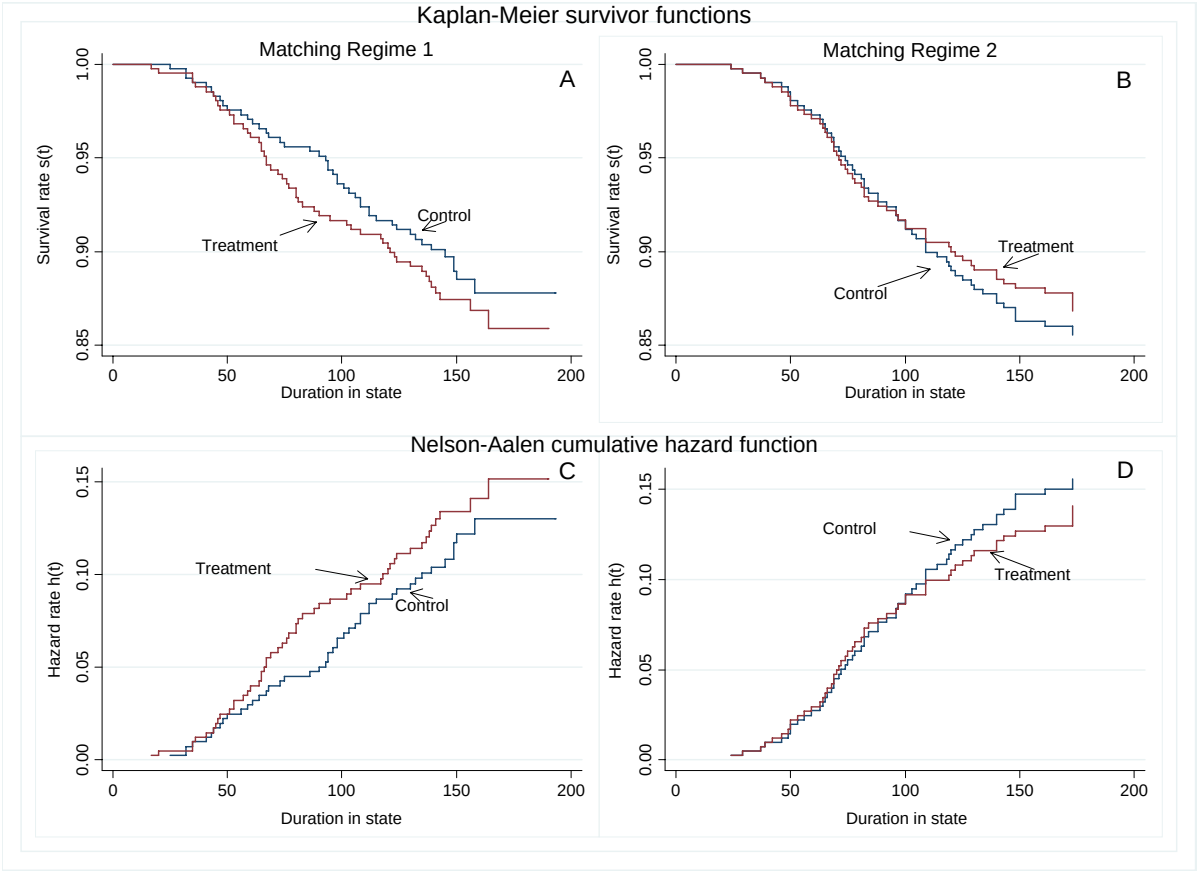


Figure 6 Survivor and cumulative hazard functions for matching regimes 1 and 2

Assuming that time is measured continuously and using treatment and control as two different strata, Figure 6 shows very similar patterns of survival and cumulative hazards over time, regardless of the chosen matching regime. Log-ranks test of the differences between the survivor functions for the treatment groups and the control groups (Table 10) show no

significant differences in the survivor functions between strata for both matching regimes 1 and 2. Thus, the first impression is that there are no differences between the firms that participated in the FRAM program and comparable firms that did not participate (the control group) and we could arrive at the conclusion that leadership training has had no impact upon the firm's chances of avoiding bankruptcies.

Table 10 Log-rank test of the difference between survivor functions

	Matching Regime 1		Matching Regime 2	
	Events observed	Events expected	Events observed	Events expected
Control	45	48.81	59	56.39
Treatment	52	48.19	54	56.61
Total	97	97.00	113	113.00
	χ^2 = 0.60		χ^2 = 0.24	
	Pr > χ^2 = 0.438		Pr > χ^2 = 0.623	

The combined impression from the graphical displays and the log-rank tests implies that there is little difference between the two preprocessing regimes. The balancing properties of matching regime 1 are inferior compared to matching regime 2 and we believe that the influences of unmeasured variables are less severe in matching regime 2 since we use yearly matches. Thus, in the further analysis we are based upon matching regime 2 only.

6.4 Results – simple regression models – discrete vs. continuous time

We can also think about the difference between the treatment group and the control group in terms of an independent dummy variable, coded one for the treatment group and zero for the control group, and inspect the effects of participation in the FRAM program by means of regression models. The regression approach facilitates comparisons between the continuous

time model and the discrete time model. Table 11 below (with an excessive number of decimals) shows the difference between estimates.

Table 11 Comparisons of the estimates from the discrete vs. continuous time model[8]

	Coef.	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
Discrete (Logistic)	-.0917061 *N = 7691*	.189972	-0.48	0.629	-.464045	.280633
Continuous (Cox)	-.0923134 *N = 816*	.188335	-0.49	0.624	-.461443	.276817

Discrete	**Odds ratio**	**Std. Err.**	**z**	**P>|z|**	**[95% Conf.**	**Interval]**
(Logistic)	.9123732	.1733256	-0.48	0.629	.6287353	1.323967
Continuous (Cox)	Haz. ratio .9118194	.1717275	-0.49	0.624	.6303732	1.318924

Despite the fact that we manually changed the data to avoid missing events in the first year after intervention, the estimates are strikingly similar, almost down to the third decimal. However, a dichotomized treatment/control variable as an independent variable only confirms what we already knew, the confidence intervals for coefficient estimates include zero and hence, it seems evident that participation in the FRAM program is unrelated to the likelihood of a future bankruptcy. Table 11 does, however, indicate that the approximation to continuous time may be appropriate. Thus, we proceed by using the continuous time model.

6.5 Results – overlooked differences and unobserved heterogeneity

The lack of difference between the treatment group and the control group may hide differences that are not easily detectable in the Cox proportional hazard model. Also, a

[8] We have omitted the results for the twelve dummy-variables involved since they are not central to our argument.

closer look at exhibit B in Figure 6 shows that the survivor function for the two groups cross and thus, the *proportional* hazard assumption is violated. Also, a closer inspection of the development of the most likely cause of bankruptcy, *short-term liabilities*, indicates that there are differences between the two groups that should affect the hazard rates.

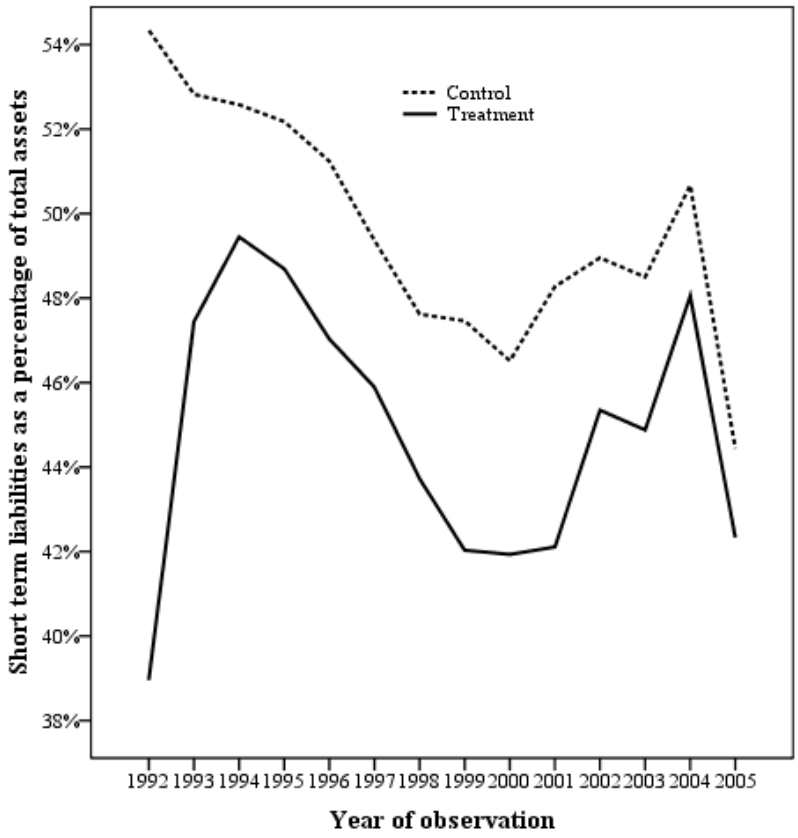


Figure 7 The evolvement of short term liabilities

The figure above shows that short time liabilities are lower in the treatment group than in the control group over the observation period but also that liabilities have increased in the treatment group in the first part of the intervention period. Note that these are fairly rough measures, simple means based on relatively few cases, in particular in the 1992 and 1993. It may imply that there are differences due to the selection processes that could affect the probability of bankruptcy that is not picked up by our preprocessing. Generally, we believe

that the *liability of newness* (Freeman, Carroll, & Hannan, 1983; Singh, Tucker, & House, 1986; Stinchcombe, 1965) is the single most important factor for explaining bankruptcies. This is also the reason why newness was included as a covariate in our matching vector. The result of the preprocessing with respect to this variable is shown in Figure 8.

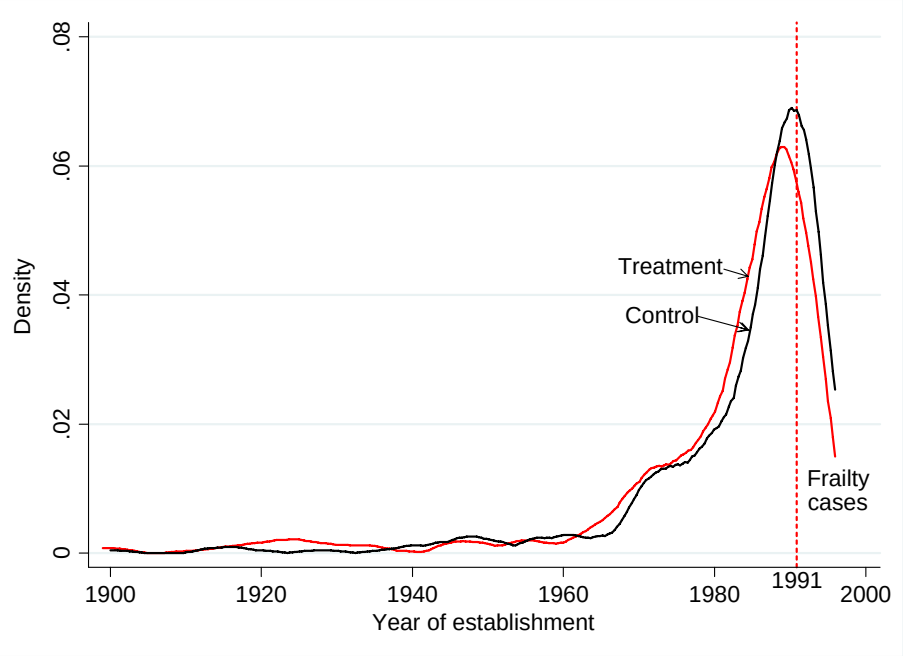


Figure 8 Newness and the frailty of firms

Figure 8 indicates that the "degree of newness" is well balanced between the treatment and the control group. Closer inspection shows that around 33% of the firms in the control group and 24% in the treatment group were established in or after 1991. In line with the liability of newness hypothesis and in the terminology of Vaupel et al. (1979) we label these firms *frailty cases* simply because we believe they are more vulnerable due to inadequate financial solidity and also because the leadership in these firms may be more receptive to new ideas about how to run the firm. The latter point, that they are more easily influenced by ideas

about leadership may mean that the consequences of their vulnerability differ between FRAM participants and non-participants.

Figure 9 Short-term liabilities for the frailty cases

The figure above indicates that solidity may be a problem for the newly established firms (the frailty cases). The solidity conditions for new firms are prevalent in both the treatment group and the control group, the question is whether participation in the FRAM program causes leaders to act differently to leaders who do not participate and to what extent these actions affect future outcomes.

Since we have very little theoretical grounds for assuming any form of duration dependence we ran a Weibull regression model with a Gamma frailty term separately for the treatment group and the control group in order to inspect possible differences between the population level hazards and the individual level hazards. The difference between the individual level and the population level hazard is, by definition, assumed away by the proportionality assumption in the Cox proportional hazard model, hence, we simply chose a very common parameterization of the duration dependence to be able to distinguish between the two levels.

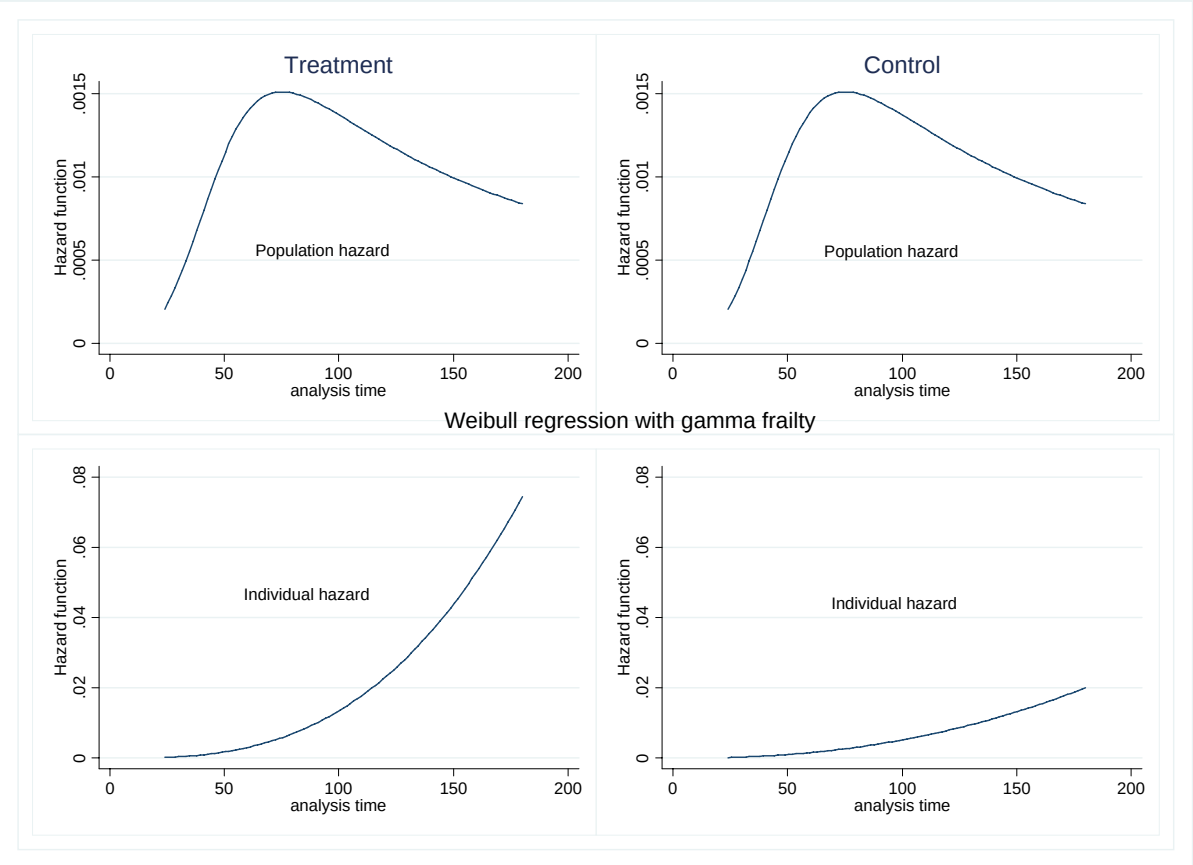


Figure 10 Weibull regressions with gamma frailty

The population hazard is the hazard averaged over all survivors while the individual hazard is the hazard that the individual unit faces. In the proportional hazard model these two magnitudes are the same since all individual units are assumed to be identical. Unobserved

heterogeneity may arise precisely because all units are *not* identical. Figure 10 shows that while the hazards at the population level for both groups follow a familiar pattern of a slight increase which levels off over time, the patterns at the individual level are different between the treatment group and the control group. The scale on the y-axes indicates that the rise-level of pattern at the population level is in fact relatively flat while differences at the individual level are of greater magnitude. Thus, what is worth noting from Figure 10 is that while the hazard falls at population level, the hazard at individual level continues to rise in the treatment group but not to the same extent in the control group. The phenomena may indicate a frailty effect caused by a heterogeneity that assures that the population hazards decline over time regardless of the shape of the hazard the individual firms face. The worst cases scenario is that this means that the FRAM program stimulates leadership behaviors which reinforce risks in the frailty group.

Clearly, the Weibull model with the frailty term is just another parameterization of duration dependence; we have not substantiated that a particular negative interaction between what is learned in the FRAM program and the hazard rate is present. We can however, test whether frailty effects are present in the statistical sense by means of a likelihood ratio test and then try out whether the presence of frailty alters estimates for other variables that may affect the hazard rate. The result from this test is shown in Table 12 below. Note that the test show the results from "shared" frailty defined as a dummy variable for firms established after 1991, whereas Figure 10 shows "unshared frailty" which was used in the graphical check of the possible presence of frailty. While Figure 10 led us to discover localized frailty by means of simply specifying a Weibull model with Gamma frailty included, "shared" frailty takes us one step further by allowing us to specify the frailty variable. Moreover, by

specifying the "shared frailty" we assume that similar observations share frailty, even though frailty may vary from group to group.

Table 12 Weibull and Cox regression with and shared frailty and frailty as a variable[9]

	Weibull	Weibull with shared Frailty	Weibull with Frailty included as variable	Cox Proportional Hazard Model
Treatment/control	-0.987*	-0.994*	-1.257**	-1.321**
	(0.531)	(0.531)	(0.546)	(0.552)
Concern ownership	0.390	0.435	0.428	0.455
	(0.423)	(0.424)	(0.425)	(0.425)
Short-time debt	0.020***	0.018***	0.020***	0.021***
	(0.005)	(0.005)	(0.005)	(0.005)
Short-time debt *& TC*	0.016**	0.018**	0.016**	0.017**
	(0.008)	(0.008)	(0.008)	(0.008)
Frail firms			0.198	0.199
			(0.270)	(0.270)
Frail firms *& TC*			0.928**	0.926**
			(0.384)	(0.384)
Constant	-12.544***	-12.448***	-12.783***	
	(0.911)	(0.936)	(0.923)	
Shape parameter ρ	1.922	1.940	1.963	
	(0.169)	(0.170)	(0.171)	
Frailty parameter θ		0.080		
		(0.098)		
Observations	816	816	816	816
Log-Likelihood	-342.08	-339.62	-333.97	-690.86

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed); Standard errors are given in parentheses
& TC means interaction with Treatment/control
Likelihood-ratio test of theta=0: chibar2(01) = 4.91 Prob >= chibar2 = 0.013

Table 12 shows that frailty is present[10] but does not have any noteworthy impact upon other estimates in the model. The table also demonstrates that there are only modest differences between the estimates from the Cox proportional hazard model and the Weibull model after including frailty and the interaction between treatment and frailty as independent variables,

[9] Note that these regressions are based upon time-varying covariates such as short time liabilities as a percentage of total assets. The Stata program facilitates the inclusion of time varying covariates both for the discrete and the continuous hazard rate models, thus in the regression models we have 8886 records for 816 subjects. For the time-constant variables we have the exact same results whether the datasets are expanded or not. The id-variable in Stata's st-setup takes care of the adjustments.

[10] Note that the frailty parameter theta is evaluated by a likelihood-ratio test as shown beneath the table, dividing the coefficient by its standard error has no meaning.

even though we are well aware that the proportionality assumption for the Cox model may be violated. Table 12 shows some consequential results; first and foremost we note that the sign of treatment/control variable now is negative and it is significantly different from zero. These results are in striking contrast to those from the log-rank test in Table 10 and the regression results in Table 11 which indicates no significant difference in hazard rate of survival rate between the treatment and the control group. Figure 11 shows that the relatively decreased hazard of bankruptcy is lower in the treatment group regardless of the estimation model used; both the parametric Weibull model and the semi parametric Cox model generate approximately the same patterns for the cumulative hazards.

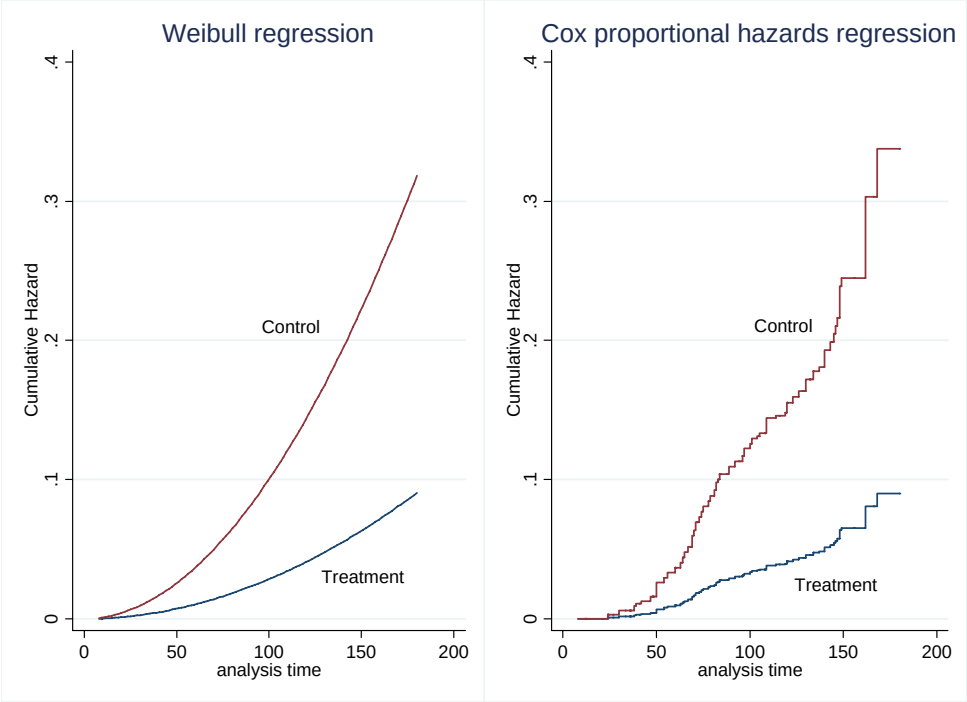


Figure 11 Cumulative Hazard for treatment and control after estimation

The most important findings reported in Table 12 are the interactions between short-term liabilities and treatment and between frailty (firms established after 1991) and treatment. It should not be a surprise that larger short-time liabilities significantly increase the likelihood

of bankruptcy but it is less reassuring that the interaction between treatment and short-time liabilities goes in the same direction, although to a lesser degree. Both estimates of the effects of short-time liabilities do, however, have insufficient power i.e., the magnitude of the coefficients' effects sizes are simply too small to be detectable with the sample size we have[11]. The estimates of the effects of the treatment/control variable and the interaction between treatment and frailty has power near one even after assuming considerable multiple correlations among the independent variables. The most important finding is, however, that the interaction between treatment and frailty which indicates that participation in the FRAM program may increase the probability of failure i.e., bankruptcy for the newly established firms. The interaction between participation and increased hazard can also be shown by a simple graph of the cumulative hazard within the treatment group and the control group as shown in Figure 12.

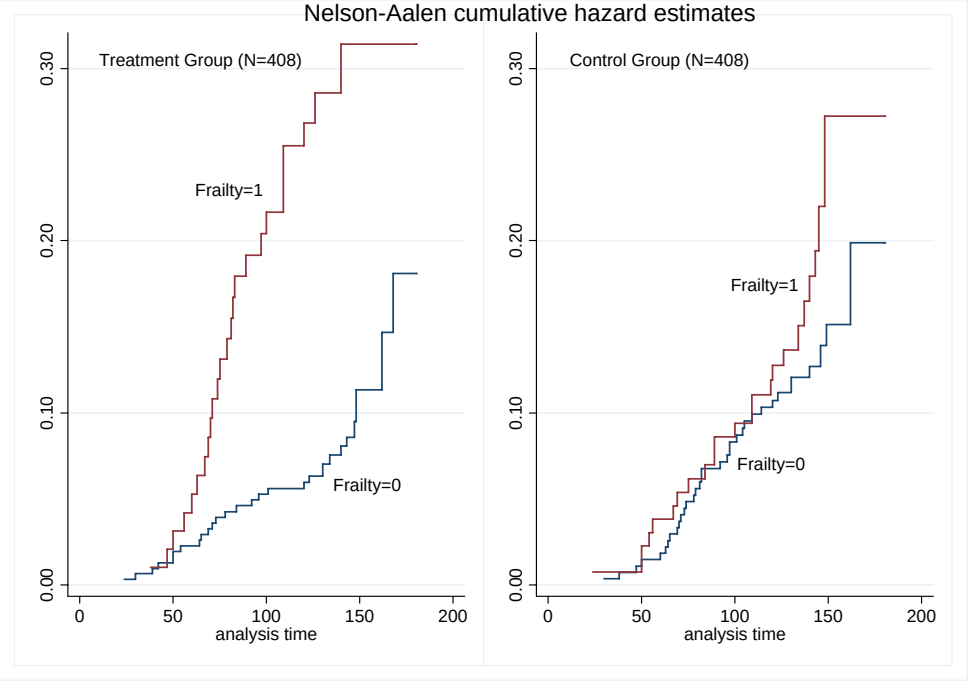


Figure 12 Cumulative hazard rates for treatment and control by frailty

[11] Power is estimated by means of Stata's stpower cox – routine.

7. Discussion

7.1 The trustworthiness of our findings

As we rely solely on administrative records, the presented findings depend heavily on our preprocessing procedures. On the other hand, observational studies like this exhibit a high degree of transparency; each individual firm can be identified by a unique number and lists of accounting information are publicly available. Thus, anyone can reproduce our matched pair data and check to what extent comparisons make sense. The experimentalists would most likely maintain that the chosen methodologies have their shortcomings. However, as pointed out by Heckman, even randomization does not remove selection bias; it balances the bias between the treatment group and the control group (Heckman & Smith, 1995). The tradition of comparing observational studies to pure experiments (Dehejia, 2005; Dehejia & Wahba, 2002; Heckman, Ichimura, Smith, & Todd, 1998; Smith & Todd, 2005) tends to claim that there is always a better answer. The tradition of considering the experiment as the "gold standard" has, however, its limitations. Apart from the problem that random assignment of leadership training programs nay not be the ideal task, selection bias may arise from sources that are hard to level out even by means of randomization. Selection effects can be generated by missing data on the common factors affecting participation and outcome, or it may occur when random assignment causes the kind of units participating in the program to differ from units participating in the program as it normally operates (Heckman et al., 1995). This phenomenon is usually labeled *randomization bias* and could easily occur if SMEs simply receive a message saying that they by lottery have been so lucky to be invited to participate in the FRAM program. Another source of randomization bias is changes in participant behavior that operates via reactions towards participation and is measurable prior to treatment. Furthermore, *substitution bias* may occur when members of

the control group gain access to close substitutes of the treatment under consideration. In e.g. training programs for small business firms, this phenomenon is likely to happen when someone in the control group recognizes that they are denied a service and react by seeking similar services offered elsewhere. Thus, experimentation is no "magic bullet" and neither is matching techniques. The best we can hope for is that we have succeeded in constructing datasets of matched pairs that are as good as possible and thus, that major confounders are eliminated.

7.2 The implications of our findings

The initial direct comparisons of the survivor functions for participants in the FRAM-program and non-participants implied no difference, i.e., initially, we found no support for our main hypothesis that participation should lower the probability of bankruptcy. However, after controlling for short-time debt, frailty defined as recently established firms and their respective interactions with treatment, the picture changes and it now appears as if the FRAM-program in fact has a *positive* impact in the sense that it lowers the probability of bankruptcies. The analyses indicate that this positive result comes at a cost; the participating firms that are established after 1991 have a considerably higher probability of failure than their non-participating comparison cases. Evaluated by the Cox model the estimate of 0.926 for the interaction of treatment and frailty means that participation in the FRAM- program would increase the probability of failure with $100*((e^{0.926})-1) \approx 152\%$. Thus, the overall positive outcome of the program results from controlling out the negative effects on the newly established firms. Thus, the most vulnerable firms are affected negatively by the program and most probably contrary to intentions.

We may speculate over the possible mechanisms underlying this result. One could be the *moral hazard* argument i.e., participants felt safe in the assumed protection of Innovation Norway and, eager to conform to the program's goal of increased turnover changed their behavior to a more debt-generating risky behavior. Other explanations could imply that lack of practice led the inexperienced leaders to literally accept everything that was taught in the seminars thereby ignoring necessary attention to their own firm's financial situation. Or, it could be the case that a level of robustness is necessary if a company is to be exposed to a change-inducing leadership program.

References

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. 2001. Implementing Matching Estimators for Average Treatment Effects in Stata. ***The Stata Journal***, 1(1): 1-18.

Allison, P. D. 1982. Discrete-Time Methods for the Analysis of Event Histories. ***Sociological Methodology***, 13: 61-98.

Bator, F. 1958. The Anatomy of Market Failure. ***Quarterly Journal of Economics***, 72(2): 311-400.

Blossfeld, H. P. a. R., G. 1995. ***Techniques of Event History Modeling***. Mahweh, N.J.: Erlbaum.

Caliendo, M., & Kopeinig, S. 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. Bonn: Forschunginstitut zur Zukunft der Arbeit.

Cohen, J. 1988. ***Statistical power analysis for the behavioral sciences*** (2nd ed.). Hilsdal, NJ.: Erlbaum.

Commission, E. 1995. The Innovation Programme: Office for Official Publication of the European Communities.

Cook, T. D., & Campbell, D. T. 1979. ***Quasi-experimentation Design & analysis issues for field settings***. Boston: Houghton Mifflin Co.

Cook, T. D., Shadish, W. R., & Wong, V. C. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. ***Journal of Policy Analysis and Management***, 27(4): 724-750.

Cook, T. D., & Steiner, P. M. 2009. Some Empirically Viable Alternatives to Random Assignment. ***Journal of Policy Analysis and Management***, 28(1): 165-166.

Cook, T. D., Steiner, P. M., & Pohl, S. 2009. How Bias Reduction Is Affected by Covariate Choice, Unreliability, and Mode of Data Analysis: Results From Two Types of Within-Study Comparisons. ***Multivariate Behavioral Research***, 44(6): 828-847.

Dehejia, R. 2005. Practical propensity score matching: a reply to Smith and Todd. ***Journal of Econometrics***, 125(1-2): 355-364.

Dehejia, R. H., & Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. ***Review of Economics and Statistics***, 84(1): 151-161.

Diaz, J. J., & Handa, S. 2006. An assessment of propensity score matching as a nonexperimental impact estimator - Evidence from Mexico's PROGRESA program. ***Journal of Human Resources***, 41(2): 319-345.

Foster, E. M. 2003. Propensity score matching - An illustrative analysis of dose response. ***Medical Care***, 41(10): 1183-1192.

Freeman, J., Carroll, G. R., & Hannan, M. T. 1983. The Liability of Newness - Age Dependence in Organizational Death Rates. ***American Sociological Review***, 48(5): 692-710.

Gavron, R., Cowling, M., & Westall, A. 1998. ***The Entrepreneurial Society***. London: Institute for Public Policy Research.

Glazerman, S., Levy, D. M., & Myers, D. 2003. Nonexperimental versus experimental estimates of earnings impacts. ***Annals of the American Academy of Political and Social Science***, 589: 63-93.

Gu, X. S., & Rosenbaum, P. R. 1993. Comparison of multivariate matching methods: structures, distances, and algorithms. ***Journal of Computational and Graphical Statistics***, 2: 405-420.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. 2007. Combining propensity score matching and group-based trajectory analysis in an observational study. ***Psychological Methods***, 12(3): 247-267.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. 1998. Characterizing selection bias using experimental data. ***Econometrica***, 66(5): 1017-1098.

Heckman, J. J. 1989. Causal Inference and Nonrandom Samples. ***Journal of Educational Statistics***, 14(2): 159-168.

Heckman, J. J. 1998. ***Characterizing selection bias using experimental data***. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., Cameron, S. V., & Schochet, P. Z. 1992. ***The determinants of public-sector and private-sector training***. Washington, D.C.: U.S. Dept. of Labor Bureau of Labor Statistics.

Heckman, J. J., & Hotz, V. J. 1989. Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs - the Case of Manpower Training. ***Journal of the American Statistical Association***, 84(408): 862-874.

Heckman, J. J., Ichimura, H., & Todd, P. E. 1997. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. ***Review of Economic Studies***, 64: 605-654.

Heckman, J. J., & National Bureau of Economic Research. 1991. ***Randomization and social policy evaluation***. Cambridge, Mass.: National Bureau of Economic Research.

Heckman, J. J., & Smith, J. A. 1995. Assessing the Case for Social Experiment. ***Journal of Economic Perspectives***, 9(2): 85-110.

Heckman, J. J., & Smith, J. A. 1998. ***Evaluating the welfare state***. Cambridge, MA.: National Bureau of Economic Research.

Hernes, G. 1972. The process of entry into first marriage. ***American Sociological Review***, 37: 173-182.

Holland, P. W. 1986. Statistics and Causal Inference. ***Journal of the American Statistical Association***(81): 945-970.

Kmenta, J. 1986. ***Elements of Econometrics*** (Second edition). New York: Macmillan.

Lewis, D. 1973. ***Counterfactuals***. Cambridge, MA: Harvard University Press.

Lundstrøm, A., Boter, H., Kjellberg, A., & Ohman, C. 1998. Svensk Småføretags-politik: Struktur, Resultat och Internationella Jamførelser. Ørebro, Sweden: FSF.

McIntosh, M. W., & Rubin, D. B. 1999. On estimating the causal effects of DNR orders. ***Medical Care***, 37(8): 722-726.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. 2004. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? ***Review of Economics and Statistics***, 86(1): 156-179.

Nesheim, T., Kvitastein, O. A., Lines, R., Grønhaug, K., & Espedal, B. 1997. Evaluering av FRAM-programmet i SND. Bergen: Foundation for Research in Economics and Business Administration (SNF).

Pearl, J. 2000. ***Causality***. Cambridge: Cambridge University Press.

Petersen, T. 1993. Recent Advances in Longitudinal Methodology. ***Annual Review of Sociology***, 19: 425-454.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. 2009. Unbiased Causal Inference From an Observational Study: Results of a Within-Study Comparison. ***Educational Evaluation and Policy Analysis***, 31(4): 463-479.

Porter, M. 1990. ***Competitive Advantage of Nations***. London: Macmillan.

Romer, P. M. 1986. Increasing Returns and Long Run Growth. ***Journal of Political Economy***(94): 1002-1036.

Rosenbaum, P., & Rubin, D. B. 1985a. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. ***American Statistician***, 39(33-38).

Rosenbaum, P. R., & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. ***Biometrika***, 70: 41-55.

Rosenbaum, P. R., & Rubin, D. B. 1984. Estimating the Effects Caused by Treatments - Comment. ***Journal of the American Statistical Association***, 79(385): 26-28.

Rosenbaum, P. R., & Rubin, D. B. 1985b. The Bias Due to Incomplete Matching. ***Biometrics***, 41(1): 103-116.

Rubin, D. B. 1993. Tasks in Statistical-Inference for Studying Variation in Medicine. ***Medical Care***, 31(5): Ys103-Ys110.

Singh, J. V., Tucker, D. J., & House, R. J. 1986. Organizational Legitimacy and the Liability of Newness. ***Administrative Science Quarterly***, 31(2): 171-193.

Smith, J. A., & Todd, P. E. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? ***Journal of Econometrics***, 125(1-2): 305-353.

Stinchcombe, A. L. 1965. Social Structure and Organizations. In J. G. March (Ed.), 1 ed.: 142-193. Chicago: Rand McNally & Company.

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., Fine, M. J., Hough, L., Karpf, M., Lave, J., Li, Y. H., Medsger, A., Redmond, C., & Ricci, E. 1995. Propensity Score Adjustment for Pretreatment Differences between Hospitalized and Ambulatory Patients with Community-Acquired Pneumonia. ***Medical Care***, 33(4): As56-As66.

Vaupel, J. W., Manton, K. G., & Stallard, E. 1979. Impact of Heterogeneity in Individual Frailty on the Dynamics of Mortality. ***Demography***, 16(3): 439-454.

Methodological Cleavages in Evaluation Research

Some Consequences for the "What If" Question of Practical Evaluations

Olav A. Kvitastein[1]
Norwegian School of Economics and Business Administration

April 2010

Abstract

The article discusses how strong directional pressures and the framing effects of the New Public Management mantra may set off institutional processes that mould the methodologies of applied evaluations. The emergence of evaluations as an integral part of New Public Management (NPM) and the substandard argument, i.e., the notion that applied evaluation research should accept lower methodological standards than discipline research, produce an unfortunate dualism. Moreover, the tension between the protagonists of naturalistic inquiry and the devotees of probabilistic causation within the evaluation community may be harmful to the public confidence in social science based evaluations. This tension is specific to the social sciences, has no parallel in other sciences, and tends to create the impression that most evaluations are disputed.

The history of evaluation has produced a number of sub-fields based on different parent disciplines. Strong paradigmatic commitment to own sub-discipline and corresponding rejection of competing world-views may lower the overall trust in evaluations. Local legitimacy in evaluations refers to the problem that some methods are recognized within specific sub-disciplines only. Since New Public Management is predominantly based on a popular notion of economics, this is the story of how pluralism meets monism.

The focus of the article is the potential for bias in the evaluations of governmental projects or programs that is created by the methodological state of affairs. The peculiarities of the various 'schools' of the field of evaluation and the question to what extent strong adherence to a particular 'school' can be a source of bias is debated.

[1] Norwegian School of Economics and Business Administration, Department of Strategy and Management, Breiviksveien 40, NO-5045 Bergen, Norway, olav.kvitastein@nhh.no

The paper pertains to *summative evaluations* of public programs and projects where the initialization, finalization and intentions are identifiable. A central theme throughout the paper is the agenda setting nature of the overt superficial "first" question of applied evaluation studies, namely, "did it work?" This question demands claims about *effects*, regardless of how well the program or project under investigation lends itself to causal investigations. Ambiguities of results provide opportunities for strong stakeholders to set the premises for debates.

While qualitative methodologies may be well suited for evaluations of ongoing program/projects where the intent is to improve performance, inquiries that have low transparency and rely heavily on subjective impressions and interpretations are less suited for summative evaluations.

Thus, the conclusion is that the final, overall conclusion about the merit or worth of a program should, whenever possible, imply a minimum of disputable subjective interpretations and follow strict rules and procedures that are backward traceable and have a high degree of transparency.

Keywords

Evaluations, Methodology, Confirmation Bias

JEL Classification Codes

C80, C81, C90

Contents

List of tables

1. Introduction

1.1. New Public Management, Accountability and Evaluation

Over the past three decades applied evaluation research has changed from being an activity mainly concerned with efforts to improve the primary schools system, (Worthen, Sanders, & Fitzpatrick, 1997) to become an integral part of the professional apparatus of the bundle of managerial reforms referred to as New Public Management (NPM) in Europe, Australia and New Zealand, and as the Reinventing Government movement in the US (Christensen & Laegreid, 1999; Christensen, Laegreid, & Wise, 2002; Hood, 1996; Pollitt & Bouckaert, 2000). New Public Management introduces a new language of public administration where the emphasis on stability, rules and responsiveness to the law, is replaced with a vocabulary that accentuates change, decentralization, responsiveness to consumers, performance, and the need to "earn" rather than to "spend" (Maor, 1999). Public demand for *accountability* is both central to the justification for New Public Management *and* a prime responsibility for the new administrative regime. Hence, the NPM movement has prepared a new place for social science research-based reasoning that has made applied evaluations one of today's fastest growing areas of social science. This revival of social science research is, however, not uncontroversial. While some researchers welcome the new financial support and the apparent resurgence of applied social science, others are more hesitant to welcome an activity they believe will necessarily introduce double methodological standards. That is, lower methodological and professional standards are unavoidably enforced in situations where most evaluation contracts demand swift answers to complicated questions. Common reasons for accepting methodological sub-standards for applied evaluations is acceptance and respect for the fact that policy-making processes follow their own pace and logic.

Thus, if social science arguments are to be taken into consideration at all, researchers have to comply with the standards the evaluation task at hands allows for.

1.2. Evaluation's problematic relation to varying parent disciplines

The conception of social science implied by applied evaluations, that researchers do not identify problems, they simply solve predefined tasks, is hard to accept for many scholars. Reactions to decreasing availability of funding for general social research, combined with growing opportunities for finance via evaluation contracts, take many forms. The formation of professional organizations like the *European Evaluation Society* and the *American Evaluation Association* and the birth of professional journals dedicated to evaluation studies, are all positive adaptive institutional responses. The tension between the research communities within the established disciplines and the emerging sub-discipline dedicated to evaluation research is, however, the other side of the coin. Even though evaluation research is well accepted as a legitimate research activity, hints of opportunism and insinuations about biases in reported outcomes, are not entirely absent. For the legitimacy of evaluation research, it is of importance to reveal the alleged sources of bias.

1.3. The paradoxical revival of the social sciences

Several prominent scholars (Haveman, 1987; Nathan, 1988; Aaron, Gramlich, Hanushek, Heckman, & Wildawsky, 1990) maintain that the status of the social sciences has been on a downward slide since the seventies. It may seem paradoxical that evaluations flourish in times when the social sciences in general have a downturn, and it is not all that clear that the new confidence in evaluations and hence, social science based public policies, implies a renewed vote of confidence in the

social sciences. As an integral part of the managerial orientation of NPM, the awakening of evaluations is interpretable, but badly in need of clarifications. What used to be called a report is now an evaluation, and thus, carries a promise of something more than just a report. Evaluation is a semantic magnet (Vedung, 2000) with a positive power that lends itself easily to a message of confidence. On the other hand, the *American Evaluation Association* is, to my knowledge, among the very few professional organizations that openly admits to and publicly discusses that a negative reputation is a problem for their profession (Donaldson, 2001). Recently, some evaluators have also expressed frustration over the tension between the quest for confidence and the feeling that a widespread acceptance of relativism characterizes both practical evaluation reports and the professional recommendations from leading journals in the field. With respect to methodology, the impression is that "anything goes" (Adelman, 1996).

1.4. The ivory tower and the pettiness of evaluations

However, applied evaluations have a built-in propensity to reveal deep-seated problems in the social sciences. The practical, interest ridden setting of applied evaluation research, tends to unveil unpleasant questions about ideological underpinnings of the theories in use, and ambiguities of the methods employed. Practical implications of the theories that guided interventions may disclose unwarranted side effects, and the prescribed methods may fail to provide trustworthy information about outcomes. Frustration in the research communities seems to disperse in two directions; towards overconfidence or retreat. Overconfidence tends to surface as an expression of near unconditional faith in theories in support of the arguments set forth and retreat can be described as the backfire of the researcher's methodological training.

When the prescribed remedies fail, the researcher renounces, not only the recommended methods, he/she rejects the entire role as a researcher and takes flight into apparently more favorable roles, such as, say, the role as a judge. The canon of opportunism in this respect, is the so-called “fourth generation evaluation” (Guba & Lincoln, 1989) where the idea of evaluation as the search for quality, merit, worth, etc., is rejected in favor of the idea that it is negotiation which is the issue (Scriven, 1993). That is, negotiation between stakeholders with different interests or worldviews is the essence of program evaluation.

1.5. New directions in evaluation research

Guba and Lincoln’s (Guba, 1990; 1989) work presents considerable creativity and insights. With the exception of their open prejudice to an undefined group of their fellow evaluation researchers who they label “positivists”, for anyone interested in research methodology, the *Fourth Generation Evaluation* (1989) is worth reading. They describe the “positivists” in the following way: “Convinced that there exists some single, true reality, driven by natural laws, open to discovery and harnessing by the methods of science, positivists reject all relativist views, of which constructivism is one, as not only seriously in error but pernicious and repugnant” (Guba et al., 1989:16). Clearly, by negatively attaching a historical philosophical position, or more precisely, a set of related philosophical positions, to a generalized third person, the “positivist” evaluator, they create a straw man, made up for the sake of the argument. By doing so, they tell us that constructivists not only reveal and recognize the existence of social constructions, they also create them. Indeed, people constantly create and communicate social constructions and it is of great importance that we seek to understand and unveil the underlying processes.

The Fourth Generation Evaluation constitutes one of the two extremes of evaluation; on the one hand, the assumption-dependent devotees of the cost-benefit methods of neo-classical economics, on the other hand, the followers of constructivist inquiry, supposedly dependent on fewer unrealistic assumptions concerning human nature and behavior. Both traditions have problems with the empirical contents of their analysis and they both run the risk of being victims of dominant stakeholders, simply because of predictable methodological flaws. Another reason for taking these two traditions as extremes is that they both seem overtly convinced of the supremacy of their positions and that they are located on opposite sides of the qualitative- quantitative divide.

1.6. Lower methodological standards for evaluations

A point of departure for the article is the recommendation put forward in the seminal article by James S. Coleman almost forty years ago (Coleman, 1972) where he claimed that policy-analyses have to accept lower methodological standards than should be accepted in mature discipline research. He pointed out that most methodological development in the social sciences has been directed toward the development of theory and testing, refinement and confirmation of these theories. Thus, there was a need to develop methods tailored for the testing and evaluation of specific policies. I argue that the qualitative movement in evaluation research and also the progress within the cost-benefit tradition tends to overlook the advances in methodologies aimed at the evaluation problem. Recent development within observational studies (Rosenbaum, 1995) and the counterfactual account of causality (Pearl, 2000) has proven to be of immense importance to evaluation studies. The fact

that most of these advances have taken place within fields like econometric, medical statistics and pure statistics, i.e., fields outside the adjacent border of evaluation studies and cost-benefit analysis signals that the boundaries between disciplines do not necessarily serve the progress of science.

1.7. The applicability of qualitative methodology for evaluations

In line with Reichenbach (1938), I agree that qualitative inquiries are usually better suited for theory development and new discoveries than most quantitative methods that seem more correctly applied to confirm what is already indicated by other methods. In some cases, however, for instance a change in infant mortality, discoveries rely exclusively on quantitative methods. I argue, however, that recently developed quantitative methodologies are far better suited for many evaluations, in particular in situations where the cry for documentation of impacts of public programs or projects dominates. Substituting qualitative methods for quantitative methods as a matter of principle does not improve the reputation of applied evaluation.

1.8. The many Schools of Evaluation

For historical reasons, evaluation research in the US is characterized by a number of "schools" that differ with respect to methodology, willingness to accept how problems are defined by principals and enthusiasm for making evaluative claims. In Europe the situation is somewhat different; whereas respect for formal rules and relatively stringent methodologies is marked within the various otherwise very different schools or sub-disciplines of evaluation research, it is hard to find a corresponding stringency within the European evaluation community. With respect to methodology we can dimly perceive American discipline and European anarchy.

The cleavage along the qualitative – quantitative division is, however, significant both in Europe and the US. The most important consequence of the multitude of "schools" for applied evaluations of public programs/projects is confusion. The different world views that underlie the methodological divisions produce divergent answers to similar questions and create impressions of widespread professional disagreement which, for the wider public audience, is difficult to interpret.

1.9. Outline of article – including expressions and definitions used

The paper contains concepts and expressions that may not be familiar outside the evaluation community. Below we explain them as they are used in the text:

Summative evaluation, the kind of applied evaluation that is the topic here is usually defined as evaluations conducted *after* completion (or stabilization of an ongoing program) of a program/project and *for* the benefit of some *external* audience or decision-maker e.g., funding agency, governmental ministry or other principal.

Formative evaluation, often contrasted with summative evaluation, is usually conducted *during* the development or improvement of a program/project, *for* the in-house-staff of the program *with the intent to improve* (Scriven, 1991).

Evaluation research simply refers to evaluation done in a serious scientific way. Since this paper only concerns summative evaluation of public programs/projects the term refers to *applied* social science research conducted with the intent to investigate the results of the initiative in question.

Naturalistic evaluation or methodology is used about an approach that minimizes the use of technical jargon, the need for prior technical knowledge about statistics and instead emphasize the use of metaphor, analogy, informal inference, reasons-explanations, interpretations, meanings and so on.

Throughout the paper I will use the terms *method* and *methodology* almost interchangeably but method will refer more to applied *techniques* while methodology will refer to the broader scope of the world views and logic that guides the kind of inquiry in question. Thus **qualitative inquiry** may refer to methodologies in terms of a range of methods of investigation that emphasize naturalism while **quantitative inquiry** refers to thinking that recognize numbers and emphasizes statistical modeling.

The article is outlined as follows. Section 2 gives a brief introduction to the contemporary context of evaluations and section 3 outlines the research questions that can be summed up in three theses: a) The substandard argument is invalid and possibly harmful to the reputation of applied evaluations. b) Confirmation bias may be a threat to evaluation studies based on qualitative methodologies. c) Strong commitment to a specific school of evaluation may produce bias in conclusions. Section 4 sketches the development of evaluations and gives a crude overview over the qualitative – quantitative debate. Section 5 discusses several sources to bias in evaluation research and section 6 sums up the presumed consequences. Section 7 is concerned with the way contractual practices are organized and section 8 completes the discussion.

2. The Context of Public Evaluations

2.1. The Move from Bureaucratic to Managerial Control

While the twentieth century was governed by the principle of a politically neutral civil service offering impartial policy advice to the elected government of the day, the last part of the 20th century has witnessed an increasing importance of public servants in process of policy formulation and implementation (Plowden, 1994). The demise of Weberian bureaucracy and the belief that a stronger, more competent and vigilant bureaucracy is the result of modernity, is among the basic premises behind many newer governmental reform initiatives (Wallis & Dollery, 1999). Over the last three decades, NPM initiatives have produced fundamental and ubiquitous institutional change in the nature of public administration in most western industrial democracies. These changes have had a variety of consequences, some of them of considerable concern for evaluation practices. The paradigmatic essence of NPM is reducing and deregulating bureaucracy, using market mechanisms and simulated markets to conduct government action, devoting responsibility downward and outward in organizations, increasing productivity, energizing agencies, and empowering employees to pursue results, improve quality, and satisfy customers (Carrol, 1998). The "worldview" of NPM presupposes that "something" fundamental happened in the 1980s that changed the field of public administration. The (assumed) appearance of new forms of governance, new relationships between citizens and their government and between the public, private, and non-governmental sectors fundamentally altered the processes of policy-making. The nature of this assumed change is explicitly expressed by former U.S. Vice-President Al Gore, when he claims that Americans view themselves as customers of the government rather than as citizens (Gore, 1993).

The new orientation implies the substitution of self-interests for the more complex norms of traditional bureaucracy as the baseline for the design of governance, and the substitution of the customer for the citizen as the basic individual unit of democratic society. These changes have altered the core concept of evaluation, *accountability*, from its many-valued meaning *democratic accountability* to the single-valued meaning *economic accountability*, and thus, brought new topicality to look at evaluations as an *agency problem.*

Agency theory assumes opportunism in the relationship between principal and agent. I do not believe that traditional bureaucracy was devoid of opportunism. My basic point is that, under the realm of NPM there may be, at least at the individual level, a better payoff for opportunism, in particular for the top civil servants (Laegreid, 2000). A market-based system for evaluation contracts implies agency problems. By its reliance on self-interests as the driving forces of governance, NPM acknowledges self-interests as a more legitimate concern for the individual than was the case under the traditional bureaucratic regime. This may be viewed as a concession to opportunism and may interfere with the choice of research strategies. To understand some implausible evaluation outcomes it is of importance to come to grip with the interplay between evaluation methodologies, the institutional mechanisms of research practices that pertain to the various methodologies, the emerging routines of evaluation practices, and the potential rewards or penalties for opportunism.

2.2. Agency theory – the heart and soul of New Public Management

Agency theory (Jensen & Meckling, 1976), the analysis of principal-agent relationships, in which one person, an agent, acts behalf of another person, a principal, lays

at the heart of NPM and can be viewed as the dominant idea behind structural reforms. Hence, when consulting companies and research institutions are competing for evaluation contracts, the contest for contracts is in line with the ideological underpinnings of the new managerialism. It is an implicit, albeit naïve assumption, that competitive bidding guarantees the best quality in the evaluation process. Apart from overlooking the differences in normative traditions between consulting companies and research institutions, this line of thinking confuses the costs of the governmental contract with the costs of the consequences of the evaluation task. Also, this line of reasoning fails to recognize that the very same contractual theory that justifies the competitive bidding process can be applied to the evaluator – evaluation management relationship. Whenever a contractual relationship can be identified, agency theory can be applied. From a methodological point of view, there are very few reasons to believe that the less costly evaluation contract produces more reliable results than the more expensive contract. It is more likely that minimum funding for the evaluation task will induce methodological shortcuts, thus undermining the trustworthiness of results. On the other hand, an abundance of research funding does not guarantee the quality of evaluations. Hence, the assumption that competition is a quality optimizer does not apply. The agency framework applied to the contractual relation between evaluator and evaluation management may, however, shed some light on the problem. There is, however, a peculiar feature with this relation, namely that it cannot be understood as a principal agency relationship *without violating the fundamental rationale for undertaking the evaluation task.* The basic rationale for evaluations is the *independence* between the researcher and the evaluation management.

2.3. Institutional theory – reinforcing human habit-taking[2]

Institutional theory (Powell & DiMaggio, 1991; Scott, 1995) offers considerable insights for understanding evaluation processes. It is, however, not easy to comprehend institutional theory as an unambiguous coherent theory. The concept of an *institution* has been used in different ways by numerous authors, and to cover diverse phenomena. In accordance with the purpose here, I use W. Richard Scott's omnibus definition: *"Institutions consist of cognitive, normative, and regulative structures and activities that provide stability and meaning to social behavior. Institutions are transported by various carriers – cultures, structures, and routines – and they operate at multiple levels of jurisdiction"* (Scott, 1995:33). In Scott's conceptualization, institutions are multifaceted systems, including symbolic systems, cognitive constructions and normative rules, and regulative processes carried out through and shaping social behavior (Scott, 1995). The view of institutions as both systems and processes facilitates discussions of the interplay between the overarching ideas from economic theory that settle what the operating notion of evaluation should be and the design of the evaluation process. It also provides for a way to interpret the cognitive mechanisms that make diverse empirical representations converge across researchers within homogeneous subgroups when methodological rules or norms are unclear or absent. Noteworthy, even though constructed and maintained by the individual, institutions assume the guise of an impersonal and objective reality. Institutional mechanisms require little or no conscious mobilization of will or effort (Scott, 1995) and hence, makes it possible to discuss the pitfalls of various research strategies without invoking accusations of deliberate distortions of assumed empirical representations.

[2] The term "habit-taking" is used since, according to Charles Sanders Peirce, concepts are habits because they relate to or organize feelings, which are themselves connective tissues of thought. (Peirce, Collected Papers, VI, 137-138)

Scott's brilliant summing up of institutional theory provides a way to systematize the many active mechanisms of institutional processes that are suggested in the vast literature on the subject. Table 1 gives a sketch of the *basis of compliance*, the *mechanisms* at work, the *logic* of the particular process, the *indicator* for the process and the *basis of legitimacy* according to what Scott labels the *three pillars of institutions*.

Table 1. The three pillars of institutions

	Regulative	***Normative***	***Cognitive***
Basis of compliance	Expedience	Social obligation	Taken for granted
Mechanism	Coercive	Normative	Mimetic
Logic	Instrumentality	Appropriateness	Orthodoxy
Indicators	Rules, laws, sanctions	Certification, accreditation	Prevalence, isomorphism
Basis of legitimacy	Legally sanctioned	Morally governed	Culturally supported, conceptually correct

Scott's table (1995:35) sweeps thousands of pages of research and his synthesizing conceptualization has a strong intuitive appeal. Although there are no well-defined boundaries that allow for precise definitions of the various concepts introduced, it is easy to grasp each and every concept as *descriptive of a social relationship* that is possible to imagine when we think about how an evaluation process is unfolding. We can reflect about the relations between the evaluation management and the evaluator, or we can envision the relations between researchers with different personalities. Scott's conceptualization provides a rich source for speculations and reflections about what the relations could be and what they could produce of outcomes that are simply due to the nature of the relations alone. This way of using concepts to induce reflections over relations e.g., Cassirer (1910) Bourdieu & Coleman (1991a) has a distinct flavor of European sociology, is easily applicable for discussing hypothetical configurations of relations, and is indicative of Scott's synthesizing capacity.

The three pillars cover diverse scholarly approaches, the discussion of institutions and different understandings of institutional processes. The *regulative* pillar clearly resembles the approaches typical of economists and economic historians, the *normative* pillar mostly that of sociologists and the *cognitive* pillar mirrors the approaches most likely to be found among psychologists and organization theorists. Scott's (1995) typology also contains suggestions about the factors or types of repositories or 'carriers' that sustain or reproduce institutions. *Cultures* as carriers transmit schemes that inform and constrain behaviors, *social structures* carry expectations connected to networks of social positions and role systems, while *routines* carry habits, standard operating procedures and other repetitive behaviors or trained incapacities[3].

Table 2. Institutional pillars and carriers

	Pillar		
Carrier	***Regulative***	***Normative***	***Cognitive***
Culture	Rules, laws	Values, expectations	Categories, typifications
Social structure	Governance systems, power systems	Regimes, authority systems	Structural isomorphism, Identities
Routines	Protocols, standard routines	Conformity, performance of duty	Performance programs, Scripts

The carriers (Scott, 1995:52) provide yet another way for reasoning about likely mechanisms or processes behind outcomes. When we observe outcomes that are most likely biased in one or another direction, it is possible to make constructive speculations about what processes would be the most likely candidates for producing such a result, without the requirement that we are able to directly observe the process

[3] The origin of this term has been credited to Randolph Bourne, Thorstein Veblen, Kenneth Burke and James G. March.

in question. Institutional theories provide frameworks for speculations and discussions.

Thus, I take agency theory as indicative of the dominant underlying epistemology of New Public Management. In its most general form, agency theory implies a coherent epistemological understanding that leaves little grounds for choosing among perspectives that exclude e.g. general equilibrium theory or welfare theory. Neither does it easily embrace other understandings that conflict with mainstream economics. These features of NPM and of agency theory facilitate strong framing effects of the chosen perspective.

Provided that we do not accept lower methodological standards for evaluation research than for discipline research, we should start looking for sources of bias where they are most likely to be found. That is, in the human imperfection as revealed in institutional theories, given the strain of the task of conducting unbiased judgment under conflict, which should be expected to prevail in most evaluations.

3. Three theses concerning sources of bias in evaluations

3.1. Is the substandard argument an acceptable excuse?

The substandard argument (Knudsen & Waerness, 2001) arises from a misinterpretation of James S. Coleman (Coleman, 1972) and maintains that policy research[4] is different from discipline research and hence, that lower standards should be accepted due to lack of time and obligations to the pace of the policy making process[5]. This point of view is, I believe, on different grounds, widely accepted among evaluation researchers. I also think it has had its impact on evaluation practices, even though the substandard argument has never been prominent in debates over methodologies. The discussions have more often focused on the connection between theories and methods. The concern has often been over inconsistent or weakly established links between theory and methods. This article follows along the same lines, but in addition maintains that in applied evaluation research the perspective has to be extended to include the peculiarities of the evaluation context. The idea that evaluation research is different from other kinds of research is neither revolutionary nor new. I argue that the impact of the substandard argument has, however, mainly been in the direction of serving as an excuse for opportunism e.g. by asserting that a program or project can be evaluated based on other grounds than the kind of reasoning that is acceptable as social science standards. Thus, evaluation research can base its legitimacy on scientific standards without having to meet the terms. Hence, the first thesis to be dealt with is:

[4] I am aware of the distinction between policy analysis and evaluation (e.g., Geva-May, I., & Pal, L. A. 1999. Good Fences Make Good Neighbours. ***Evaluation***, 5(3): 259-277.). However, even though I agree that a difference exists, I hold the substandard argument to be valid also for evaluations.

[5] Noteworthy, Coleman in the same article rejects the substandard argument.

I. Accepting substandard evaluations leaves the debate over evaluation methodology obsolete and makes it hard to establish the status of evaluations.

By accepting the substandard dictum we may only contribute to worsen the reputation of evaluation research; a problem that recently justified a special issue of the *American Journal of Evaluation* (Donaldson, 2001). In the long run, acceptance of substandard may also affect the legitimacy of social science in general and hence, be erosive to the very foundation of evaluations.

3.2. Does qualitative inquiry make the position of the researcher harder?

I believe that the qualitative-quantitative dichotomy is in many respects a false dichotomy that stimulates an unfortunate debate. I do, however, think there are some features of this debate that may help us to identify sources of what is usually labeled *confirmation bias* (Baron, 1981; Jones & Sugden, 2001; Krems & Zierer, 1994; Nickerson, 1998; Pollard, 1983). The term confirmation bias implies unidirectional bias. That is, the bias goes mainly in favor of dominant stakeholders. Hence, I contend that the context of evaluations tends to influence the outcomes of evaluations and that researches who see qualitative methods as the only valid way, give away means for resisting pressure towards concordance. That is, I do not suggest any difference between the devotees of qualitative inquiry and other researchers with respect to moral courage. I simply suggest that the qualitative researcher lacks the opportunity to appeal to method as independent judgment. The lack of support from method as a "third person" may hold back the researcher's opportunity for creating a distance between him/herself as a professional and the results he/she presents. This distance may be essential in situations with conflict between researchers and stake-

holders, however weak the distinction between the researcher and this "third person" might be. Hence, my second thesis of inquiry is:

II. *Qualitative inquiry may produce confirmation bias in evaluation studies by amalgamating the research and the researcher, and hence confusing psychological self-defensive mechanisms and professional argumentation. Thus, conflict may be experienced as insult and support for argument as praise, constituting a mechanism that increases the probability of agreement between the researcher and the dominant stakeholders.*

By leaving little or no room for methodology, the researcher risks being understood as a lawyer who continuously writes and rewrites the law he/she practices. Needless to say, this is a situation where any blame will be directed towards the person, not the rules. Thus, by invoking a discussion of this kind, the researcher leads the debate towards the fallacy of *argumentum ad hominem*, in soccer jargon, the fallacy of taking the man instead of the ball.

3.3. Is the division into "schools" of research a likely source of bias?

Also, the researcher's dedication to a specific sub-discipline or "school" of research may serve as a source of bias insofar that loyalty to the common understanding limits the range of valid outcomes. Thus, by excluding arguments that do not conform to the epistemological and ontological basis of the 'school', the researcher may cause disciplinary loyalty to become a source of bias in the direction of central beliefs held by the "school". The bias is hypothesized to be in the direction of "within school consensus", and reinforced when conclusions coincide with the interest of dominant stakeholders. Hence, my third thesis of inquiry is:

III. Strong commitment to a specific 'school' of thought may produce bias in a direction that ensures paradigmatic support.

Noteworthy, this proposal reflects my impression that at least some qualitative methodologies seem to be more directed towards paradigmatic support than towards traditional investigation. That is, some methodologies seem to search more for theory-laden observations than others. I do, however, accept the notion that theory-laden observations are hardly entirely avoidable regardless of methodologies. By the same token, actively seeking observations that are good candidates for confirmation and excluding observation that could lead to refutation violates the very notion of doing research.

3.4. The Scope of the Questions Proposed

In accordance with Coleman (Coleman, 1972) I believe that the substandard argument should be rejected in any kind of evaluation studies. That is, regardless of what kind of evaluation is carried out, there is no legitimacy for accepting lower standards for evaluation research than for discipline research. The argument that qualitative inquiry may produce confirmation bias in evaluation studies is however, only taken to be valid insofar that a question about outcomes, i.e., the effects of a project/program or its merit in other respects is invoked. That is, the argument may be applicable to both formative and summative evaluations, but for formative evaluations only to the extent the relative merit or worth of the project/program is questioned. For summative evaluations where the demand for evaluative claims is explicit, problems concerning bias are indubitably of importance. For evaluations where the intentions behind the evaluation is of a different nature, say, program adjustment based on analyses of implementation processes or other more intermediate

concerns, problems of bias in conclusions may still apply, but is of a different nature. Thus, producing exaggerated expectations is different from exaggerating claims about outcomes. By the same token as the distinction between formative evaluations, process evaluations, summative evaluations, effect evaluation is fuzzy due to a great many different uses of these two concepts. For the sake of simplicity it is most convenient to restrict the scope of our arguments to be valid only for summative evaluations.

4. Some development paths of evaluations

4.1. The Substandard argument

James S. Coleman was prescient when he, 31 years ago, stated that "There is no body of methods, no comprehensive methodology, for the study of the impact of public policy as an aid to future policy". … "The systematic methods they (the social scientists) have developed are methods for aiding the disciplinary development, not for such externally-imposed irritants as the evaluation of public policies. The philosophical bases of the methods are all oriented in this direction – toward the development of "theory", toward the generation and testing of hypotheses to confirm, refine and enlarge theory"… "A central implication is that a coherent and self-conscious methodology for studying impacts of public policy must be developed, if the social sciences are to function as policy sciences" (Coleman, 1972). Clearly, it is unfortunate that his foresight has been turned into an argument for accepting lower standards for evaluations than for others fields of social research. It is also clear that the tremendous impact of Coleman's article is also rightfully justified because of other insights offered. In the aftermath, especially the insights elaborated around the four features of policy research he considered most important for describing the difference between policy research and discipline research, namely time, language, conflict and information.

1. *Time*. Policy research has to follow the time-schedule of decisions in the world of action in accordance with the pace of politics, and accept to work based on the information available at the time when answers are demanded.

2. *Language*. Researchers who are doing policy research have to communicate with people who are not familiar with and do not master the concepts and jargon of the specialized researchers.
3. *Conflict*. Policy research is characterized by conflicting interests. Results may interfere with existing power-relations and resource allocation. It is difficult to avoid the researchers being dragged into or affected by conflicts.
4. *Information*. In the world of action, comprehensive explanations and additional information may be of major importance. The discipline's research requirements for elegance and parsimony do no apply. In the world of action one has to use simple models despite prevailing complexities. Policy research has to accept "the world as it is" in a way very different from what can be expected to be valued in the world of discipline research.

According to Coleman (1972:2), it follows from these differences that: "it is important at the very outset to distinguish sharply between a methodology that has as its philosophic base the testing and development of theories, and a methodology that has as its philosophic base a guide to action". Needless to say, this statement is completely in line with the fundamental premises and inspirations for this article. I truly agree with the argument that in many respects, there is a distinct difference between applied evaluation research[6] and discipline research. I also recognize the importance of a sharp distinction between a methodology for evaluations and a methodology for testing and developing theories. My reasons for not accepting Coleman's reasoning as a basis for the substandard argument is that there has been tremendous development in methodology the last thirty years aimed at precisely

[6] Noteworthy, Coleman does not insist on a sharp distinction between evaluation research and policy research.

these problems. Clearly, this development has had, as should be expected, a development marked by optimism over new methods, criticism, setbacks, and retries.

The first serious, and may be the most well known attempt to solve the methodology puzzle was Campbell and Stanley's (1963) celebrated work on quasi-experimentation. Their book started a wave of optimism and progress and moved the logic of experimentation and causality from the well-developed, isolated experimental setting of physics to the open field setting, carrying with it the lessons learned from psychological experimentation. A decade or so later, more problems are listed and optimism is less prominent in the follow up volume by Cook and Campbell (1979). The long list of possible threats to the validity of quasi-experiments reveals one major problem; how to analyze the impact of a project and program when randomized experiments are not a feasible option? You cannot, on moral grounds, expose people to random treatment for school achievements, health problems or other problems that seriously affects their fate, even if you were confident that the new knowledge gained would be to the benefit of mankind. In most cases you cannot assign random treatment for legal reasons. Hence, the problem that most public projects and programs involved non-random treatment, even when participants had to qualify for the program, remained. Of the early serious attempts to solve the problem was the work of Charles Judd and David Kenny (1981) and Christopher Achen (1986) who mainly sought the solutions within the econometric tradition. The major breakthrough came with the work of the Harvard statisticians Paul Rosenbaum and Donald Rubin (1983) but was not well communicated to the social science community before it was criticized, reinterpreted and reiterated by econometricians (Heckman, Smith, & Clements, 1997). Hence, the methodological developments requested by Coleman

in the early seventies were established first nearly thirty years later. Regrettably however, the peculiarities of evaluations, so clearly spelled out by Coleman were discovered by others and taken as an argument for accepting substandard research. Unfortunately, the solution to the methodological problems seems to have appeared so late that the evaluation disciplines and sub-disciplines that have surfaced in the meantime seem to have developed impenetrable boundaries.

Noteworthy, incorporating the new methodological insights may reduce some of the problems caused by the four features of policy research that according to Coleman (1972) constitute the major difference between policy research and discipline research, namely *time*, *language*, *conflict* and *information*. *Time* is likely to remain a major problem, but *language* may be less of a problem since the logic of experimentation is easy to communicate. *Conflict* may be reduced, both as a consequence of the ease of communication and because the new methods may increase perceived legitimacy and authority of results. Hence, there are many reasons to reject the substandard argument, and thus, to invoke a renewal of the debate over methodologies.

The logic of experimentation may, however, be more convincing than it deserves to be and the counterfactual accounts of causality are, from a variety of points of view, rightfully contested. I do, however, argue that the new quantitative methodologies outperform older methods in their ability to answer rather simple questions in a more reliable way. The new methodologies are, however, not well suited to answer the more complex questions that are often the ambitions of qualitative methods. Hence, it can be argued that the new quantitative methodologies in some respects are inferior to qualitative methods. The point to put forward here is that the evaluation con-

text may not be the best time and place for the more complex questions. Dominant stakeholders will tend to give priority to the more simple questions concerning the correspondence between goals and outcomes, regardless of the extent to which researchers find such questions to be of any relevance.

4.2. The qualitative-quantitative methods debate

More than two decades ago John Van Maanen stated that "the label *qualitative methods* has no precise meaning in any of the social sciences" (Van Maanen, 1979:520). The absence of a precise meaning does not, however, disqualify this loosely connected bundle of methods, "since qualitative researchers tend to regard social phenomena as more particular and ambiguous than replicable and clearly defined" (Van Maanen, 1979). By the same token, it is not so easy to give an exact account of what methods should qualify for the label *quantitative methods*. Further, there is no reason to assume that researchers who use quantitative methods *a priori* take social phenomena to be less complex and ambiguous than the qualitative researcher. It is, however, probably not entirely wrong to say that, despite many assertions that qualitative and quantitative methods are not mutually exclusive and in usual should strengthen analysis when combined, the tension between the devotees of each camp have not leveled off over the years. On the contrary, many new textbooks express attitudes that come close to hostility towards any use of numbers besides necessary paging. The deadly serious, humorless rhetoric employed hints to a hermeneutics of suspicion[7], indicative of a linguistic turn[8], where, however, the text

[7] This term was coined by Paul Ricoeur, P. 1970. ***Freud and Philosophy: An Essay on Interpretation***. New Haven: Yale University Press.:27) to describe the three key intellectual figures of the twentieth century, who, in their different ways, sought to unmask, demystify, and expose the real from the apparent, namely Marx, Nietzsche, and Freud, the leading figures of the school of suspicion.

to be dissected is missing.

The book on Grounded Theory has obtained the status of a classic text (Glaser & Strauss, 1967). Many of the newer textbooks also give valuable contributions to the understanding of methodologies and the limitation of methods. Some are extremely well organized, e.g. Flick (2002) and others open new perspectives for the organization researcher by relating both research strategies and methods to established research traditions (Creswell, 1998). Many of these books could indeed be included on more curricula. The point is that qualitative methods open for insights that *add* to, not *substitute* other insights. As additions, new perspectives should receive a warm reception while as alternatives they can be erosive.

A particular feature of many textbooks on qualitative methods is the overwhelming number of methodologies and perspectives introduced. Creswell (Creswell, 1998) encourages students to become familiar with the research tradition of *biography*, *phenomenology*, *grounded theory*, *ethnography* and *case studies*. Instructions for all five traditions, usually taught in different courses to students of philology and seldom offered to social science students, are all covered in a single textbook. As shown in Table 3, the different perspectives introduced cover research traditions that require a wide range of training and skills. *Biography* is something that is usually understood as quite apart from what in general occupies the social scientist. The focus is on the life of an individual, a theme that is even further from the traditional tasks of the evaluation researcher. *Phenomenology*, a tradition heavily criticized for

[8] Attributed to the excellent, humorous phrase by the econometrician Arjo Klamer Klamer, A. 2001. Making sense of economice: from falsification to rhetoric and beyond. ***Journal of Economics Methodology***, 8(1): 69-75. "It was about then that I made my linguistic turn".

departing from Husserl's original intentions, is included with a focus on understanding. *Grounded theory* is introduced as a methodology with a focus on theory development, *ethnography* with a focus on cultural interpretation and *case study* methodology is introduced as a means for in-depth analysis of single or multiple cases.

The ambitions of a training project of these dimensions are praiseworthy, even though the realism of gaining thorough understanding of all these traditions can be questioned. Indeed, my experience from discussions with scholars from ethnography and cultural studies indicates that many of the courses offered in business schools tends to imprint the content of courses in ethnography and phenomenology with instrumental epistemologies, alien to the original theories. Case studies[9], on the other hand, are more familiar to the business student, but mostly applied as a teaching device, to a lesser extent as a research methodology.

A very informative way to organize and understand the different methodologies is introduced by Morgan (Morgan & Smircich, 1980). The different methodologies are organized along a continuum from subjective to objective approaches to social science, differentiated by their assumed core ontological assumptions and their basic epistemological stances. His rough typology is a helpful device for discussion, al-

[9] Noteworthy, however, Creswell, J. W. 1998. ***Qualitative Inquiry and Research Design: Choosing Among Five Traditions***. London: Sage. has no reference to Eisenhardt, K. M. 1989. Agency Theory: An assessment and review. ***Academy of Management Review***, 14(1): 57-73., the elsewhere most cited article on the theory building aspect of case studies.

though he has added some spice to the debate, in particular by using the term "concrete" quite frequently and using Skinner (1953) as an example of the "positivist".

Table 3. Dimensions of five research traditions in qualitative research

Dimension	**Biography**	**Phenomenology**	**Grounded Theory**	**Ethnography**	**Case Study**
Focus	Exploring the life of an individual	Understanding the essence of experiences about a phenomenon	Developing a theory grounded in data from the field	Describing and interpreting a cultural and social group	Developing an in-depth analysis of a single case or multiple cases
Discipline origin	Anthropology Literature History Psychology Sociology	Philosophy, Sociology, Psychology	Sociology	Cultural anthropology Sociology	Political science, sociology, evaluation, urban studies, other social sciences
Data collection	Primarily interviews and documents	Long interviews with up to 10 people	Interviews with 20-30 individuals to "saturate" categories and detail a theory	Primarily observations and interviews with additional artifacts during extended time in the field (e.g., 6 months to a year)	Multiple sources- documents, archival records, interviews, observations, physical artifacts
Data analysis	Stories Epiphanies Historical content	Statements Meanings Meaning themes General description of the experience	Open coding Axial coding Selective coding Conditional matrix	Description Analysis Interpretation	Description Themes Assertions
Narrative form	Detailed picture of an individual's life	Description of the "essence" of the experience	Theory or theoretical model	Description of the cultural behavior of a group or an individual	In-depth study of a "case" or "cases"

Source: (Creswell, 1998)

Combing Table 3 and Table 4 opens up large fields of inquiry that should inspire imagination and clarify how the same phenomenon could look very different dependent on how we choose to observe it.

Table 4. Basic Assumptions of the subjective – objective debate

	Subjectivist Approaches to Social Science					**Objectivist Approaches to Social Science**
Core Ontological Assumptions	Reality as a Projection of Human Imagination	Reality as a social construction	Reality as a Realm of symbolic Discourse	Reality as a Contextual field of Information	Reality as a concrete process	Reality as a concrete structure
Assumptions About Human Nature	Man as pure spirit, consciousness, being	Man as a social constructor, the symbol creator	Man as an actor, the symbol user	Man as an Information Processor	Man as an adaptor	Man as a responder
Basic Epistemological Stance	To attain phenomenological insight, revelation	To understand how social reality is created	To understand patterns of symbolic discourse	To map context	To study systems process, change	To construct a positivist science
Some Favored Metaphors	Transcendental	Language game, accomplishment text	Theater Culture	Cybernetic	Organism	Machine
Research Methodology	Exploration of pure Subjectivity	Hermeneutics	Symbolic analysis	Contextual analysis of Gestalten	Historical Analysis	Lab experiments, surveys

Source: (Morgan et al., 1980)

The data analysis part of Table 3 gives a rough impression of how one should go about data collection and a less than lucid guidance to data analysis. Table 4 demonstrates that explicating core ontological assumptions and basic epistemologies clarifies both the choice of research strategies and what kind of question will have priority in a given study.

Clearly, questions concerning the impact of a given project or program do not have high priority in a qualitative methods framework. On the contrary, questions concerning *understanding* of the nature of a given project or program are emphasized. Also, theory construction is an essential objective of qualitative inquiries.

4.3. Evaluation as a Domain of Schools

Evaluation, with its origin in education, psychology and sociology has over time evolved into a variety of "schools". These "schools" reflect *both* their origins, i.e. the basic epistemologies of their parent disciplines, such as economics, psychology, sociology and so forth, *and* the diverging ways of thinking in various subgroups *within* disciplines that constitute sub-disciplines occupied with evaluation research. The nature of the sub-disciplines varies from sociologists that on principle grounds renounce quantitative techniques, theorists of economic organization specializing in transaction cost economics to economists who prefers non-empirical, theoretical deductions. Sub-disciplines may be more or less institutionalized and boundaries may be more or less pragmatically upheld. Sub-disciplines can be identified and categorized based on the methodologies applied, basic epistemological beliefs, and how they view the evaluation task. Without any ambition of a comprehensive listing, only a brief sketch of the most important will be described here.

The distinction between the various schools *within* the evaluation community introduced here originates from Michael Scriven's work that sums up a long experience from evaluations at the theoretical, the institutional and the practical level (Scriven, 1993). Scriven emphasizes that his standpoint is the *consumer's* point of view, that is, his approach is a consumer-weighted view, rather than a management-weighted approach.

The 'decision support' school maintains that program evaluation is a part of the decision process of rational program management. This perspective is in particular accentuated in the work of Ralph Tyler (Stufflebaum, Guba, & Tyler, 1971) the

founder of educational evaluation. It is made explicit in the CIPP model (Context, Input, Process, Product) and plays a leading role in evaluation, in particular in educational evaluation.

1. **C***ontext evaluation* supports planning decisions. The clarification of needs facilitates more comprehensive definitions of goals.
2. **I***nput evaluation* serves as a structuring device for decision making by identifying available resources and relevant strategies for action. Design of projects/programs is then based on the plans or strategies with the best prospect for goal realization.
3. **P***rocess evaluation* supports decisions about implementations. How good were the implementation strategies? What are the threats to success? What revisions should be undertaken? After answering such questions, procedures can be monitored and adjusted.
4. **P***roduct evaluation* aims at the re-circulation of decisions. What is achieved? To what extend are the needs fulfilled? When should the mission be regarded as completed?

Many contemporary evaluation theorists basically share the 'decision support' view of evaluation although it is important to recognize that the CIPP model goes far beyond the basic idea of decision support into detailed checklists that cover most of what is involved in program evaluation (Scriven, 1993).

The 'relativistic' approach (Rossi, Freeman, & Lipsey, 1999) maintains the view that evaluation should be done by using the client's values as a framework, without any commitment by the evaluator to those values (Scriven, 1993). This view is probably the most common among evaluators, and implies that the idea of the possibility of value-free research is to some extent preserved. That is, there is a basic belief that the researcher possesses relatively neutral methodologies that can be used for providing the necessary answers, independently of the nature of the questions

asked. 'Relativistic' is Scriven's term, but implies nothing more than the main line in most social science departments, at least within the U.S. As far as I know, this is Scriven's personal opinion, not a verified widespread attitude. In the case, say, where a researcher was asked to evaluate only the beneficial outcomes of a project/program, I believe most researchers would reject the invitation, based on the nature of the question asked. If not, Scriven's (1993) term 'relativistic' is indeed appropriate.

The 'rich description' approach, often associated with the North Dakota school, (Kemmis & Stake, 1988; Stake, 1975, 1986, 1995; Stake, Easley, & Anastasiou, 1978) sees evaluation as a kind of ethnographic or journalistic enterprise, where evaluators report what they see without trying to make evaluative statements that infer on evaluative conclusions. This view also has wide support from many UK theorists and is characterized by a research strategy where the researcher has a detached, outside-looking in and neutrally observing kind of attitude toward the evaluation problem. They pretend to report neutrally what they observe, without drawing any conclusions. There is a flavor of relativism about the "*rich description school*" and it has much in common with the "*relativistic approach*" in the sense that both schools pretend toward neutrality and objectivity, avoid value issues and shrink from or ensure that they do not produce any evaluative claims. I believe the "*rich description*" approach has considerable merit as a strategy for meta analyses, but is otherwise a step away from what is traditionally regarded as program evaluation. Such meta-analyses can be of immense value for reflections over policies, but may produce ambiguities when confused with substantive evaluations.

The 'social process' school originates from a distinguished group of Stanford academics led by Leo Cronbach (1980). It is known for its rejection of the importance of evaluation for decision support purposes, for downplaying the need for accountability, and for the substitution of 'understanding' for evaluating them in the ordinary sense (Scriven, 1993). The approach emphasizes measurement models, and is probably close to ways of thinking that are common in the mainstream of the sociologically oriented members of the *American Evaluation Association.* The central point for Cronbach and his followers was not evaluation in the consequence-oriented meaning of the word as it is used in this paper, but rather evaluation as the *understanding* of the project or program in question. To use a simple patient metaphor: For the '*social process'* school it is more important to understand what the symptoms express than to reduce the patient's pain. The followers of the '*social process'* school do not take value issues for granted and do not avoid discussing such issues. Ethical issues are often explicitly debated and they do not seek to avoid evaluative claims. Even though it is more than twenty years since the '*social process'* school was at its height and a lot of methodological development has taken place over the years, this school should probably be regarded as one of the directions that have contributed most to quantitative modeling within evaluation studies.

The '*constructivist*' or '*fourth generation*' approach has gained a lot of attention the last ten years (Guba et al., 1989) and has many followers, in particular in Europe. This point of view rejects evaluation as the search for quality, merit, worth, etc., in favor of the idea that it is negotiation which is the issue (Scriven, 1993). That is, negotiation between stakeholders with different interests is the essence of program evaluation. The core of the '*fourth generation*' approach is constructivism, a radical

position in contemporary epistemology and research methodology. The term 'constructivism' refers loosely to the idea that reality is not 'out there' but is constructed by each of us, and is rooted in the hermeneutic school (Bontekoe, 2000). The purpose of evaluation is to negotiate the conflicting interpretation of reality that is assumed to exist among stakeholders. The approach is relativistic in the sense that it is not a priori possible to define what apprehension of the evaluation question that is the most correct or objective one. Interpretations and understanding of problems and results may vary between stakeholders and there are no objective criteria that can aid in deciding what the best or the most correct one should be. Thus, negotiations are the most appropriate strategy for evaluations. – This is a radical view of the essence of evaluation that implies a complete re-conceptualization of the very task of evaluations and changes the very role of the evaluator from that of a researcher to the role of a judge. The approach does, however, build upon well recognized and established theoretical positions such as grounded theory (Glaser et al., 1967), interpretation theory e.g. (Eco, 1992) and hermeneutics, e.g. (Bernstein, 1983). I believe the approach is of great value for meta-evaluations, but to lesser extent suited for outcome oriented applied evaluations.

Cost-benefit analysis, or benefit-cost analysis; the more recent positive wording of the discipline, is the preferred method of evaluation for many governmental agencies. *Cost-benefit analysis* can be characterized both as a 'school' of evaluation and as an aspect of evaluation since the cost-effectiveness of the project/program in question is its major concern. Due to traditions, general trust in the discipline of economics, and the fact that the educational background of many bureaucrats is economics, many tender documents specifically require cost-benefit analysis to be inte-

gral part of the evaluation. That is, whatever kind of evaluation the researcher wants to carry out, whatever 'school' the evaluator subscribes to and whatever methodology supports the conclusions, the final report *has to* include an assessment of benefits and costs in the form of a cost-benefit analysis. Clearly, this requirement should be expected to have some impact upon the choice of evaluation methods. There are, however, surprisingly many evaluation reports that manage to carry through a cost-benefit analysis even though they do not include convincing attempts at assessing the impacts of the project/program in question.

In its general form, cost-benefit analysis has an appealing logical structure that resembles sound accounting principles. The benefits refer to the changes in the allocation of resources brought about by a project or program by comparing the situation before and after the installation of the project/program. Given some norm of calculating social welfare, the two situations can be compared. The common measure of the returns from a project/program is the net present value (*NPV*) of the project. Given that costs, benefits, and the social rate of discount are all accurately measured, *NVP* has a straightforward interpretation. When *NVP* is positive, the project/program under consideration provides a desirable yield to society, and hence, the evaluator can conclude that the decision to undertake the project/program indeed was a correct one.

The political use of cost-benefit analysis during the deregulation era in the early eighties during the Carter and Reagan administrations is an outstanding example of the use of evaluations to restrain regulatory agencies. Mandatory use of cost-benefit analyses "unless specifically prohibited by law" was used in order to dissolve even

health related regulations provided that *NVP* was below unity. The exemplar political use of cost-benefit analysis is the case of the cotton dust standard (Tolchin, 1987). Recent research has revealed that cost-benefit analysis is in such a troubled state that its usefulness in evaluations could be questioned (Frank, 2000; Kornhauser, 2000; Posner, 2000; Richardson, 2000; Sen, 2000). "The Statement of Principles on cost-benefit analysis" from the American Enterprise Institute (AEI) states that "Benefit-cost analysis should be required for all major regulatory decisions, but agency heads should not be bound by a strict benefit-cost test. Instead, they should be required to consider available benefit-cost analysis and to justify the reasons for their decisions in the event that the expected cost of a regulation far exceeds the expected benefits"[10].

Clearly, the reason why so many public agencies demand cost benefit analysis as an integral part of *any* evaluation report is, at the surface level, a need for clarification of the costs associated with demonstrated outcomes. For the policy maker it is of importance to know not only to what extent a given project or program is successful, but also if there is a reasonable proportionality between costs and the results achieved. Usually, this boils down to the simple question of whether the benefit cost ratio is the greater one or not. At a deeper, more subtle level, it may be questioned whether it is not the accountancy-like logic that constitutes a rhetorical beauty so perfectly suited for policy discourses that is the reason. It is disturbing that a technique, well proven to be seriously flawed and void of any scientific merit, still has its camp of devotees and is even regarded as a requirement for a complete evaluation.

[10] The statement is signed by Kenneth J. Arrow, Robert W. Hahn and Robert N. Stavins

The *causal-counterfactual approach* (Pearl, 2000) to evaluations is a rather recent invention. In essence, the inspiration for this approach can be found in several shortcomings of the established quasi-experimental approach (Cook et al., 1979). First and foremost, there has been a longstanding agreement among most scholars that the only appropriate way to talk about the effect of a public intervention is to speculate about the hypothetical situation that would have prevailed in the absence of the intervention. This line of reasoning introduces the term *counterfactual* and the interpretation and quantification of impact or effect of the intervention as the difference between the *factual* and the *counterfactual*. The development in *possible world semantics* (Lewis, 1973) provided the logic for counterfactual reasoning, and work in statistics along the same line of reasoning (Rubin, 1974, 1978, 1990) made the technical solutions available. The technical/statistical solutions also facilitated a way out of those problems that had laid earlier analysis of the quasi-experiments on shaky grounds, in experimental language, the non-random assignment of treatment, and the effects of cases (individuals, firms) being singled out for treatment for special reasons, the so-called selection effect.

Lewis' original (Lewis, 1973) formulation of the counterfactual theory of causation was spelled out under the assumption of determinism and had to be modified to allow for chancy causation. This work led to the more general notion of the causal counterfactual. *"Where **c** and **e** are distinct events, **e** causally depends on **c** if and only if, if **c** had not occurred, the chance of **e**'s occurring would have been much less than it actually was (given that **c** occurred)"* (Lewis, 1986). This more common sense notion of causality as related to events in the past, laid the ground for joining the logical structure of causation with the probabilistic notion of causation. In a

number of works, mostly in the field of medical statistics, Paul R. Rosenbaum and Donald B. Rubin (Rosenbaum & Rubin, 1985; Rosenbaum, 1995; Rosenbaum et al., 1983) developed their framework for the so-called *observational studies* where the properties of random assignments are emulated by means of matching procedures. Throughout the 1990s this line of reasoning was followed up, criticized and refined by a number of statisticians (Robins, 1989, 1997), a number of econometricians (Heckman, 1998; Heckman, Ichimura, & Todd, 1997; Heckman & Smith, 1995a; Heckman & Smith, 1995b) and sociologists (Winship & Morgan, 1999). The causal-counterfactual approach follows the language of experimentation, similar to the literature on quasi-experimentation. Contrary to the quasi-experimental tradition, the causal-counterfactual approach facilitates statistical analysis in practical settings, not merely discussions of potential and pitfalls.

Of the seven 'schools' listed above, the first one, the *'decision support'* school comes close to the common-sense view of evaluation. That is, evaluations of programs and projects are not that unrelated to other kinds of evaluations, say testing of products, or medical checks. There is no attempt to avoid evaluative claims; the most noteworthy with this point of view is that they see evaluations as an integrated part of the evaluating agency's decision problem. The *'relativistic'* approach avoids direct evaluative conclusions in favor of relativistic ones. The *'rich description'* approach avoids them in favor of non-evaluative descriptions. The *'social process'* school avoids evaluative conclusions in favor of insights and understanding of social phenomena, and the '*constructivist*' or 'fourth generation' avoids them along with all factual claims (Scriven, 1993). *Cost-benefit analysis* and the *causal counterfac-*

tual approach are somehow not as easily categorized as 'schools' as the other five. For public interventions, most bureaucrats consider cost-benefit analysis a must, regardless of whatever 'school' the evaluation task is performed under. This follows from the overarching understanding of the society as an *economy*, a perspective embedded in the NPM paradigm. The causal counterfactual approach is included as a 'school' for a slightly different reason. It is the only perspective listed that allows for the analysis of effects without assuming effects from the outset, and hence the 'school' that facilitates analysis based on the smallest number of untestable assumptions.

Table 5 sums up the basic features of the seven 'schools' I have briefly described above. A hypothetical seminar with, say, five devotees of each of the seven different 'camps' reveals the problem for evaluation research thought of as a discipline. Assume for the sake of simplicity that each participant was educated in such a way that he/she new only his/hers own perspective. It is highly unlikely that any kind of communication would take place. It is more likely that well-known psychological mechanisms of self-defense would be mobilized in accordance with what is known from the institutional perspective. Hence, it is not unreasonable to look for sources of bias in evaluation research, not only in the methodologies employed, but also in the human factor. The fact that evaluation research is conducted in settings that unavoidably includes not only conflicting interests and world-views, but also relations between fallible humans with limited rational capacities is a good reason for tightening the methodological grip, thereby leaving less space for accusations that 'anything goes'.

Table 5. 'Schools' and dominating traits of the agenda setting processes

	AGENDA SETTING THEMES				
'SCHOOLS'	**Evaluative Claims**	**Relation to Stakeholders**	**Relation to Evaluation Management**	**Basis for Validity claims**	**Research Strategy / Methodology**
The 'decision support' school	Yes	Biased towards contractor	Insider	Rational action	Predominantly quantitative
The 'relativistic' approach	Yes	Objective, but receptive towards the goals of the contractor	Outsider, but receptive towards the goals of the contractor	Design	Qualitative/ Quantitative
The 'rich description' approach	No	Independent	Outside, overlooking	Understanding	Qualitative
The 'social process' school	Yes	Independent	Independent	Design and measures	Predominantly quantitative
The 'Fourth generation'	No	Negotiator	Independent judge	Weighted views	Qualitative
Cost-benefit analysis	Yes	Independent	Independent	Neo-classical Economic Theory	Quantitative/ semi-empirical
The Causal-counterfactual approach	Yes	Independent	Independent	Design	Predominantly quantitative

4.4. Methodologies, ideas and basic beliefs

In accordance with many other scholars e.g., Morgan et al. (1980), I agree that the qualitative-quantitative divide is a misnomer that covers up a variety of underlying ideological[11], epistemological and ontological questions. When such questions are clarified, the link between methodologies, i.e., strategies of inquiry and method, i.e., techniques of investigations, becomes more transparent and explicit. The concept of 'ideology' as used here refers to ideas that stick in the mind regardless of the extent to which they have been proven erroneous. It does not necessarily have to be associated with a specific political ideology and it will be used interchangeably with the term 'idea'. Epistemology refers both to an underlying theory of knowledge and as used here, especially to beliefs about how knowledge or truth can be comprehended.

[11] By the term 'ideology' we refer to the pre-Marx concept of ideology that was first coined by Destutt de Tracy in 1796 to refer to the 'science of ideas'.

The term ontology refers to beliefs and theories of reality or existence, i.e., one's ontological beliefs determine how one thing about reality and about what exists in fact and what exists only in thought. I also introduce the terms 'pragmatic ontology' and 'ontological commitment' where the first concept implies an acceptance of phenomena and events that have causal powers regardless of the extent to which we are able to establish the underlying process that brought these phenomena or event into existence. The term 'ontological commitment' implies that ontologies are socially embedded, and thus, subject to commitment by the researchers who inhabit the different positions. The term 'pragmatic ontology' further implies that I do not follow the traditional doctrine of epiphenomenalism, that mental phenomena can have no causal effects. On the contrary, I accept all phenomenon and events that probes to have causal effects as valid explanatory factors. As institutional processes lies at the heart of the kind of phenomenon I believe to be related to outcomes in evaluation studies, and cognitive processes may entail mental constructs, variables based on emotional or other mental constructs are of importance. By the same token, arguments concerning 'ontological commitment' assume that commitment works via personal involvement and hence, can best be explained by cognitive theory based uncovering of the institutional processes involved. Also, I do not accept the notion of methodological individualism since I acknowledge concepts based on characteristics of institutional arrangements where constructs cannot be deconstructed in a way that makes it probable to trace the individual motives or decisions. That is, such concepts should be considered valid as long as they have explanatory power. Few of the core concepts introduced are new to the debate over methodologies and my interpretations are not particularly controversial. The point to be made is that the interpretations are not value free. The ideological stance of the researcher is usually

made clear in their outline of core concepts that clarify their basic beliefs about how reality can be reached and interpreted. This constitutes problems with the individual researcher's attitude towards research. Moral judgments set aside, it could be argued that opportunism is unavoidable, due to external pressures in terms of demands on earnings and loyalty towards one's colleagues earning expectations, when funding is received from related sources. We should hope that we never reach the point where consultants and researchers openly admit that they would have given an honest answer, could they only afford to do so. It is, however, fortunate that, in particular among young career-oriented researchers, a self-imposed naiveté eases the burdens of blatant opportunism.

4.5. The paradigm debate

The debate over qualitative versus quantitative inquiry in evaluations research has mainly evolved around an assumed strong tie between paradigms and methods. The notion of a paradigm is in most cases directly inspired by Kuhn's (Kuhn, 1962) concept, although discussions seem to reveal different interpretations of Kuhn's conceptualization. Mostly, the term paradigm seems to be taken as synonymous with "philosophical world view". This expansion of the concept unfortunately serves antagonism and fuels debates more effectively than the more modest idea of paradigm as "the psychological phenomenon related to believing that the description or explanation is correct, and the sociological phenomenon surrounding the coordinated enterprise of instrumentation, graduate education, textbook writing, reading, and 'problem-solving' according to the suggestions of the theory's agenda", called "normal science" by Kuhn (Poslby, 1998:202). A second misinterpretation of Kuhn is the assertion of a close connection in general between a single paradigm and an entire

academic discipline; an interpretation that departs substantially from Kuhn's intentions but is well suited for justifications of sharp delimitation between disciplines or sub-disciplines. Also, a small and rather insignificant sub-field may look more prominent when it is introduced as a new paradigm, alternative to a larger, established discipline. At any measure, it is hard to be confident that it is the question of qualitative versus qualitative inquiry that constitutes a paradigmatic divide. It seems more to be the case that the debate over paradigms is used as a rhetorical device for the sake or the argument. Adding descriptions of the epistemological and ontological beliefs presumed to be essential to the various "philosophical world view" a debate where polarization is a likely outcome is generated since a "straw-man" extreme and representative of historical well-known philosophical positions but characteristic of only few living researchers, is constructed. Many evaluation researchers characterize this debate as "unfortunate" (Worthen et al., 1997).

The debate over methods may also be seen as bearing the imprint of the history of evaluation. In the 1950s and 1960s evaluation was mainly something that concerned primary schools. The few people that were engaged in evaluation research were schoolteachers. With increased attention to the problems of evaluating the effects of new reform programs for schools and new pedagogical techniques for making education more effective, the psychologist, trained in the experimental tradition entered the scene. The work of Campbell and Stanley (1969) gave considerable impetus and inspiration to evaluation research and made experimental and quasi-experimental the predominant approach. At that time, in the middle of the sixties, there was a recognized criticism of the experimental tradition within the psychological profession. This disagreement was mainly due to the tension between the established Freudian

psychoanalytic tradition and experimental psychology. In the education community, reactions toward the new evaluation techniques was also probably inspired by the pedagogy of scholars like Paolo Freire and others whose main concern was very different from those of the proponents of experimentation. Hence, due to its roots in the efforts to reform in the primary schools system, it is likely that the conflict between qualitative and quantitative methodologies has a history that reflects the contrast between managerial and pedagogic thinking in the Great Society era of the late sixties and early seventies. Hence, the conflict shows path dependence, a heritage long forgotten by today's actors in the debate.

4.6. Sub-discipline loyality

It is generally accepted that the term *social sciences* covers several basic disciplines, such as economics, sociology, anthropology, organization studies, and so forth. Within or between these basic disciplines a variety of sub-fields or sub-disciplines have emerged over the years. Each of these up-and-coming sub-disciplines can be viewed as having an inner drive towards maturity, although some of them may turn out to be short-lived fads. The development towards maturity implies a number of institutional processes, such as organizing conferences and the establishment of new journals in the field. The ambitions of the sub-fields may vary from the more modest to the full-blown ambitions of professionalism with formal educational requirements and well-defined boundaries for membership. In Niklas Luhmann's (1995) terms this development implies a kind of autopoiesis[12] where the growth of an academic sub-field may partly depend upon the new field's ability to draw the demarcation lines which distinguish members from non-members. The notion of being a member of a

[12] Autopoiesis refers to a process whereby an organization produces itself.

distinguished academic field seems to be of importance for academic identity (Henkel, 2005). Academic identity determines the character of the language the scholar is expected to use in written and oral presentation, i.e. what terms are anticipated to be excessively repeated and emphasized. Moreover, academic identity offers a relief from the burden of an otherwise perceived constant demand for acquiring new knowledge. Within what can loosely be termed "the evaluation community" patterns of sub-divisional grouping is very visible. It is visible in the conference programs of both the European Evaluation Society (EES) and the American Evaluation Association (AEA). Terms like "evidence-based evaluations" (e.g.,Sanderson, 2002) would in some meetings create reactions equivalent to the division of the Red Sea.

The debate over qualitative versus quantitative methods appears to be cemented by a sharp assignment of priority for either the first or the latter-mentioned kind of methodology. Thus, as evaluation becomes increasingly divided into distinct sub-fields, questions concerning methods become apparent. The world-views inherent in the sub-disciplines usually reflect those of a parent discipline. The methodological peculiarity of the parent discipline is in usual carried over to its respective sub-disciplines. As new theoretical projects, sub-disciplines tend to have a self-protective drive towards maturity, manifested by the development of core concepts and research methodologies. As such, some of the new sub-disciplines may have traits that are not well suited for investigations where the quest for causal conclusions is unavoidably strong. Attempts to comply with the demand for causal claims such as Lawrence B. Mohr's (1999) re-introduction of *causation without variance* correctly point out causation does not require statistical models. Also, the APA standard re-

quest for effect-size reporting has produced efforts to include concepts such as statistical power, effect-size and validity in qualitative studies (Onwuegbuzie, 2003; Onwuegbuzie & Leech, 2007a, b). Thus, naturalistic inquiries seem to be subject to external pressure, e.g. from academic journals, for including types of reasoning usually associated with quantitative methodologies. The attempts to conform to these demands indicate that there is still a considerable amount of work to be done (Onwuegbuzie et al., 2007a).

5. Expectations, confirmation bias, and self-delusions

5.1. Sources of bias – cognitive perspectives

Human perceptions are fallible and the direction of bias is usually in support of our preconceived beliefs, i.e., confirmation bias. Confronted with our erroneous predictions the common reactions to the alternative better predictions may be resistance, rationalization of one's own prediction and to cast doubt upon the foundation of the undesirable prediction. The extent of the literature on confirmation bias is considerable (e.g. Baron, 1981; Benshakhar, 1995; Davis, 1994; Gadenne, 1982, 1984; Jonas, Schulz-Hardt, Frey, & Thelen, 2001; MacCoun, 1998; Mynatt, Doherty, & Tweney, 1977; Nickerson, 1992) and is relevant for evaluation research as well as other research. For our discussion here, however, we will only scratch the surface.

We can coarsely distinguish between two separate potentially bias producing processes within the evaluation setting: *a)* those typical of the research process itself and, *b)* those communication processes that cause clearly biased results to become accepted by governmental agencies. Clearly, these processes are related, but for the purpose here, it is convenient to split the discussion into two parts. The *first part* concerns the relationship between the researcher and his/her interpretation of data. This involves not only the peculiarities of the individual researchers but also traits that are present in the research community that embeds the researcher. The *second part* concerns the stakeholders in the evaluation process itself. The most prominent of these is the evaluation management and other relevant stakeholders, the policy-makers and other parts of the government agencies involved in or affected by the outcomes of the evaluation.

Researcher will most likely approach the relevant stakeholders affected by the project/program in question with some pre-determinate ideas or hypotheses about what reactions or expectations are central to the evaluation. The quantitative oriented researcher may express his/her ideas in the form of written questions in a questionnaire or as instruction in an interview guide. The qualitative oriented researcher is, however, less assertive of own capabilities to settle in advance what the right questions might be, and want to figure this out by interacting with the people involved, either via direct conversation or by other means that allow for the most direct communication. Interaction and direct involvement with those affected by the public project/program in question, seem to be salient features that distinguish qualitative inquiry from quantitative inquiry. Clearly, the personal features of the qualitative inquiry procedures are in line with cognitive psychology reasoning insofar that it recognizes limited cognitive capacity. The qualitative methods literature is, however, more reluctant to incorporate newer research on expectation confirmation in interpersonal relations. Several theories concerning expectation confirmation may apply. The hot cognition hypothesis (Abelson, 1963) posits that all social concepts that have been frequently used in the past become “affectively charged” positively or negatively tagged, with the affective charge linked directly to the concept in memory (Chen & Bargh, 1997). The most well-known type of expectation confirmation is the self-fulfilling prophecy where our expectations about others, whatever their origins, tends to elicit the very behavior that is expected. Coincidentally, this line of research has much in common with evaluation research, as they both originate from studies of the primary school system. One of the first hints of expectancy confirmation was the demonstrations that teachers who were led to expect particular levels of performance from students in their classroom, acted in ways that elicited perform-

ances that confirmed the initial expectations (Rosenthal & Jacobson, 1968). More recent research shows that the processes of expectancy confirmation and disconfirmation involves a complex intertwining of cognitive, motivational, and behavioral activities in social interaction (Snyder & Stukas, 1999) but also that the phenomena follows traceable paths that make it possible to map the mechanisms at work. Decomposition of the elements of the mechanisms into a series of steps suggests the following sequence: *(a)* perceivers adopt beliefs about targets; *(b)* perceivers behave toward targets as if these beliefs were true; *(c)* targets fit their behavior to perceivers' overtures; and *(d)* perceivers interpret targets' behavior as confirming their beliefs (Kelly, 1992). Expectations can be categorized according to their *properties*, such as certainty, accessibility, explicitness, and importance. Increases in one of these properties should lead perceivers to increase their tendency to act on expectations in ways that increase the likelihood of confirmation (Olson, Roese, & P., 1996). The studies of expectancy confirmation complement the more philosophy of science oriented debates over the problem of "theory-laden" observations (Estany, 2001; Franklin et al., 1989; Holman, 1979; Moore & Barresi, 1993) i.e., the influence of theory on what see (Scaff & Ingram, 1987).

5.2. Bias due to scholarly training

In the Anglo-Saxon world researches are generally trained in a tradition where theory guides observations and tests, that is; observations are, by definition, decided upon by theory and thus, researchers may see what they are looking for. This view of the relation between theory and data is different in other spheres of the world. The favored method of e.g. Pierre Bourdieu (1984; 1993; 1998; 1999; 1991b) is correspondence analysis (Benzecri, 1969; Greenacre, 1993; Hill, 1974); a method where

the iteration between data an theory is explicitly stated as an integral part of the research process. Thus, the absolute priority of theory is not absolute. Moreover, in practical research, the iterative switching between data and theory is usually the case despite the fact that this practice may violate the logic of statistical tests. It could be maintained that strict commitment to a chosen theory which induces a temptation to have a selective attitude towards the data is more of a problem for the Anglo-Saxon researcher than theory-laden observations. Rigorous theories tend to rebut rigorous empirical tests.

The file-drawer problem (e.g.,Rosenthal, 1979; Rotton, Foos, Vanmeek, & Levitt, 1995; Shadish, Doherty, & Montgomery, 1989) is well known to researchers. In its most extreme version it insinuates that "journals are filled with studies that show Type I error while the file-drawers are filled with the 95% that show non-significant results" (Rosenthal, 1979:638). Researchers are trained to expect that only studies that confirm results will become published. The tendency to prefer studies that produce significant results *may* carry over to the evaluation setting, regardless of qualitative or quantitative methods orientation.

Moreover, it could be argued that the qualitative researcher, as an interpreter of "data" takes the role of an *expert*. As pointed out by Robyn Dawes (1979), simple linear models may in many cases (Kuhn, 1962) outcompete the clinical intuition of presumably skilled expertise.

5.3. The Measurement Question – a source of confusion

Summative evaluations assume outcome assessments, i.e., one or other kind of measurement of what is achieved by the project or program in question have to be

acknowledged. If we for simplicity imagine three different scholars, one with his/her basic methodological training in psychometrics, one in econometrics and one trained in qualitative methods only, their conception of what to measure would surely differ. The psychometric inclined researcher would most likely have a theoretical concept as his point of departure and spend considerable time on explaining the measurement procedure and to what extent he/she succeeded[13] in measuring what was intended. The econometrician would most likely look for things like number of new jobs created, the meaning of significant events such as bankruptcies or mergers or one other predefined measure in money terms. Both the psychometrically inclined and the econometrically inclined would look for measures that are embedded within a field-specific theoretical framework, but while the psychometrically inclined would emphasize measurement, the econometrically inclined would look for some magnitude "already out there" and emphasize the data-generating process, i.e. the process that left the observable tracks[14]. The qualitative methods inclined researcher would most likely also take theory[15] as his/her point of departure and, much in line with the psychometrician he/she would emphasize measurement. However, while the traditional psychometrician would most likely prefer predefined items, preferably tested by other researchers in several other studies, the qualitative method researcher would prefer to rely on his/her personal interpretation. Thus, while the psychometricians rely on a worldview that research is cumulative or at least self-correcting in the sense that he/she has to believe that the many efforts to establish an item-bank that over time makes constructs converge towards the semantic domain of the theoretical concept, the qualitative researcher has to rely more on

[13] They usually succeed, otherwise it is not reported.
[14] The business student, trained in a hybrid of the psychometric and the econometric tradition would be utterly confused and invariably report Cronbach's coefficient alpha.
[15] It is not always clear what is meant by theory; here we at a minimum think of something like a model, analytical framework or a conceptual scheme.

his/her personal intuition. Moreover, because qualitative methodology is not a single method but a bundle of many different and to some extent incomparable heuristics for making sense of the empirical world as perceived by her/him personally, the urge to discover something new as opposed to the drive towards seeing something recognized by others, involves a kind of self-exposure which is not experienced by the psychometrician. Clearly, new surprising discoveries may be more risky than well recognized findings and any deviation from what is expected e.g., by the evaluation, management might trigger a quest for validity. While validity assessment is routinized in the psychometrician's research process, it is less well developed within a qualitative inquiry framework (Onwuegbuzie et al., 2007b). Moreover, the evaluation management and other stakeholders are quite likely to take validity assessment as a "quality-clue" and thus, in a conceivable battle over conflicting interests among stakeholders, the psychometrician may score in terms of legitimacy, despite the qualitative methods, the researcher's inclusion of the term "qualitative", which is easily and even unconsciously associated with "quality", is the abstract aspect of research that signals trustworthiness to the layman. Both the psychometrician and qualitative researcher are primarily equipped with methodological apparatuses unambiguously oriented towards an understanding of the empirical world as interpreted by human experience i.e., they use instruments aimed at eliciting human perceptions. Thus, the double nature of their common endeavor is to interpret interpretations; the psychometrician primarily by means of statistical models that extend what is already known and the qualitative researcher by means of systematic inquiries, on the hunt for new interpretations.

Compared to the psychometrician and the qualitative researcher, the econometrician tends to come across as naïve and ignorant, in particular in the view of the part of the new generation of qualitative inquiry scholars that from the outset have chosen to renounce so-called "positivist methods" altogether. Heavy reliance upon the behavioral assumption of the "economic man" may be read as unawareness concerning the human condition for interpreting the outer world. Psychometricians seem to have forgotten their debt to econometricians, in particular the works of Goldberger in the early seventies (e.g., Goldberger, 1972; Joreskog & Goldberger, 1975). The principal disagreement between psychometricians and econometricians over the measurement question is of a different nature than the disagreement between qualitative methods researchers and the econometricians; while psychometricians consider econometricians to go too light on the measurement question and rely too heavily on behavioral assumptions, the qualitative method researcher appears to understand econometricians as representatives of a kind of contemporary "Vienna Circle"; more or less today's logical positivists. This angle on discussions does, by its sweeping through the older philosophical debates over questions of human access to the empirical world, in fact produce interesting reflections over the development of social science methodology; unfortunately, it also prevents recent developments within econometrics from being communicated to the evaluation community at large.

The most undercommunicated aspect of summative evaluation is the consequence of the inherent counterfactual nature of the established jargon, at least within the evaluation of EU-framework programs (Luukkonen, 2000) where *additionality* is a central concept. Additionality means that observed outcomes would not have been

present without the program in question and, if positive additionality is said to exist, the cost-benefit ratio has to exceed unity. Since additionality is established in the terminology as *the* criteria for success of a chosen policy, the discourse following the presentation of an evaluation report is structured as if observed outcomes are *caused* by the program/project that is evaluated. This is likely to happen regardless of whether a method which justifies this kind of reasoning is applied or not.

The idea that qualitative inquiry makes the evaluation task easier is not alien to researchers. This is not a very productive idea. To use a phrase from Thomas A. Schwandt, a prominent scholar of qualitative inquiry, such a position is doubly tragic:

> "Sadly, some researchers seem drawn to qualitative inquiry for the simple fact that they do not wish to "deal with numbers." This is doubly tragic. First, it is based on faulty reasoning— there is nothing inherent in the epistemologies of qualitative inquiry that prohibits the use of numbers as data. Second, such a stance can be based in the illusion that so-called qualitative inquiry is somehow "easier" to do than so-called quantitative inquiry. But it is hard to imagine what criteria might be employed to determine that the level of effort and thought required for writing field notes, conducting and transcribing interviews, interpreting different kinds of qualitative data, and so on is somehow lower (or higher, for that matter) than that required for designing and executing a careful and meaningful test of a statistical hypothesis. These inquiry tasks simply require different kinds of awareness, knowledge, and skills." (Schwandt, 2000:206)

The econometrician's presumed naivety with respect to the measurement question is frequently emphasized as an important difference between naturalistic inquiries and statistically oriented quantitative methodologies. The 'brute fact' argument and the 'theory-laden' observations argument is occasionally introduced as a hint to the consequences of ignoring the inevitable effect of theory, namely that you see what you

are looking for. Thus, as indicated by the quotation, below, naïve belief in the possibility of objectivity could lead to worthless measurements.

> "Since the classic studies of Hanson (1958) and Kuhn (1962), it has become commonplace to note that there is no such thing as a brute 'fact' or a completely neutral 'observation language' simply 'there' in the world, existing independently of all conceptualization and serving as definitive test of our theoretical generalizations. The 'facts' composing the world of our observations are always 'theory-laden' (partly quoted from Hanson, 1958:19)." (Scaff et al., 1987:235)

The problem with this line of reasoning is that it is hard to find a quantitative oriented researcher who would completely disagree. Some would find the tendency to absolutism in the quotation a bit out of place, in particular since the quotation demonstrates that the 'brute fact' argument is not well understood[16]. The naturalists' argument that "there exists multiple realities which are, in the main, constructions existing in the minds of people; they are therefore intangible and can be studied only in holistic, and idiosyncratic, fashion" (Guba & Lincoln, 1988:81-82) does not rule out the possibility for measurements that many people may comprehend similarly. In summative evaluations it is of grave importance that central measurements can be communicated and are easily understood in the same manner by stakeholders and the wider audience.

In 1966 Donald D. Campbell and his associates (Webb, Campbell, Schwartz, & Sechest, 1966) made an important contribution, apparently long forgotten by the evaluation research community, by introducing the somewhat strange term '*unobtrusive measures*'. Webb et al. (Webb et al., 1966) defined *reactivity* as *obtrusiveness*, thus making the search for unobtrusive measures the search for methods that *do not*

[16] The term "brute facts" was first coined by Gertrude Elizabeth Margaret Anscombe (Anscombe, G. E. M. 1958. On Brute facts. ***Analysis***, 18: 69-72.) and was meant as a term for that which has no explanation outside the institutional framework in which it is embedded such as e.g., a hundred dollar bill, which clearly, without an institutionalized money system is just a piece of paper. Thus, brute facts are not suited for *explanations* (Fahrbach, L. 2005. Understanding brute facts. ***Synthese***, 145(3): 449-466.

affect or distort the data that are collected. Both quantitative and qualitative methodologies presuppose the participation of individuals as research objects. Thus, regardless of the data collection methods being interviews, surveys, rating scales or open questions in questionnaire based studies; people have to volunteer to participate in the endeavor of data collection. For this process to succeed there must be consent among the participants that it is worthwhile to offer the time and effort necessary to help the researcher to gather the data that is needed for the study. In the psychology literature this spurs a variety of questions, some of an ethical nature e.g. (Atwell, 1982), others of methodological character e.g. (Rosenthal, 1976). It is undoubtedly unethical to force people to participate in a study and it is questionable to engage people in studies where they disagree with the purpose of the study or would not have participated had they known the purpose of the study. Clearly, attitudes toward a study may also affect how people choose to relate and react to various questions. Thus, the way a given study relates to its study objects is a methodological problem and a potential source of bias such as e.g. strategic answering due to positive/negative interests or indifference (careless respondents) towards the outcomes of the study (e.g.,Schmitt & Stults, 1985).

Administrative records are unobtrusive in the sense that there is no reason to expect reactions towards the collection of administrative data in general, in particular when administrative archives would exist independent of the project/program under scrutiny. Objection to the use of administrative records by participants in the project/program that is evaluated is a different issue.

A property related to unobtrusiveness is stability. By stability we here think of stability over time, location and groups. Comparisons of outcomes over time, between various locations and groups require that we measure the same phenomena every time. Since summative evaluations aim at detecting differences between identifiable participants and non-participants, a degree of neutrality of measures is necessary for meaningful comparisons. Since stability across subjective interpretations may be hard to achieve, even for the experienced researcher, naturalism may be the best choice for measurements in summative evaluations. Thus, traditional scientific methodology may be the preferred choice when stability across repetitions is of importance.

6. Presumed consequences – a summing up

This section sums up some probable answers to the three theses proposed and other peculiarities of the circumstances that confront evaluation researchers when the task at hand is applied summative evaluations and the demand for causal claims is present albeit not always explicitly contractually stated.

6.1.The three theses

6.1.1. Consequences of the substandard argument

Concerning the substandard argument I have suggested that "*accepting substandard evaluations leaves the debate over evaluation methodology obsolete and makes it hard to establish the status of evaluations*".

As previously stated, I think the substandard argument should be rejected. The sensible line of reasoning put forward by Coleman (1972) four decades ago was based on the contemporary development of evaluation methodology. Since then, evaluation methodology has made substantially progress and it is possible to present trustworthy analyses of presumed outcomes of programs and projects; analyses that are trustworthy in the sense that they fulfill the transparency requirement and can, within reasonable limits of probability, establish whether intended program goals are attained or not. Methodological progress is not, however, our main reason for rejecting the substandard argument. It is more the expected *consequences* of the acceptance of lower methodological standards in an increasingly routinized use of evaluations. The last two decades have seen a substantially growth in use of evaluations. In the US, the introduction of the Government Performance and Results Act of 1993, which made evaluations mandatory for programs and projects with expenditures

above a specified sum, had a major impact. Moreover, evaluations are now both an integrated part of NPM and an important ingredient in the media-strategy of the agency responsible for the program/project in question. The extensive use of evaluations in combination with the impression that "anything goes" with respect to methodology (Adelman, 1996) is not a good scenario. The impression that the fact that an evaluation is carried out is sufficient is disturbing. The remaining part of the process of legitimating the program/process is handled by the spin doctors and the media strategy. Such a development is unfortunate, both from a reputation of the social sciences point of view and a professional evaluator's point of view. Within the individualizing logic of NPM, this development is understandable; the insistency that an identifiable *person* should carry the responsibility for the result of public programs/projects is to some extent a rather recent invention, born from media convenience. Regardless of whether this person was responsible for the initiative to the project/program in question, to what extent he/she was in charge of the practical implementation or not, the need for attributing success of failure to a person that can be interviewed, asked for comments an so on, is crucial for the construction of media coverage. Thus, summative evaluation may take the character of a verdict, both with respect to the program/project under scrutiny *and* the person assumed responsible. Evidence and verdict are connected, both in legal terms and in the folk psychology that produces reactions from the media audience. Verdicts based on flawed evidence are undesirable and so are declarations of successes which appear to be more marketing than research-based documentation.

Evidence and methodology is an inseparable duality. Moreover, the transparency of the methods applied is critical for external judgments of the quality of the evidence

presented. Thus, the growth and extensive use of evaluations that has emerged over the two recent decades implies that the idea that we should accept lower standards for evaluations than for discipline research is not a good idea for summative evaluations.

6.1.2. Qualitative methods for summative evaluations: The amalgamation problem

My proposal concerning the sole reliance upon qualitative methods for summative evaluations states that: "*Qualitative inquiry may produce confirmation bias in evaluation studies by amalgamating the research and the researcher, and hence confusing psychological self-defensive mechanisms and professional argumentation. Thus, conflict may be experienced as insult and support for argument as praise, constituting a mechanism that increases the probability of agreement between the researcher and the dominant stakeholders.*"

Clearly, the research/researcher amalgamation problem is a difficulty of concern not only for the qualitative researcher. Generally, the research process implies considerable personal involvement. The researcher who feels that the research task is executed in a proper way is usually prepared to defend his/her work. The critical difference between qualitative and quantitative inquiry is rooted in the subjective, interpretative aspects unique to qualitative inquiries. While most quantitative inquiries operate at the molar level of observations and restrict evaluative causal claims to this level, qualitative inquiries may go beyond this level and try to understand the underlying mechanisms beyond the observable. It is not all that clear whether the qualitative analysts accept the notion that causal assertions are meaningful at the molar level even when the ultimate micromediation is not known (Cook et al., 1979). Most facets of qualitative inquiries do, however, define their activities as characterized by

the difference between what they do and what quantitative researches do. Thus, quantitative inquiries serve as a reference base. While molar causation is central to quantitative analyses e.g. analyses based upon measures against the counterfactual, it has no place in qualitative inquiries. Whereas quantitative analyses emphasize models, rules and formal procedures, qualitative analyses accentuate the narrative. Thus, a report based upon quantitative analyses may appear as sterile, demanding reading while the report from the qualitative analyst tells a story that may even be interesting to read. Moreover, the quantitative report may contain equations and statistical expressions unfamiliar to the reader. In sum, the qualitative report may be more communicative than the quantitative report.

However, in summative evaluations, the first question is: Did the project/program in question turn out according to intention? And second: Did it produce the intended results at a cost that makes it beneficial to society? In my opinion, a worrisome number of summative evaluations provide a positive answer to both questions. The sheer fraction of yes's compared to no's indicates confirmation bias (Jones et al., 2001; Nickerson, 1992; Pollard, 1983) of the kind usually termed *researcher bias* (Bamberger, Rugh, Church, & Fort, 2004). To my knowledge no study of confirmation bias exists which involves a qualitative/quantitative distinction. The point to be made here is that summative evaluations force a decision concerning the success or failure of a project/program and also force the conclusion that these outcomes can be causally attributed to the project/program in question. Any disagreement concerning this conclusion asks for a thorough investigation of the evidence behind this decision. Hence, under fierce conflict between stakeholders, methods which are transparent in the sense that every step in the research process is traceable, is to the

researcher's advantage. The researcher can find shelter under the protective umbrella of objectivity only as far as he/she can demonstrate, step by step, exactly how the conclusions were produced. The appeal to objectivity is probably the best possible escape route for the researcher in a situation of fierce conflict. The only alternative, the appeal to authority, would probably, according to Aaron et al., (1990) be of less use. Appeal to objectivity, however problematic it may be, is a less viable possibility for the qualitative researcher. Accusations of subjectivity are hard to beat off when personal interpretation and understanding is an integral part of the method applied. Thus, since conflict is a frequently occurring situation in summative evaluation, methods that permit for a kind of detachment, i.e., that make it possible to create at least some distance between the researcher and his/her work, in short, traditional scientific methods, may be preferable.

Lately, the study of evaluations per se has gained attention from scholars, and a number of academics seem to endorse E. R. House's (1999) conclusion that[17]:

"Society expects for evaluation to be based on scientific authority . . . the more objective and less ideological evaluation becomes, the more useful it is and the more it threatens established authority. A useful practice would provide sound evaluations that are not ideological. . . The evaluation theory developed so far is too ideological." (House et al., 1999:30)

6.1.3. Schools and subdisciplines: Consequences of scholarly contradicting world views

The many subdisciplines of evaluation offer a smorgasbord of alternatives where the nature of the evaluation report can be picked according to taste. As shown in Table 5, the different 'schools' vary with respect to their relationship to management, their relationship to stakeholders, and their willingness to make evaluative claims. For

[17] Quotation from Henry, G. T. 2001. How modern democracies are shaping evaluation and the emerging challenges for evaluation. ***American Journal of Evaluation***, 22(3): 419-429.

formative evaluations, the smorgasbord may be useful. Pluralism is positive in the sense that is prevents the dominance of a single perspective, thus some aspect of the program/project under scrutiny that would otherwise be neglected could be brought to attention. On the other hand, for summative evaluations, pluralism in perspectives may be negative if it opens up for predictable outcomes, i.e., evaluation management is supplied with a mean for choosing the results that suit their preferred agenda. Provided that each perspective is assigned near equal legitimacy and the peculiarities of each different 'school' is known by the evaluation management, opportunism may win through.

6.2. The irrereversibility of customs and habits

Institutionalization processes are runaway trains that come into action simply as time passes and activities are repeated. The logic of NPM is based upon mimetic processes (Dimaggio & Powell, 1983; Galaskiewicz & Wasserman, 1989) where public sector programs and projects mimic what is perceived as parallel tasks in the private sector. The evaluation part, which is not correspondingly common in the private sector, is included as justified by the demand for accountability and as an act or responsibility due to the fact that actions are undertaken on behalf of the public interest. The basic view of the private enterprise as driven by the *logic of consequentiality*, i.e., rational, informed self-interest is central to NPM. Thus, when the rules of rational action are transformed to the public sector, productivity is enhanced via incentives since "man's natural proclivity is to pursue his own interests" (Brennan & Buchanan, 1985:ix).

However, in the public sector the demand for legitimacy may oust the demand for accountability and the actions required for establishing legitimacy may differ from the procedures that facilitate accountability. Thus, as public agencies copy the practices of other public agencies and evaluations become commonly routinized due to media-driven language games and the 'semantic magnet' properties of the term 'evaluation' (Vedung, 2000), institutional processes cement evaluations as part of the public agency's modus operandi. As part of the going concern of a public agency, that status of evaluation changes and becomes sustained by the *logic of appropriateness* (March & Olsen, 2004). Evaluation is now commonplace and acceptance of the new routine becomes part of the rules that guide appropriate behavior. Thus, the identity and role as a bureaucrat now encapsulates the acceptance of evaluations and objections to the new routine or to the quality of reports may be taken as unwillingness to comply with agency culture and the community of colleagues.

It is not likely that the number of evaluation reports produced in the near future will drop. Attempts to reduce the number of summative evaluations at a stage of program development so early that no results should be expected to be observable, would most probably raise objections from many public agencies. Thus, even though the immoderate optimism and faith in liberal economics that gave birth to NPM and the Reinvention Government movement in the early eighties has lost its momentum, evaluations as an essential part of public programs and projects prevail. Thus, since a growth in the production of evaluations is predictable, quality consideration should be essential both to the evaluation research community and the public at large.

7. Evaluation commissioning

7.1. Competitive tendering

The general rule for the contracts for public evaluations is open competitive tendering. Within the European Union and the EEA countries announcement of tenders are subject to strict rules to make sure that contracts above a minimum value are subject to open competition between relevant research institutions and consulting companies. There are few restrictions with respect to formal qualification for applicants and no prerequisites in forms of certificates or other statements of eligibility for contractors. The tenderer is free to contract with any offeree found suitable for the purpose at hand. There are no restrictions in terms of required procedural arrangements for the bidding process, e.g. independent assessment of what should be considered the best offer or rules concerning repeated use of contractors. In most cases, contractual decisions are handled by the office or agency responsible for the projects or programs under examination. Contract prices are normally fixed or nearly fixed and the evaluation management is often in charge of both developing the tender documents and making the decision about whom to hire for the evaluation. Thus, contractual agreements are usually based on assessment of the perceived quality of the bidder's proposal about how to solve the task at hand. Thus, the reputation of the offeree and expectations about the nature of the final product play central roles for the formation of contracts. Reputations and expectations may primarily appear as recognition of the quality of research and the ability to carry out eminent evaluations, but it may also include the rumor that dominant stakeholders usually get the answer they want. If experience indicates that re-contracting depends more upon sufficient support for the interests of central stakeholders than on the quality of the

research that carries the conclusions of summative evaluations, we may have a situation where open competitive tendering systematically deflates the value of evaluations. Institutionalization of summative evaluation as the end of line unit within the logic of NPM may, under unfavorable circumstances where research institutions depend heavily upon funding from evaluation contracts, encourage a kind of conformity that renders summative evaluations uninformative.

7.2. Specification of the evaluation task

The basic question in any evaluation is to provide trustworthy evidence about the merit or worth of the program or project in question. The reasons for using phrases like *evaluation research* and the reasons for engaging a *research institution* to do the job are rooted in the need for trustworthy conclusions. Public agencies invoke the legitimacy of science by asking for answers based upon *scientific methods*.

A question is an expression of intellectual anxiety and an answer an attempt at resolution of that anxiety (Myhill, 1951). From a logical point of view, we can distinguish between two kinds of questions, *formal* and *informal* questions. A formal question carries with it the form of its answer, i.e., the social context is such that the criterion of the acceptability of the answer is known and agreed upon by both questioner and answerer in abstraction from the answer itself (Myhill, 1951). The purest kind of formal question is the question of the truth or falsity of a mathematical theorem within a known system. The criteria for being a proof within the system are exactly specified and agreed upon by both questioner and answerer. In empirical research formal questions are a less pure kind of formal questions. Criteria of confir-

mation are less specifiable than for the mathematical proof and hence, not so easily agreed upon by the research community.

An informal question is one where the form of the answer is not known either by the answerer or the questioner in abstraction from the answer itself. The dictum that the meaning of a proposition is the method of its verification, does not apply to propositions which answer informal questions, since part of the meaning of such questions is to question what the form of its answer should be (Myhill, 1951:58). Hence, part of the meaning of the question "What are the merits of this particular project" is "What form of answer is best suited to resolve the anxiety expressed by this question"? That is, to the extent a question is formal, to an equivalent extent the questioner will be prepared to state precisely the kind of evidence it would take to convince him/her of the of the truth of any proposed answer. Hence, a formal question asks for the matter of its answer but provides a specification of the form of the answer while an informal answer asks for both. The ambiguity of the informal question opens the opportunity for the researcher as contractor to advocate his/her preferred form of answer and the tenderer receives a menu of possible solutions where it is fairly easy to predict reactions from various audiences such as the media, the policy makers or various research communities. Thus, an evaluation concerning, say, toxic waste does not require much knowledge of chemistry and an evaluation of a vaccination program may not specify the need for epidemiological expertise. Indeed, the paradox of the demand for qualitative methods in the era of NPM is not a paradox but more of a confirmation of the dictum that management is better left to the managers.

7.3.The integrity of research

The legal language of most evaluation contracts is usually aimed at protecting the interests of the evaluation management. Thus, the evaluation management refrains from specifying the form of the answer to the evaluation problem, have the right to accept any methodological solution suggested and have legal monopoly on the use of the reports produced. In sum, conducting evaluation makes weighty demands upon the integrity of research. In an ideal world, researchers should, by upbringing or education, have internalized a commitment to present only what they at a minimum themselves believe to be valid knowledge. In philosophy of science jargon[18] this should imply that:

- The researcher should not acknowledge any higher authority than the authority of the best argument.
- The researcher should not accept any other directional force, whether it be social or human based, than the institution of the "argumentative game" itself. That is, the logic of the argument should not be contaminated by considerations regarding social roles, status or position. Also, there should be unrestrained freedom of speech and symmetry between all parts involved in the argumentative dialog.
- The free, argument-driven fellowship and what it produces should serve as measures for actions and institutional arrangements and also as a critical test of what should be considered as valid, legitimate and rationally acceptable.
- The top priority should be the awareness of self-deception[19], the most common plague for the social scientist.

[18] Adapted from Habermas, J. 1990. Discourse Ethics: Notes on a Program of Philosophical Justification. In S. Benhabib, & F. R. Dallmayr (Eds.), ***The Communicative ethics controversy***: viii, 378 s. Cambridge, Mass.: MIT Press., page 118.
[19] This argument is from Wittgenstein, L. 1953. ***Philosophical Investigations***. New York: MacMillan..

Many researchers would discard such pronouncement as naïve or even exorbitant. Some may admit that the four points above are actually a proper description of their own, basic thoughts about doing research, although that they found them idealistically naïve. Others would say that the four points have a blatant flavor of objectivism and that they personally were working in a field where the prevailing paradigm barred against beliefs in the possibility of objectivism. Many researchers would acknowledge that publishing in leading journals of their field gave much of the raison d'être for their scholarly efforts, not their inner feelings of intellectual honesty.

8. Concluding remarks

I clearly recognize a place for qualitative inquiries in evaluation studies but also that the multiple realities supposition and the emphasis on subjective interpretation may produce unwarranted ambiguities in summative evaluations. Many problems perceived to exist between proponents of qualitative and quantitative research may stem from misunderstandings. Even well conducted, backward traceable and replicable quantitative studies have a difference between *reconstructed logic* and *logic in use* (Kaplan, 1964). The logic in use during the research may take many forms, may go back and forth in iterative steps, may involve interpretations and may not look very different from what the qualitative researcher is doing, with the exception that it usually involves numbers. The reconstructed logic, which succeeds the actual research process, is more stringent since it orders the sequence of the work and fills in what is actually done at each step of the quantitative research schemata. Thus, the research process may invariably have the two facets Reichenbach (1938) coined *context of discovery* and *context of justification*. It is worth noting that it is the reconstructed logic that makes analyses replicable and backward traceable. In short, in summative evaluations a few elements should be recognized, such as:

- acknowledgement of the reliance on the senses in the research process, and, a
- mandatory a priory statement of methodological principles, an
- awareness to the replicability principle, combined with
- sensibility concerning the communicability of results, and finally,
- it should be possible to falsify any hypothesis introduced.

Attention towards agreed upon scientific standards does not remove the dilemmas that frequently meets the evaluator. Time constraints and a lack of access to relevant data may compel substandard methodologies. Summative evaluations of programs and projects concerning, say, aid to developing countries, where relevant data in some cases is near inaccessible, may imply an implicit ultimatum to accept lower methodological standards. Any evaluative claim would require the construction of a baseline and a likely counterfactual. The implicit third requirement is that the baseline and the counterfactual are constructed in accordance with recognized methods, i.e., that the source of authority is *science* in one or other sense. Lack of methodological rigour may in the long run undermine the legitimacy of social science based research.

The mere growth in the number of evaluations conducted worldwide and the growing number of consultancy companies implies that strong institutional forces are cementing evaluations both as an industry and evaluations as an institution itself. From a commercial point of view the cost-effectiveness of evaluation methodologies is an issue. Despite the numerous Evaluation Standards issued from organizations like the World Bank, the European Commission and national accounting offices it is not necessarily the case that future evaluation practices invariable contribute to public trust in social science based research.

References

Abelson, R. P. 1963. Computer Simulation of "Hot Cognition". In Tomkins, & Messick (Eds.), ***Computer Simulation of Personality***: 277-298. New York: Wiley.

Achen, C. H. 1986. ***The statistical analysis of quasi-experiments***. Berkeley: University of California Press.

Adelman, C. 1996. Anything Goes. ***Evaluation***, 2(3): 291-305.

Anscombe, G. E. M. 1958. On Brute facts. ***Analysis***, 18: 69-72.

Atwell, J. 1982. Human rights in human subjects research. In A. Kimmel (Ed.), ***Ethics of human subject research***: 81-91. San Francisco: Jossey-Bass.

Bamberger, M., Rugh, J., Church, M., & Fort, L. 2004. Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. ***American Journal of Evaluation***, 25(1): 5-37.

Baron, J. 1981. An Analysis of Confirmation Bias. ***Bulletin of the Psychonomic Society***, 18(2): 53-53.

Benshakhar, G. 1995. Confirmation Bias in Judgments of Cqt-Polygraph Examiners - Hypothetical Description and Preliminary-Results of the Examiner Decision-Maker Process. ***Integrative Physiological and Behavioral Science***, 30(1): 102-103.

Benzecri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In S. Watanabe (Ed.), ***Methodologies of Pattern Recognition***: 35-60. New York: Academic.

Bernstein, R. J. 1983. ***Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis***. Philadelphia: University of Pensylvania Press.

Bontekoe, R. 2000. ***Dimensions of the Hermeneutic Circle***. New York: Humanity Books.

Bourdieu, P. 1984. ***Distinction : a social critique of the judgement of taste***. London: Routledge & Kegan Paul.

Bourdieu, P. 1993. ***Sociology in question***. London: Sage.

Bourdieu, P. 1998. ***Acts of resistance : against the tyranny of the market***. New York: The New Press.

Bourdieu, P. 1999. ***The Weight of the World : social suffering in contemporary society***. Cambridge: Polity Press.

Bourdieu, P., & Coleman, J. S. (Eds.). 1991a. ***Social theory for a changing society***. Boulder, CO, US: Westview Press.

Bourdieu, P., & Coleman, J. S. 1991b. ***Social theory for a changing society***. Boulder, Colo.: Westview Press.

Brennan, G., & Buchanan, J. 1985. ***The reason of rules : constitutional political economy***. Cambridge: Cambridge University Press.

Campbell, D. T., & Stanley, J. C. 1963. ***Experimental and quasi-experimental designs for research***. Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. 1969. ***Experimental and quasi-experimental designs for research*** (5th print ed.). Chicago ,.

Carrol, J. D. 1998. Book Review. ***American Review of Public Administration***, 28(4): 402-407.

Cassirer, E. 1910. ***Substanzbegriff und Funtionsbegriff***. Berlin: B. Cassirer.

Chen, M., & Bargh, J. A. 1997. Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. ***Journal of Experimental Social Psychology***, 33(5): 541-560.

Christensen, T., & Laegreid, P. 1999. New Public Management - Design, Resistance, or Transformation. ***Public Productivity & Management Review***, 23(2): 169-193.

Christensen, T., Laegreid, P., & Wise, L. R. 2002. Transforming Administrative Policy. ***Public Administration***, 80(1): 153-178.

Coleman, J. S. 1972. ***Policy research in the social sciences***. Morrist., N.J. ,.

Cook, T. D., & Campbell, D. T. 1979. ***Quasi-experimentation Design & analysis issues for field settings***. Boston: Houghton Mifflin Co.

Creswell, J. W. 1998. ***Qualitative Inquiry and Research Design: Choosing Among Five Traditions***. London: Sage.

Cronbach, L. J. 1980. ***Towards Reform in Program Evaluation: Aims, Methods, and Institutional Arrangements*** (1st ed.). San Francisco: Jossey-Bass Publishers.

Davis, N. M. 1994. Combating Confirmation Bias. ***American Journal of Nursing***, 94(7): 17-17.

Dawes, R. M. 1979. Robust Beauty of Improper Linear-Models in Decision-Making. ***American Psychologist***, 34(7): 571-582.

Dimaggio, P. J., & Powell, W. W. 1983. The Iron Cage Revisited - Institutional Isomorphism and Collective Rationality in Organizational Fields. ***American Sociological Review***, 48(2): 147-160.

Donaldson, S. I. 2001. Overcoming our Negative Reputation: Evaluation Becomes Known as a Helping Profession. ***American Journal of Evaluation***, 22(3).

Eco, U. 1992. ***Interppretation and Overinterpretation***. Cambridge: Cambridge University Press.

Eisenhardt, K. M. 1989. Agency Theory: An assessment and review. ***Academy of Management Review***, 14(1): 57-73.

Estany, A. 2001. The thesis of theory-laden observation in the light of cognitive psychology. ***Philosophy of Science***, 68(2): 203-217.

Fahrbach, L. 2005. Understanding brute facts. ***Synthese***, 145(3): 449-466.

Flick, U. 2002. ***An introduction to qualitative research*** (2nd ed.). London: Sage.

Frank, R. H. 2000. Why is cost-benefit analysis so controversial? ***Journal of Legal Studies***, 29(2): 913-930.

Franklin, A., Anderson, M., Brock, D., Coleman, S., Downing, J., Gruvander, A., Lilly, J., Neal, J., Peterson, D., Price, M., Rice, R., Smith, L., Speirer, S., & Toering, D. 1989. Can a Theory-Laden Observation Test the Theory. ***British Journal for the Philosophy of Science***, 40(2): 229-231.

Gadenne, V. 1982. Confirmation Bias and Rationality of Cognitive-Processes. ***Psychologische Beitrage***, 24(1): 11-25.

Gadenne, V. 1984. Confirmation Bias and Confirmation Error - an Indispensable Differentiation. ***Psychologische Beitrage***, 26(4): 736-738.

Galaskiewicz, J., & Wasserman, S. 1989. Mimetic Processes within an Interorganizational Field - an Empirical-Test. ***Administrative Science Quarterly***, 34(3): 454-479.

Geva-May, I., & Pal, L. A. 1999. Good Fences Make Good Neighbours. ***Evaluation***, 5(3): 259-277.

Glaser, B., & Strauss, A. 1967. ***The discovery of grounded theory: Strategies for qualitative research***. Chicago: Aldine.

Goldberger, A. S. 1972. Structural equation models in the social sciences. ***Econometrica***, 40: 979-1001.

Gore, A. 1993. Creating a Government that Works Better and Costs Less: Report of the National Performance Review. Washington D.C.: U.S. Government Printing Office.

Greenacre, M. 1993. ***Correspondence Analysis in Practice.*** London: Academic Press.

Guba, E. G. 1990. ***The Paradigm dialog***. Newbury Park, Calif.: Sage Publications.

Guba, E. G., & Lincoln, Y. S. 1988. Naturalistic and rationalistic enquiry. In J. P. Keeves (Ed.), ***Educational Research, Methodology, and Measurement: An International Handbook***. Oxford: Pergamon Press.

Guba, E. G., & Lincoln, Y. S. 1989. ***Fourth generation evaluation***. Newbury Park, Calif.: Sage Publications.

Habermas, J. 1990. Discourse Ethics: Notes on a Program of Philosophical Justification. In S. Benhabib, & F. R. Dallmayr (Eds.), ***The Communicative ethics controversy***: viii, 378 s. Cambridge, Mass.: MIT Press.

Hanson, N. R. 1958. ***Patterns of Discovery***. Cambridge: Cambridge University Press.

Haveman, R. H. 1987. ***Poverty Policy and Poverty Research.*** Madison: University of Wisconsin Press.

Heckman, J. J. 1998. ***Characterizing selection bias using experimental data***. Cambridge, MA: National Bureau of Economic Research.

Heckman, J. J., Ichimura, H., & Todd, P. E. 1997. Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. ***Review of Economic Studies***, 64: 605-654.

Heckman, J. J., Smith, J., & Clements, N. 1997. Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts. ***Review of Economics Studies***, 64: 487-535.

Heckman, J. J., & Smith, J. A. 1995a. Assessing the Case for Social Experiment. ***Journal of Economic Perspectives***, 9(2): 85-110.

Heckman, J. J., & Smith, J. A. 1995b. Assessing the Case for Social Experiments. ***Journal of Economic Perspectives***, 9(2): 85-110.

Henkel, M. 2005. Academic identity and autonomy in a changing policy environment. ***Higher Education***, 49(1-2): 155-176.

Henry, G. T. 2001. How modern democracies are shaping evaluation and the emerging challenges for evaluation. ***American Journal of Evaluation***, 22(3): 419-429.

Hill, M. O. 1974. Correspondence Analysis - Neglected Multivariate Method. ***Journal of the Royal Statistical Society Series C-Applied Statistics***, 23(3): 340-354.

Holman, E. L. 1979. Problem of Theory-Laden Perception. ***Philosophical Studies***, 35(1): 91-99.

Hood, C. 1996. Exploring variations in public management reforms of the 1980s'. In H.A:G.M. Bekke, J.L. Perry, & T. A. J. Toonen (Eds.), ***Civil service systems in comparative perspective***. Bloomington: Indiana University Press.

House, E. R., & Howe, K. R. 1999. ***Values in evaluation and social research***. Thousand Oaks, Calif.: Sage.

Jensen, M., & Meckling, W. 1976. Theory of the firm: Managerial behavior, agency costs, and ownership structure. ***Journal of Financial Economics***, 3: 305-360.

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. 2001. Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. ***Journal of Personality and Social Psychology***, 80(4): 557-571.

Jones, M., & Sugden, R. 2001. Positive confirmation bias in the acquisition of information. ***Theory and Decision***, 50(1): 59-99.

Joreskog, K. G., & Goldberger, A. S. 1975. Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. ***Journal of the American Statistical Association***, 70(351): 631-639.

Judd, C. M., & Kenny, D. A. 1981. ***Estimating the effects of social interventions***. Cambridge ,: Cambridge university press.

Kaplan, A. 1964. ***The conduct of inquiry methodology for behavioral science***. Scranton, Pa.: Chandler.

Kelly, H. H. 1992. Common-sense psychology and scientific psychology. ***Annual Review of Psychology***, 43: 1-23.

Kemmis, S., & Stake, R. E. 1988. ***Evaluating curriculum***. Geelong, Vic.: Deakin University.

Klamer, A. 2001. Making sense of economice: from falsification to rhetoric and beyond. ***Journal of Economics Methodology***, 8(1): 69-75.

Knudsen, K., & Waerness, K. 2001. Kontant evaluering. ***Tidsskrift for velferdsforskning***, 4(4): 252-258.

Kornhauser, L. A. 2000. On Justifying Cost-Benefit Analysis. ***The Journal of Legal Studies***, xxix(2): 971-1004.

Krems, J. F., & Zierer, C. 1994. Are Experts Immune to Cognitive Bias - the Dependence of Confirmation Bias on Specialist Knowledge. ***Zeitschrift Fur Experimentelle Und Angewandte Psychologie***, 41(1): 98-115.

Kuhn, T. S. 1962. ***The structure of scientific revolutions***. Chicago: University of Chicago Press.

Laegreid, P. 2000. To Civil Servants Under Contract. ***Public Administration***, 78(4): 879-896.

Lewis, D. 1973. ***Counterfactuals***. Cambridge, MA: Harvard University Press.

Lewis, D. K. 1986. ***Counterfactuals*** (Repr. with corr. ed.). Oxford: Blackwell.

Luhmann, N. 1995. ***Social systems***. Stanford, Calif.: Stanford University Press.

Luukkonen, T. 2000. Additionality of EU framework programmes. ***Research Policy***, 29(6): 711-724.

MacCoun, R. J. 1998. Biases in the Interpretation and use of Research Results. ***Annual Review of Psychology***, 49: 259-287.

Maor, M. 1999. The Paradox of Managerialism. ***Public Administration Review***, 59(1): 5-18.

March , J. G., & Olsen, J. P. 2004. The logic of appropriateness: 1-28. Oslo: Centre for European Studies, University of Oslo.

Mohr, L. 1999. The Qualitative Method of Impact Analysis. ***American Journal of Evaluation***, 20(1): 69-84.

Moore, C., & Barresi, J. 1993. Knowledge of the Psychological States of Self and Others Is Not Only Theory-Laden but Also Data-Driven. ***Behavioral and Brain Sciences***, 16(1): 61-62.

Morgan, G., & Smircich, L. 1980. The Case for Qualitative Research. ***Academy of Management Review***, 5(4): 491-500.

Myhill, J. R. 1951. On the Ontological Significance of the Löwenheim-Skolem Theorem. In M. White (Ed.), ***Academic Freedom, Logic and Religion***: 57-70. Pittsburg: The University of Pennsylvania Press.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. 1977. Confirmation Bias in a Simulated Research Environment - Experimental-Study of Scientific Inference. ***Quarterly Journal of Experimental Psychology***, 29(FEB): 85-95.

Nathan, R. P. 1988. ***Social Science in Government: Uses and Misuses***. New York: Basic Books.

Nickerson, R. S. 1992. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. ***Review of General Psychology***, 2(2): 175-220.

Nickerson, R. S. 1998. Confirmation Bias: A Ubiguitous Phenomenon in Many Guises. ***Review of General Psychology***, 2(2): 177-220.

Olson, J. M., Roese, N. J., & P., Z. M. 1996. Expectancies. In E. T. Higgins, & A. W. Kruglanski (Eds.), ***Social Psychology: Handbook of Basic Principles***. New York: Guilford.

Onwuegbuzie, A. J. 2003. Effect sizes in qualitative research: A prolegomenon. ***Quality & Quantity***, 37(4): 393-409.

Onwuegbuzie, A. J., & Leech, N. L. 2007a. A call for qualitative power analyses. ***Quality & Quantity***, 41(1): 105-121.

Onwuegbuzie, A. J., & Leech, N. L. 2007b. Validity and qualitative research: An oxymoron? ***Quality & Quantity***, 41(2): 233-249.

Pearl, J. 2000. ***Causality***. Cambridge: Cambridge University Press.

Plowden, W. 1994. ***Ministers and Mandarins***. London: Institute for Public Policy Research.

Pollard, P. 1983. Confirming Confirmation Bias. ***Behavioral and Brain Sciences***, 6(2): 258-259.

Pollitt, C., & Bouckaert, G. 2000. ***Public Management reform: a comparative analysis***. Oxford: Oxford University Press.

Poslby, N. W. 1998. Social Science and Scientific Change: A Note on Thomas S. Kuhn's Contribution. ***Annual Review of Political Science***(1): 199-210.

Posner, R. A. 2000. Cost-benefit analysis: Definition, justification, and comment on conference papers. ***Journal of Legal Studies***, 29(2): 1153-1177.

Powell, W. W., & DiMaggio, P. J. 1991. ***The New Institutionalism in Organizational Analysis***. Chicago and London: University of Chicago Press.

Reichenbach, H. 1938. ***Experience and Prediction: An Analysis of the Foundatins and Structure of Knowledge***. Chicago, Illinois: University of Chicago Press.

Richardson, H. S. 2000. The Stupidity of the Cost-Benefit Analysis. ***The Journal of Legal Studies***, xxix(2): 971-1004.

Ricoeur, P. 1970. ***Freud and Philosophy: An Essay on Interpretation***. New Haven: Yale University Press.

Robins, J. M. 1989. The analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitutional studies. In L. Sechrest, H. Freeman, & A. Mulley (Eds.), ***Health Service Research Methodology: A Focus on AIDS***: 113-159. Washington DC: US Public Health Service.

Robins, J. M. 1997. Causal inference from complex longitudinal data. In M. Berkane (Ed.), ***Latent Variable Modeling and Applications to Causality: Lecture Notes in Statistics***. New York: Springer-Verlag.

Rosenbaum, P., & Rubin, D. B. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. ***American Statistician***, 39(33-38).

Rosenbaum, P. R. 1995. ***Observational studies***. New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. ***Biometrika***, 70: 41-55.

Rosenthal, R. 1976. ***Experimenter effects in behavioral research***. New York: Irvington.

Rosenthal, R. 1979. The File Drawer Problem and Tolerance for Null Results. ***Psychological Bulletin***, 86(3): 638-641.

Rosenthal, R., & Jacobson, L. 1968. ***Pygmalion in the classroom : teacher expectation and pupils' intellectual development***. New York: Holt Rinehart and Winston.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. 1999. ***Evaluation : a systematic approach.*** (6th ed.). Thousand Oaks, Calif.: Sage Publications.

Rotton, J., Foos, P. W., Vanmeek, L., & Levitt, M. 1995. Publication Practices and the File Drawer Problem - a Survey of Published Authors. ***Journal of Social Behavior and Personality***, 10(1): 1-13.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. ***Journal of Educational Psychology***, 66: 688-701.

Rubin, D. B. 1978. Bayesian inference for causal effects. ***Annals of Statistics***, 6: 34-58.

Rubin, D. B. 1990. Formal Modes of Statistical Inference for Causal Effects. ***Journal of Statistical Planning and Inference***, 25: 279-292.

Sanderson, I. 2002. Evaluation, policy learning and evidence-based policy making. ***Public Administration***, 80(1): 1-22.

Scaff, L. A., & Ingram, H. M. 1987. The Influence of theory on what you see. In D. J. Palumbo (Ed.), ***The Politics of Program Evaluation***. Newbury Park: Sage.

Schmitt, N., & Stults, D. M. 1985. Factors Defined by Negatively Keyed Items - the Result of Careless Respondents. ***Applied Psychological Measurement***, 9(4): 367-373.

Schwandt, T. A. 2000. Three Epistemological Stances for Qualitative Inquiry
Interpretivism, Hermeneutics, and Social Constructionism. In N. K. Denzin, & Y. S. Lincoln (Eds.), 2nd ed.: 189-213. Thousand Oaks, Calif: Sage.

Scott, W. R. 1995. ***Institutions and organizations***. Thousand Oaks: Sage Publications.

Scriven, M. 1991. ***Evaluation thesaurus*** (4th ed.). Newbury Park, Calif.: Sage Publications.

Scriven, M. 1993. ***Hard-won lessons in program evaluation***. San Francisco: Jossey-Bass.

Sen, A. 2000. The Discipline of Cost-Benefit Analysis. ***The Journal of Legal Studies***, xxix(2): 931-952.

Shadish, W. R., Doherty, M., & Montgomery, L. M. 1989. How Many Studies Are in the File Drawer - an Estimate from the Family Marital Psychotherapy Literature. ***Clinical Psychology Review***, 9(5): 589-603.

Skinner, B. F. 1953. ***Science and human behavior***. New York: Macmillan.

Snyder, M., & Stukas, A. A. 1999. Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. ***Annual Review of Psychology***, 50: 273-303.

Stake, R. E. 1975. ***Evaluating the arts in education : a responsive approach***. Columbus, Ohio: Merrill.

Stake, R. E. 1986. ***Issues in research on evaluation : improving the study of transition programs for adolescents with handicaps***. Champaign, Ill.: College of Education University of Illinois.

Stake, R. E. 1995. ***The art of case study research***. Thousand Oaks: Sage Publications.

Stake, R. E., Easley, J. A., & Anastasiou, C. J. 1978. ***Case studies in science education***. Urbana

Washington: Center for Instructional Research and Curriculum Evaluation University of Illinois at Urbana-Champaign ;

for sale by the Supt. of Docs. U.S. Govt. Print. Off.

Stufflebaum, D., Guba, E. G., & Tyler, R. 1971. ***Educational Evaluation and Decision Making***. New York: Peacock Publishers.

Tolchin, S. J. 1987. The political uses of evaluations research: Cost-benefit analysis and the cotton dust standard. In D. J. Palumbo (Ed.), ***The Politics of Program Evaluation***. Newbury Park: Sage.

Van Maanen, J. 1979. Reclaiming Qualitative Methods for Organizational Research: A preface. ***Administrative Science Quarterly***, 24(4): 520-526.

Vedung, E. 2000. ***Public Policy and Program Evaluation***. New Brunswick: Transaction Publishers.

Wallis, J., & Dollery, B. 1999. ***Market failure, government failure, leadership and public policy***. Basingstoke: Macmillan.

Webb, E., Campbell, D., Schwartz, R., & Sechest, L. 1966. ***Unobtrusive measures***. New York: Rand-McNally.

Winship, C., & Morgan, S. L. 1999. The Estimation of Causal Effects from Observational Data. ***Annual Review of Sociology***, 25: 659-706.

Wittgenstein, L. 1953. ***Philosophical Investigations***. New York: MacMillan.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. 1997. ***Program Evaluation Alternative: Approaches and Practical Guidelines*** (Second Edition ed.). New York: Longman.

Aaron, H. J., Gramlich, E. M., Hanushek, E. A., Heckman, J. J., & Wildawsky, A. 1990. Social Science Research and Policy. ***The Journal of Human Resource***, 25(2): 297-304.

Appendix A

Rosenbaum Bounds - Initial Matched Pairs – Original and corrected samples[1]

Regional Venture Capital Loans
Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0001	437.5	437.5	217.7	680.5
1.05	0.0003	393.5	485.5	171.0	734.5
1.10	0.0010	352.5	530.0	128.5	788.5
1.15	0.0031	310.0	571.0	88.0	838.5
1.20	0.0082	272.5	612.8	51.5	889.5
1.25	0.0189	237.0	656.5	13.5	944.5
1.30	0.0383	202.5	701.0	-23.0	998.3
1.35	0.0702	167.0	740.5	-59.5	1047.0
1.40	0.1170	133.0	782.5	-95.5	1101.0
1.45	0.1798	102.5	821.0	-128.0	1154.0
1.5	0.2573	74.0	858.0	-161.0	1203.5

N=343

Regional Venture Capital Loans
Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0004	365.0	365.0	148.5	580.0
1.05	0.0014	324.0	401.5	110.5	626.0
1.10	0.0040	286.8	436.5	75.0	670.0
1.15	0.0097	254.1	475.5	43.0	711.5
1.20	0.0210	222.5	511.0	8.0	751.0
1.25	0.0407	190.7	541.5	-25.5	792.0
1.30	0.0715	157.5	572.0	-60.5	829.5
1.35	0.1155	128.5	602.5	-94.4	867.5
1.40	0.1734	99.5	641.0	-123.5	908.0
1.45	0.2440	73.5	671.5	-153.0	950.5
1.50	0.3248	50.0	703.0	-185.0	988.0

N=279

Investment Grants
Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	738.5	738.5	514.0	983.0
1.05	0.0000	675.3	802.5	457.0	1051.5
1.10	0.0000	618.5	863.5	404.5	1120.0
1.15	0.0000	564.0	923.5	356.0	1188.5
1.20	0.0000	514.0	983.0	309.5	1258.5
1.25	0.0000	466.5	1039.7	264.5	1324.5
1.30	0.0000	423.0	1095.5	223.5	1392.5
1.35	0.0001	381.0	1152.5	183.5	1462.0
1.40	0.0003	342.0	1209.0	145.0	1528.0
1.45	0.0011	304.5	1265.5	109.0	1592.4
1.50	0.0033	268.0	1320.0	73.5	1658.5
1.55	0.0086	234.5	1375.0	40.5	1724.5
1.60	0.0198	201.0	1430.0	9.0	1790.5
1.65	0.0407	169.5	1485.0	-21.0	1857.0
1.70	0.0753	138.7	1538.5	-50.5	1926.5
1.75	0.1269	109.5	1591.0	-78.5	1993.5
1.80	0.1966	81.0	1645.5	-106.0	2060.5
1.85	0.2829	54.0	1697.0	-133.0	2129.5
1.90	0.3813	27.5	1750.0	-160.0	2200.0
1.95	0.4851	3.0	1802.2	-185.5	2268.0
2.00	0.5874	-20.5	1856.6	-210.5	2334.5

N=619

Investment Grants
Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	636.5	636.5	428.5	863.5
1.05	0.0000	586.0	687.4	381.0	919.0
1.10	0.0000	540.0	738.5	337.5	973.5
1.15	0.0000	495.0	787.0	296.0	1026.9
1.20	0.0000	453.5	834.8	257.0	1079.5
1.25	0.0000	415.0	879.0	219.5	1131.5
1.30	0.0001	378.0	923.0	184.5	1183.0
1.35	0.0003	342.5	966.5	149.5	1235.0
1.40	0.0009	309.5	1009.5	116.5	1288.0
1.45	0.0026	278.0	1050.0	84.0	1338.0
1.50	0.0063	248.0	1092.5	53.0	1387.0
1.55	0.0138	218.0	1134.5	23.0	1434.5
1.60	0.0274	190.5	1174.5	-4.0	1483.0
1.65	0.0499	162.0	1216.5	-31.5	1532.5
1.70	0.0838	136.0	1258.5	-57.0	1578.5
1.75	0.1308	110.5	1297.0	-82.5	1623.5
1.80	0.1915	85.0	1336.5	-108.0	1669.0
1.85	0.2646	61.5	1374.0	-132.0	1716.6
1.90	0.3472	37.0	1412.0	-156.0	1765.0
1.95	0.4354	15.0	1451.0	-178.5	1811.0
2.00	0.5247	-7.0	1487.0	-201.5	1857.5

N=470

[1] Original samples include duplicated records in both treatment cases and the control cases. Duplicates and outliers are removed in the corrected samples.

Rosenbaum Bounds - Initial Matched Pairs – Original and corrected samples

Both Venture Capital Loans & Investment Grants
Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	696.8	696.8	445.5	968.8
1.05	0.0000	636.5	758.0	389.5	1041.0
1.10	0.0000	578.0	819.5	333.0	1113.5
1.15	0.0000	526.0	881.0	283.0	1185.1
1.20	0.0001	472.5	938.5	235.5	1255.5
1.25	0.0003	425.0	997.0	185.5	1322.5
1.30	0.0009	380.0	1053.0	141.0	1387.5
1.35	0.0029	334.5	1111.5	98.6	1452.5
1.40	0.0078	294.5	1168.0	56.0	1515.2
1.45	0.0181	255.5	1225.5	16.5	1584.0
1.50	0.0374	215.5	1279.5	-22.6	1649.5
1.55	0.0695	178.5	1333.5	-60.5	1708.3
1.60	0.1176	143.1	1384.5	-99.5	1772.5
1.65	0.1830	109.5	1434.5	-136.5	1833.0
1.70	0.2645	77.0	1486.5	-173.0	1901.5
1.75	0.3582	42.0	1537.0	-210.5	1961.0
1.80	0.4584	12.2	1591.4	-248.0	2020.0
1.85	0.5585	-17.5	1641.5	-284.0	2079.5
1.90	0.6525	-48.0	1689.0	-321.0	2142.8
1.95	0.7359	-79.5	1736.5	-354.5	2204.0
2.00	0.8061	-108.0	1786.9	-390.0	2261.5

N=491

Both Venture Capital Loans & Investment Grants
Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	718.0	718.0	478.0	981.0
1.05	0.0000	668.0	772.0	430.0	1041.5
1.10	0.0000	619.0	821.8	384.0	1102.5
1.15	0.0000	572.5	873.0	340.0	1165.8
1.20	0.0000	531.5	920.5	299.0	1225.5
1.25	0.0000	489.5	967.0	262.5	1282.5
1.30	0.0001	449.5	1015.1	225.4	1339.5
1.35	0.0002	415.0	1061.5	189.0	1395.5
1.40	0.0006	379.0	1109.0	155.0	1451.5
1.45	0.0016	345.0	1159.3	120.5	1505.7
1.50	0.0036	312.6	1205.0	88.0	1562.5
1.55	0.0076	283.0	1249.5	55.0	1615.0
1.60	0.0149	255.0	1293.5	25.0	1668.0
1.65	0.0268	227.5	1337.0	-3.5	1719.8
1.70	0.0451	198.3	1379.6	-33.0	1771.6
1.75	0.0714	172.0	1419.5	-60.0	1822.0
1.80	0.1068	146.0	1462.5	-89.5	1876.5
1.85	0.1520	120.5	1505.5	-116.5	1929.5
1.90	0.2066	96.0	1548.3	-142.5	1978.5
1.95	0.2695	72.5	1588.5	-167.5	2027.8
2.00	0.3389	47.0	1629.5	-195.5	2081.0

N=376

Financial Schemes: Entire Dataset
Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	644.0	644.0	505.5	791.8
1.05	0.0000	585.1	703.5	451.0	856.0
1.10	0.0000	532.4	761.5	401.0	918.5
1.15	0.0000	482.0	819.0	353.5	980.5
1.20	0.0000	435.5	875.0	308.5	1041.0
1.25	0.0000	392.0	930.0	265.5	1100.0
1.30	0.0000	350.5	984.5	225.0	1159.1
1.35	0.0000	310.5	1037.5	186.0	1217.5
1.40	0.0000	273.0	1090.0	148.5	1274.5
1.45	0.0001	237.0	1141.5	113.0	1331.5
1.50	0.0007	202.0	1193.0	78.0	1387.5
1.55	0.0039	169.0	1243.3	45.0	1442.8
1.60	0.0156	137.0	1292.5	12.5	1498.3
1.65	0.0477	106.0	1343.0	-19.5	1552.5
1.70	0.1158	75.5	1391.0	-50.5	1608.5
1.75	0.2300	47.0	1439.5	-80.5	1662.5
1.80	0.3843	18.5	1487.5	-109.0	1716.1
1.85	0.5549	-9.0	1535.0	-138.0	1771.1
1.90	0.7119	-36.5	1582.8	-166.5	1824.5
1.95	0.8337	-62.5	1630.5	-194.0	1878.5
2.00	0.9144	-88.5	1676.5	-221.0	1932.0

N=1453

Financial Schemes: Entire Dataset
Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	584.5	584.5	455.5	722.5
1.05	0.0000	538.0	633.5	411.5	774.5
1.10	0.0000	493.0	681.0	368.5	825.0
1.15	0.0000	451.5	727.5	328.0	874.0
1.20	0.0000	413.0	772.5	290.1	923.0
1.25	0.0000	376.0	816.2	254.5	970.5
1.30	0.0000	340.5	859.5	219.5	1017.0
1.35	0.0000	306.6	902.0	186.5	1063.0
1.40	0.0000	275.0	943.0	154.5	1108.5
1.45	0.0000	245.0	983.5	123.5	1154.7
1.50	0.0003	214.5	1024.0	93.5	1198.7
1.55	0.0014	186.5	1063.2	65.0	1242.0
1.60	0.0054	159.0	1102.0	37.0	1285.0
1.65	0.0170	132.0	1141.5	10.2	1329.0
1.70	0.0438	106.5	1180.0	-16.0	1371.0
1.75	0.0954	81.0	1217.5	-42.0	1411.5
1.80	0.1790	57.1	1254.0	-67.0	1453.5
1.85	0.2940	33.0	1290.5	-92.0	1493.6
1.90	0.4309	11.0	1328.0	-115.5	1534.0
1.95	0.5730	-11.5	1363.9	-139.0	1574.5
2.00	0.7030	-34.0	1398.0	-161.5	1615.0

N=1125

Rosenbaum Bounds - Initial Matched Pairs – Original and corrected samples

The FRAM Program

Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0072	221.5	221.5	43.0	396.0
1.05	0.0223	179.5	261.5	5.0	433.5
1.10	0.0556	142.0	299.0	-33.0	471.0
1.15	0.1159	106.5	334.0	-68.5	508.0
1.20	0.2071	71.0	367.0	-104.0	543.5
1.25	0.3253	40.0	399.0	-140.0	579.5
1.30	0.4587	9.5	428.5	-171.5	615.0
1.35	0.5918	-20.5	458.5	-205.0	650.0
1.40	0.7109	-49.0	488.5	-237.0	681.0
1.45	0.8075	-77.0	516.0	-267.5	712.5
1.50	0.8792	-105.0	544.5	-299.0	743.0

N=430

The RFAM Program

Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0104	203.0	203.0	30.5	376.5
1.05	0.0290	164.5	240.5	-5.0	412.5
1.10	0.0669	130.0	277.0	-40.0	447.0
1.15	0.1312	95.5	311.0	-73.0	479.5
1.20	0.2240	64.0	341.0	-107.0	512.5
1.25	0.3402	35.0	371.5	-140.0	544.5
1.30	0.4685	7.5	400.0	-169.0	575.5
1.35	0.5954	-21.0	427.0	-201.5	608.0
1.40	0.7089	-46.5	453.0	-229.5	638.5
1.45	0.8019	-72.5	479.3	-259.0	667.0
1.50	0.8722	-99.0	505.0	-287.5	693.0

N=391

The Network Program

Rosenbaum Bounds - Original Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	903.7	903.7	567.3	1260.5
1.05	0.0000	756.1	1053.6	429.5	1421.1
1.10	0.0001	621.0	1200.5	300.0	1580.0
1.15	0.0010	494.3	1344.5	178.0	1736.0
1.20	0.0090	376.0	1487.3	64.0	1890.0
1.25	0.0474	263.5	1626.0	-45.0	2040.4
1.30	0.1579	157.5	1764.5	-150.9	2189.5
1.35	0.3590	57.0	1900.0	-251.5	2338.0
1.40	0.6016	-39.9	2032.8	-350.6	2484.5
1.45	0.8035	-133.5	2164.0	-446.3	2630.0
1.50	0.9238	-223.0	2295.6	-541.5	2773.0

N=1482

The Network Program

Rosenbaum Bounds - Corrected Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	671.5	671.5	362.0	1004.3
1.05	0.0002	544.1	803.5	240.0	1144.5
1.10	0.0026	426.0	932.0	126.0	1281.5
1.15	0.0188	313.5	1057.5	17.5	1417.3
1.20	0.0818	209.5	1181.2	-85.0	1551.1
1.25	0.2309	110.3	1300.7	-182.8	1682.5
1.30	0.4582	15.5	1420.0	-278.0	1812.0
1.35	0.6922	-75.0	1537.3	-369.4	1936.8
1.40	0.8616	-161.5	1652.4	-459.2	2061.5
1.45	0.9508	-244.5	1766.5	-547.0	2183.0
1.50	0.9861	-326.5	1878.0	-632.1	2306.0

N=1405

Over the next pages we show the development of the Rosenbaum Bounds for the Financial Schemes for matched pairs over time. The changes in the number of cases and Rosenbaum Bounds reflect the effects of attrition. This development is shown for the entire Financial Schemes dataset only. A complete detailed overview would require 17*5* = 170 different tables.

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1990- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	244.5	244.5	146.0	347.5
1.05	0.0000	217.0	272.0	118.5	378.0
1.10	0.0001	191.0	299.0	93.5	406.5
1.15	0.0004	166.0	325.0	69.5	435.0
1.20	0.0019	143.0	350.5	46.5	464.0
1.25	0.0071	120.5	376.0	24.5	491.0
1.30	0.0214	99.5	399.5	3.5	518.0
1.35	0.0530	79.5	423.5	-16.5	545.0
1.40	0.1106	60.5	447.0	-37.0	569.5
1.45	0.1995	41.5	470.0	-56.5	594.5
1.50	0.3166	23.5	492.5	-74.5	619.5
1.55	0.4511	6.5	514.5	-93.0	643.0
1.60	0.5871	-10.5	537.0	-111.0	667.0
1.65	0.7098	-27.5	557.0	-128.0	691.0
1.70	0.8093	-43.5	577.5	-145.5	714.5
1.75	0.8828	-58.5	597.5	-162.5	737.0
1.80	0.9324	-73.0	617.5	-178.5	761.5
1.85	0.9633	-88.0	637.5	-194.5	783.5
1.90	0.9812	-103.0	656.5	-210.5	806.5
1.95	0.9909	-117.5	676.0	-226.0	828.0
2.00	0.9958	-131.0	695.5	-241.5	850.0

N=654 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1991- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	275.0	275.0	144.5	410.5
1.05	0.0002	239.0	310.5	108.5	449.5
1.10	0.0009	204.5	346.0	75.0	487.5
1.15	0.0043	172.5	380.0	44.0	524.5
1.20	0.0152	142.5	413.5	13.5	561.5
1.25	0.0423	112.5	445.0	-15.5	598.0
1.30	0.0965	84.5	477.0	-44.5	633.5
1.35	0.1853	58.5	508.0	-72.0	668.5
1.40	0.3073	33.0	538.0	-98.5	702.5
1.45	0.4499	8.5	568.5	-124.5	737.5
1.50	0.5944	-15.7	598.0	-148.0	773.0
1.55	0.7230	-39.0	627.0	-173.0	806.5
1.60	0.8247	-61.5	655.5	-196.5	840.0
1.65	0.8971	-83.5	683.0	-219.0	873.0
1.70	0.9438	-105.0	711.5	-242.0	906.0
1.75	0.9713	-126.0	740.0	-264.0	939.0
1.80	0.9863	-145.5	768.0	-285.0	970.0
1.85	0.9938	-165.5	796.0	-306.0	1002.5
1.90	0.9974	-184.5	823.5	-327.0	1034.0
1.95	0.9989	-203.5	849.0	-347.0	1065.0
2.00	0.9996	-221.5	876.0	-367.0	1097.0

N=640 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1992- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	448.0	448.0	286.0	626.0
1.05	0.0000	405.5	491.5	243.0	675.5
1.10	0.0000	365.5	535.5	202.5	723.5
1.15	0.0000	326.5	580.5	165.5	771.0
1.20	0.0002	289.0	623.0	130.0	819.5
1.25	0.0009	254.0	663.5	94.5	865.5
1.30	0.0034	219.0	703.5	61.5	914.0
1.35	0.0101	188.0	741.0	30.0	961.0
1.40	0.0258	158.0	781.5	-1.0	1008.0
1.45	0.0563	129.0	820.5	-30.5	1053.0
1.50	0.1076	99.5	858.0	-60.0	1098.0
1.55	0.1832	73.0	897.5	-87.5	1143.5
1.60	0.2815	47.5	935.5	-114.5	1187.0
1.65	0.3960	21.5	973.0	-140.5	1232.0
1.70	0.5166	-3.5	1012.0	-165.5	1274.0
1.75	0.6324	-27.0	1047.5	-191.0	1316.0
1.80	0.7345	-50.5	1083.5	-216.0	1356.0
1.85	0.8178	-73.0	1120.5	-240.0	1395.5
1.90	0.8810	-95.5	1155.5	-263.5	1435.5
1.95	0.9259	-117.0	1190.0	-285.5	1475.0
2.00	0.9560	-137.0	1227.0	-307.0	1512.5

N=597 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1993- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	676.5	676.5	475.0	902.0
1.05	0.0000	622.0	733.0	425.5	965.0
1.10	0.0000	572.0	787.0	377.5	1026.0
1.15	0.0000	525.5	841.0	333.0	1086.0
1.20	0.0000	481.5	893.5	290.0	1145.5
1.25	0.0000	441.0	945.5	249.0	1204.5
1.30	0.0000	399.5	996.5	211.0	1262.0
1.35	0.0001	362.0	1047.0	173.5	1318.5
1.40	0.0004	327.0	1094.5	139.0	1375.5
1.45	0.0012	291.5	1143.0	104.5	1432.5
1.50	0.0035	259.0	1190.0	71.5	1487.0
1.55	0.0089	226.5	1237.0	39.5	1542.5
1.60	0.0198	197.5	1284.0	9.5	1596.5
1.65	0.0398	167.0	1329.5	-19.5	1648.5
1.70	0.0725	139.0	1375.0	-48.5	1700.5
1.75	0.1208	111.0	1420.5	-76.0	1751.0
1.80	0.1859	85.0	1464.5	-102.5	1801.5
1.85	0.2666	59.0	1509.0	-130.5	1849.0
1.90	0.3593	34.5	1553.0	-155.5	1899.0
1.95	0.4584	10.0	1595.5	-180.0	1944.5
2.00	0.5576	-13.5	1636.5	-205.0	1991.0

N=579 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1994- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	1001.0	1001.0	755.0	1263.0
1.05	0.0000	936.3	1066.5	696.5	1336.5
1.10	0.0000	875.0	1131.8	640.5	1409.0
1.15	0.0000	817.0	1193.0	585.5	1483.0
1.20	0.0000	763.5	1253.0	536.5	1552.5
1.25	0.0000	715.5	1314.0	487.0	1618.5
1.30	0.0000	668.5	1373.0	440.0	1685.5
1.35	0.0000	623.0	1433.0	397.0	1753.0
1.40	0.0000	579.0	1492.5	352.5	1815.5
1.45	0.0000	540.0	1548.5	311.5	1879.0
1.50	0.0000	499.5	1602.0	271.0	1944.5
1.55	0.0000	460.5	1655.0	234.5	2003.0
1.60	0.0002	425.5	1710.5	196.5	2061.5
1.65	0.0005	388.5	1763.0	161.0	2121.0
1.70	0.0013	354.0	1814.0	127.5	2180.5
1.75	0.0032	320.0	1864.0	94.0	2239.5
1.80	0.0071	288.0	1916.0	59.5	2298.0
1.85	0.0143	258.5	1965.0	29.0	2356.0
1.90	0.0265	228.5	2012.0	-3.5	2415.5
1.95	0.0459	198.5	2058.0	-34.5	2475.0
2.00	0.0742	170.5	2106.0	-64.5	2527.0

N=573 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1995- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	905.0	905.0	637.0	1193.0
1.05	0.0000	836.5	977.5	572.0	1270.5
1.10	0.0000	770.0	1046.0	511.0	1345.5
1.15	0.0000	707.5	1114.0	451.5	1420.0
1.20	0.0000	649.5	1179.5	395.5	1495.0
1.25	0.0000	595.0	1242.5	342.0	1569.0
1.30	0.0000	543.0	1305.0	291.5	1642.5
1.35	0.0000	495.0	1367.5	245.0	1714.0
1.40	0.0002	447.0	1426.5	199.0	1785.5
1.45	0.0007	401.5	1487.0	156.0	1856.5
1.50	0.0021	358.0	1546.0	116.0	1927.5
1.55	0.0054	317.0	1607.0	73.0	1995.5
1.60	0.0126	276.5	1665.0	34.5	2061.0
1.65	0.0262	239.5	1722.5	-2.5	2126.5
1.70	0.0496	202.0	1780.0	-38.5	2191.0
1.75	0.0857	168.5	1836.0	-76.0	2253.5
1.80	0.1369	135.5	1892.5	-112.5	2319.0
1.85	0.2036	102.0	1948.0	-146.5	2379.5
1.90	0.2842	69.0	2003.0	-179.5	2439.5
1.95	0.3749	38.5	2053.0	-212.5	2499.0
2.00	0.4706	9.2	2105.0	-245.5	2557.5

N=563 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1996- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	717.0	717.0	423.0	1026.5
1.05	0.0000	641.5	793.0	351.5	1109.5
1.10	0.0000	570.0	866.5	283.0	1189.5
1.15	0.0003	501.0	939.5	216.5	1271.5
1.20	0.0012	436.5	1013.0	154.5	1348.5
1.25	0.0043	376.0	1081.0	96.0	1423.5
1.30	0.0128	318.5	1147.5	39.5	1498.0
1.35	0.0320	263.0	1214.0	-15.5	1574.0
1.40	0.0685	211.0	1279.5	-69.5	1648.5
1.45	0.1281	160.5	1341.0	-123.5	1717.5
1.50	0.2132	113.0	1403.5	-175.5	1787.5
1.55	0.3202	67.0	1462.5	-221.5	1855.5
1.60	0.4407	21.0	1522.5	-269.0	1924.5
1.65	0.5631	-22.0	1583.0	-314.0	1993.0
1.70	0.6764	-66.5	1643.0	-360.0	2060.0
1.75	0.7728	-108.5	1698.5	-405.0	2126.0
1.80	0.8486	-150.5	1753.0	-449.5	2191.5
1.85	0.9041	-191.5	1810.0	-491.5	2256.5
1.90	0.9421	-227.0	1864.5	-535.5	2320.5
1.95	0.9666	-265.5	1918.5	-576.5	2383.5
2.00	0.9816	-301.0	1973.0	-617.5	2442.5

N=566 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -1997- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	736.3	736.3	379.0	1106.5
1.05	0.0002	647.5	827.0	291.0	1209.0
1.10	0.0009	561.0	913.0	204.0	1306.0
1.15	0.0036	476.0	995.5	128.5	1399.0
1.20	0.0116	402.5	1077.0	56.5	1495.0
1.25	0.0310	329.5	1162.5	-17.0	1582.0
1.30	0.0695	260.5	1243.5	-89.0	1670.0
1.35	0.1339	191.5	1319.5	-156.5	1761.0
1.40	0.2264	130.5	1396.3	-220.5	1845.5
1.45	0.3424	72.0	1476.5	-287.5	1930.0
1.50	0.4709	13.0	1547.0	-351.5	2018.5
1.55	0.5983	-44.5	1615.5	-409.5	2096.0
1.60	0.7126	-101.5	1687.5	-468.0	2172.5
1.65	0.8060	-155.5	1759.5	-524.0	2258.0
1.70	0.8763	-206.0	1829.0	-586.5	2338.5
1.75	0.9254	-258.5	1893.5	-645.5	2416.5
1.80	0.9572	-311.0	1963.0	-698.5	2498.0
1.85	0.9767	-361.0	2031.0	-754.5	2581.5
1.90	0.9879	-406.5	2093.0	-809.0	2656.5
1.95	0.9940	-454.0	2151.5	-862.0	2731.0
2.00	0.9971	-497.5	2215.5	-918.0	2808.0

N=536 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 1998 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	1139.0	1139.0	707.0	1607.5
1.05	0.0000	1028.5	1250.0	602.0	1728.5
1.10	0.0000	927.0	1357.5	505.5	1847.0
1.15	0.0001	833.5	1458.0	414.0	1970.5
1.20	0.0002	747.3	1561.0	322.0	2089.0
1.25	0.0010	660.5	1655.0	237.5	2198.5
1.30	0.0033	579.0	1756.5	158.5	2313.0
1.35	0.0091	501.5	1851.5	82.0	2426.0
1.40	0.0217	426.5	1953.0	13.0	2533.0
1.45	0.0456	355.0	2048.0	-59.5	2636.5
1.50	0.0854	284.5	2137.5	-126.0	2744.0
1.55	0.1443	219.5	2226.5	-193.5	2848.5
1.60	0.2228	155.5	2317.0	-260.0	2950.0
1.65	0.3179	95.0	2407.5	-324.5	3044.0
1.70	0.4233	38.0	2495.5	-383.5	3143.0
1.75	0.5313	-18.0	2574.5	-440.0	3244.5
1.80	0.6342	-72.5	2656.0	-498.5	3345.5
1.85	0.7259	-125.0	2742.0	-555.0	3441.5
1.90	0.8026	-178.5	2822.5	-609.0	3537.5
1.95	0.8632	-232.0	2901.5	-664.0	3628.0
2.00	0.9088	-282.8	2980.5	-721.5	3729.0

N=506 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 1999 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	1188.5	1188.5	698.5	1714.0
1.05	0.0000	1073.5	1306.5	583.0	1847.5
1.10	0.0000	967.5	1422.5	475.0	1984.5
1.15	0.0003	861.5	1538.5	373.0	2129.0
1.20	0.0010	760.5	1648.5	276.0	2271.0
1.25	0.0034	662.5	1753.0	181.0	2404.5
1.30	0.0095	571.0	1863.0	90.0	2539.0
1.35	0.0230	485.0	1973.5	8.5	2676.0
1.40	0.0483	402.0	2087.5	-76.5	2812.0
1.45	0.0902	323.0	2198.5	-152.5	2937.0
1.50	0.1519	248.0	2308.5	-225.0	3061.0
1.55	0.2334	173.5	2415.5	-294.5	3184.0
1.60	0.3311	101.0	2521.5	-359.5	3311.0
1.65	0.4383	36.5	2632.5	-429.5	3434.0
1.70	0.5470	-30.5	2736.5	-496.5	3561.0
1.75	0.6493	-94.0	2841.0	-558.5	3677.0
1.80	0.7393	-154.0	2939.0	-624.5	3810.5
1.85	0.8139	-210.0	3037.0	-686.0	3938.0
1.90	0.8722	-262.5	3134.0	-744.0	4065.0
1.95	0.9155	-318.0	3227.5	-802.5	4195.5
2.00	0.9460	-367.5	3327.3	-860.0	4317.0

N=477 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds -2000 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0000	1148.5	1148.5	599.0	1767.5
1.05	0.0001	1013.5	1286.5	475.5	1927.0
1.10	0.0004	898.5	1420.5	363.5	2083.5
1.15	0.0017	787.5	1545.5	259.5	2233.0
1.20	0.0057	674.5	1671.5	154.5	2384.0
1.25	0.0153	574.0	1805.5	59.0	2543.5
1.30	0.0355	474.0	1929.5	-39.0	2695.0
1.35	0.0716	385.5	2051.0	-123.0	2834.0
1.40	0.1282	301.5	2170.0	-209.5	2979.0
1.45	0.2067	215.0	2298.0	-290.5	3133.5
1.50	0.3044	136.0	2414.5	-370.0	3268.0
1.55	0.4145	62.5	2539.5	-449.0	3405.0
1.60	0.5280	-15.5	2652.5	-523.5	3542.5
1.65	0.6361	-84.5	2769.0	-598.0	3682.5
1.70	0.7315	-150.5	2881.5	-670.0	3826.5
1.75	0.8103	-218.0	2994.0	-734.0	3964.5
1.80	0.8715	-281.5	3115.0	-802.5	4096.0
1.85	0.9164	-344.5	3215.0	-870.0	4238.0
1.90	0.9477	-405.5	3325.5	-930.5	4379.5
1.95	0.9684	-464.5	3434.5	-1000.5	4511.0
2.00	0.9816	-522.0	3539.5	-1062.5	4643.5

N=457 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2001- Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0008	839.5	839.5	322.5	1413.0
1.05	0.0032	720.0	964.0	207.0	1545.0
1.10	0.0105	608.5	1083.5	93.5	1679.0
1.15	0.0279	506.0	1198.5	-12.5	1816.5
1.20	0.0624	407.5	1310.5	-111.5	1958.0
1.25	0.1204	313.5	1422.5	-206.0	2087.5
1.30	0.2044	221.5	1528.0	-296.5	2216.5
1.35	0.3112	133.5	1636.0	-382.0	2343.5
1.40	0.4321	46.5	1741.5	-461.0	2467.5
1.45	0.5553	-34.5	1851.5	-538.0	2598.0
1.50	0.6695	-112.0	1959.0	-614.5	2726.0
1.55	0.7668	-189.0	2057.5	-684.0	2858.0
1.60	0.8435	-259.5	2160.0	-756.5	2980.5
1.65	0.9000	-326.5	2263.5	-831.5	3110.0
1.70	0.9390	-391.5	2361.5	-904.0	3234.5
1.75	0.9643	-452.3	2456.3	-967.5	3344.5
1.80	0.9800	-513.5	2556.0	-1031.5	3467.5
1.85	0.9892	-571.5	2655.5	-1105.5	3589.5
1.90	0.9944	-631.5	2756.5	-1173.5	3698.0
1.95	0.9972	-682.5	2855.0	-1229.5	3818.5
2.00	0.9986	-736.0	2951.5	-1296.0	3930.5

N=429 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2002 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0268	591.0	591.0	-7.5	1180.5
1.05	0.0656	461.0	717.0	-123.0	1311.5
1.10	0.1337	338.0	835.5	-244.0	1444.5
1.15	0.2336	219.5	954.0	-359.0	1572.5
1.20	0.3588	110.0	1059.5	-465.0	1709.5
1.25	0.4954	2.5	1169.0	-579.0	1847.0
1.30	0.6274	-85.5	1273.5	-683.5	1974.0
1.35	0.7418	-181.0	1373.5	-791.5	2097.5
1.40	0.8318	-273.0	1477.0	-899.0	2211.0
1.45	0.8967	-362.0	1577.0	-1004.5	2329.0
1.50	0.9400	-445.5	1684.5	-1111.0	2433.0
1.55	0.9669	-535.0	1790.0	-1210.0	2541.5
1.60	0.9826	-616.0	1894.0	-1306.5	2657.0
1.65	0.9913	-696.5	1989.0	-1402.0	2766.0
1.70	0.9958	-778.5	2084.5	-1501.0	2880.5
1.75	0.9981	-862.5	2173.0	-1596.5	2987.5
1.80	0.9991	-945.5	2262.5	-1690.5	3092.0
1.85	0.9996	-1025.0	2349.5	-1777.0	3202.0
1.90	0.9998	-1105.0	2428.0	-1880.0	3299.0
1.95	0.9999	-1181.0	2511.0	-1978.0	3389.5
2.00	1.0000	-1256.5	2591.0	-2073.5	3474.0

N=397 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2003 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0325	637.5	637.5	-38.5	1314.5
1.05	0.0755	496.5	778.5	-180.0	1462.5
1.10	0.1475	362.0	912.5	-310.5	1605.0
1.15	0.2496	233.0	1046.5	-437.5	1747.5
1.20	0.3742	113.0	1161.5	-559.5	1890.0
1.25	0.5079	-7.0	1279.5	-676.0	2025.5
1.30	0.6356	-116.5	1399.0	-785.0	2160.0
1.35	0.7459	-224.5	1512.0	-888.0	2292.5
1.40	0.8328	-326.0	1618.5	-995.5	2415.0
1.45	0.8959	-423.0	1729.5	-1099.5	2543.0
1.50	0.9385	-518.5	1836.8	-1206.5	2660.0
1.55	0.9654	-609.5	1945.5	-1316.0	2771.5
1.60	0.9814	-694.0	2048.5	-1424.0	2880.0
1.65	0.9904	-780.5	2153.0	-1529.0	2988.5
1.70	0.9953	-861.5	2256.0	-1629.5	3092.5
1.75	0.9977	-942.5	2350.5	-1728.0	3205.0
1.80	0.9990	-1018.5	2445.0	-1821.0	3312.5
1.85	0.9995	-1098.0	2540.5	-1911.5	3427.0
1.90	0.9998	-1181.5	2630.5	-1994.5	3537.5
1.95	0.9999	-1263.0	2714.0	-2078.5	3641.0
2.00	1.0000	-1343.0	2794.8	-2160.0	3742.0

N=375 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2004 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0147	733.5	733.5	69.5	1488.0
1.05	0.0375	597.0	874.5	-57.5	1660.5
1.10	0.0802	465.0	1021.5	-181.5	1817.5
1.15	0.1483	344.5	1162.5	-304.0	1969.5
1.20	0.2420	229.5	1290.5	-423.0	2117.5
1.25	0.3555	118.0	1431.5	-549.5	2269.5
1.30	0.4785	17.0	1556.5	-669.0	2404.0
1.35	0.5989	-76.5	1691.5	-796.5	2535.0
1.40	0.7068	-174.0	1811.0	-920.5	2661.0
1.45	0.7960	-267.5	1924.5	-1046.5	2788.5
1.50	0.8647	-359.5	2045.5	-1166.5	2916.5
1.55	0.9142	-452.0	2154.0	-1286.0	3045.5
1.60	0.9478	-549.5	2270.0	-1412.5	3179.5
1.65	0.9695	-641.0	2375.5	-1524.5	3289.0
1.70	0.9828	-733.5	2470.5	-1643.5	3426.0
1.75	0.9907	-825.0	2574.5	-1763.5	3555.5
1.80	0.9951	-923.5	2663.5	-1873.5	3681.0
1.85	0.9975	-1017.0	2760.0	-1980.0	3823.5
1.90	0.9988	-1107.0	2853.0	-2097.5	3941.0
1.95	0.9994	-1195.5	2956.0	-2205.0	4076.0
2.00	0.9997	-1287.0	3047.0	-2309.0	4205.5

N=353 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2005 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0004	1226.0	1226.0	505.5	2024.5
1.05	0.0016	1077.5	1375.5	362.0	2190.0
1.10	0.0050	934.0	1517.0	236.0	2348.5
1.15	0.0128	804.0	1663.0	104.0	2499.5
1.20	0.0287	684.0	1797.3	-20.0	2648.5
1.25	0.0565	574.0	1931.0	-138.5	2801.0
1.30	0.1000	474.0	2067.0	-254.0	2960.0
1.35	0.1610	361.0	2190.0	-362.5	3118.5
1.40	0.2388	262.5	2308.5	-469.5	3270.5
1.45	0.3299	166.0	2425.0	-574.0	3419.5
1.50	0.4289	67.8	2542.5	-675.5	3552.0
1.55	0.5291	-23.5	2655.5	-778.0	3702.5
1.60	0.6245	-116.5	2767.5	-874.5	3851.5
1.65	0.7103	-200.0	2891.0	-979.5	4005.5
1.70	0.7835	-284.5	3006.0	-1075.0	4147.5
1.75	0.8432	-366.5	3125.8	-1164.5	4275.0
1.80	0.8897	-446.0	3239.5	-1256.5	4389.0
1.85	0.9246	-525.5	3347.5	-1347.5	4525.0
1.90	0.9498	-600.5	3458.5	-1443.0	4655.0
1.95	0.9674	-676.0	3552.5	-1530.5	4783.5
2.00	0.9793	-754.8	3665.3	-1625.0	4914.0

N=330 Matched Pairs

All Kinds: Venture Capital & Investment Grant

Rosenbaum Bounds - 2004 - Sample

Γ	p-critical	Hodges-Lehmnann point estimate			
		t_{max}	t_{min}	CI_{max}	CI_{min}
1.00	0.0010	1290.8	1290.8	461.5	2187.5
1.05	0.0033	1133.0	1447.5	314.0	2371.0
1.10	0.0088	991.5	1602.0	171.5	2550.0
1.15	0.0202	851.5	1747.5	37.0	2731.0
1.20	0.0410	721.5	1898.0	-96.0	2915.5
1.25	0.0746	599.5	2044.0	-210.5	3099.0
1.30	0.1235	472.5	2171.5	-318.5	3280.5
1.35	0.1883	361.0	2301.5	-422.5	3465.0
1.40	0.2673	256.0	2449.0	-530.0	3649.0
1.45	0.3568	149.0	2580.0	-635.0	3840.5
1.50	0.4515	49.0	2713.0	-736.5	4030.5
1.55	0.5459	-50.0	2854.5	-835.5	4216.5
1.60	0.6349	-138.5	2987.0	-937.0	4398.0
1.65	0.7148	-223.0	3120.0	-1036.0	4574.0
1.70	0.7834	-303.5	3254.0	-1131.0	4770.0
1.75	0.8398	-378.0	3394.0	-1239.0	4941.5
1.80	0.8845	-461.5	3526.0	-1338.0	5114.0
1.85	0.9188	-536.0	3655.5	-1436.0	5286.5
1.90	0.9442	-609.0	3801.5	-1541.0	5450.0
1.95	0.9625	-684.0	3940.5	-1637.0	5611.0
2.00	0.9753	-754.3	4069.5	-1735.0	5777.0

N=291 Matched Pairs

Appendix B

Panel data estimates – difference –in-differences

Tab **B1a** Regional Venture Capital Loans – Treatment Period 1995 - 2006

Independent variables	Fixed effect estimates	95% Conf. Interval	GLS random effect estimates	95% Conf. Interval	RE-FE difference
Treatment-Control	-		742.01		
			(553.11)		
Period After Intervent.	782.29		793.80		-11.51
	(164.13)		(163.36)		
Treatm x Intervent.	**733.97**	**[278.93 1189.01]**	**754.52**	**[301.73 1207.32]**	**-20.55**
	(232.11)		(231.02)		
Constant	3434.22		2813.08		
	(83.65)		(391.11)		
F	*54.03*				
p-value	*0.000*				
Wald chi-square			*117.43*		
p-value			*0.000*		
Hausman chi-square					*4.60*
p-value					*0.100*
Number of obs =	*5442*				
Number of groups =	*422*				

Tab **B1b** Regional Venture Capital Loans – Treatment Period 1991 - 2006

Independent variables	Fixed effect estimates	95% Conf. Interval	GLS random effect estimates	95% Conf. Interval	RE-FE difference
Treatment-Control	-		220.66		
			(591.03)		
Period After Intervent.	432.11		450.23		-18.12
	(211.77)		(210.70)		
Treatm x Intervent.	**1125.64**	**[538.52 1712.77]**	**1135.40**	**[551.40 1719.41]**	**-9.76**
	(299.49)		(297.97)		
Constant	3251.03		2830.62		
	(134.40)		(417.92)		
F	*29.14*				
p-value	*0.000*				
Wald chi-square			*65.15*		
p-value			*0.000*		
Hausman chi-square					*2.43*
p-value					*0.296*
Number of obs =	*5442*				
Number of groups =	*422*				

Tab **B2a** Investment Grants – Treatment Period 1995 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		1759.53		
			(1011.69)		
Period After Intervent.	1869.90		1926.95		-57.05
	(416.99)		(414.77)		
Treatm x Intervent.	**1615.50**	**[459.55 2771.46]**	**1620.98**	**[471.31 2770.64]**	**-5.48**
	(589.72)		(586.57)		
Constant	6178.56		4610.28		
	(224.54)		(715.37)		
F	*44.99*				
p-value	*0.000*				
Wald chi-square			*102.50*		
p-value			*0.000*		
Hausman chi-square					*3.87*
p-value					*0.144*
Number of obs =	*11912*				
Number of groups =	*856*				

Tab **B2b** Investment Grants – Treatment Period 1991 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		582.51		
			(689.74)		
Period After Intervent.	1597.178		1608.96		-11.77
	(211.15)		(210.71)		
Treatm x Interventi.	**888.49**	**[303.14 1473.83]**	**898.59**	**[314.55 1482.62]**	**-10.10**
	(298.61)		(297.98)		
Constant	3863.68		3497.59		
	(138.42)		(487.72)		
F	*97.90*				
p-value	*0.000*				
Wald chi-square			*204.39*		
p-value			*0.000*		
Hausman chi-square					*3.27*
p-value					*0.1948*
Number of obs =	*11912*				
Number of groups =	*856*				

Tab **B3a** Both Investment Grants & Venture Capital Loans – Treatment Period 1995 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		1363.25		
			(528.15)		
Period After Intervent.	1737.38		1751.22		-13.84
	(198.06)		(196.77)		
Treatm. x Intervent.	**1814.99**	**[1265.97 2364.02]**	**1829.45**	**[1284.04 2374.87]**	**-14.46**
	(280.09)		(278.28)		
Constant	4940.12		3957.46		
	(103.53)		(373.46)		
F	*199.33*				
p-value	*0.000*				
Wald chi-square			*430.63*		
p-value			*0.000*		
Hausman chi-square					*1.96*
p-value					*0.375*
Number of obs =	*12414*				
Number of groups =	*920*				

Tab **B3b** Both Investment Grants & Venture Capital Loans – Treatment Period 1991 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		351.88		
			(595.36)		
Period After Intervent.	1510.11		1508.98		1.13
	(270.60)		(268.15)		
Treatm x Interventi.	**2409.68**	**[1659.56 3159.80]**	**2414.80**	**[1671.54 3158.05]**	**-5.12**
	(382.68)		(379.21)		
Constant	4169.13		3634.47		
	(174.00)		(420.98)		
F	*120.49*				
p-value	*0.000*				
Wald chi-square			*265.88*		
p-value			*0.000*		
Hausman chi-square					*0.01*
p-value					*0.993*
Number of obs =	*12414*				
Number of groups =	*920*				

Tab **B4a** The FRAM Program – Treatment Period 1997 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		124.79		
			(258.74)		
Period After Intervent.	1673.33		1659.52		13.81
	(90.72)		(90.43)		
Treatm. x Intervent.	**358.71**	**[106.57 610.85]**	**369.82**	**[118.59 621.04]**	**-11.11**
	(128.63)		(128.18)		
Constant	3636.69		3509.86		
	(44.11)		(182.93)		
F	*418.39*				
p-value	*0.0000*				
Wald chi-square			*837.09*		
p-value			*0.0000*		
Hausman chi-square					*3.65*
p-value					*0.1613*
Number of obs =	*10964*				
Number of groups =	*760*				

Tab **B4b** The FRAM Program – Treatment Period 1992 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		-145.38		
			(288.70)		
Period After Intervent.	1626.95		1607.13		19.82
	(120.07)		(119.88)		
Treatm x Interventi.	**516.79**	**[183.64 849.92]**	**522.25**	**[189.73 854.77]**	**-5.46**
	(169.95)		(169.66)		
Constant	2987.36		2940.14		
	(76.89)		(204.11)		
F	*250.65*				
p-value	*0.0000*				
Wald chi-square			*495.61*		
p-value			*0.0000*		
Hausman chi-square					*12.35*
p-value					*0.0021*
Number of obs =	*10964*				
Number of groups =	*760*				

Tab **B5a** The Network Program – Treatment Period 1997 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		2120.05		
			(712.85)		
Period After Intervent.	2166.02		2149.04		16.98
	(176.82)		(176.37)		
Treatm. x Intervent.	**3254.64**	**[2764.50 3744.77]**	**3241.60**	**[2752.73 3730.46]**	**13.04**
	(250.07)		(249.43)		
Constant	8415.87		7298.92		
	(81.97)		(504.06)		
F	*544.92*				
p-value	*0.0000*				
Wald chi-square			*1106.03*		
p-value			*0.0000*		
Hausman chi-square					*7.44*
p-value					*0.0243*
Number of obs =	*38058*				
Number of groups =	*2728*				

Tab **B5b** The Network Program – Treatment Period 1993 - 2006

Independent variables	**Fixed effect estimates**	**95% Conf. Interval**	**GLS random effect estimates**	**95% Conf. Interval**	**RE-FE difference**
Treatment-Control	-		1505.77		
			(731.85)		
Period After Intervent.	2431.69		2414.05		17.64
	(194.33)		(194.06)		
Treatm x Interventi.	**2707.81**	**[2169.15 3246.47]**	**2695.87**	**[2157.96 3233.77]**	**11.94**
	(274.82)		(274.45)		
Constant	7380.38		6448.16		
	(116.58)		(517.50)		
F	*428.03*				
p-value	*0.0000*				
Wald chi-square			*871.48*		
p-value			*0.000*		
Hausman chi-square					*11.52*
p-value					*0.0031*
Number of obs =	*38058*				
Number of groups =	*2728*				

Appendix C Size distributions for Costs of Financial Schemes

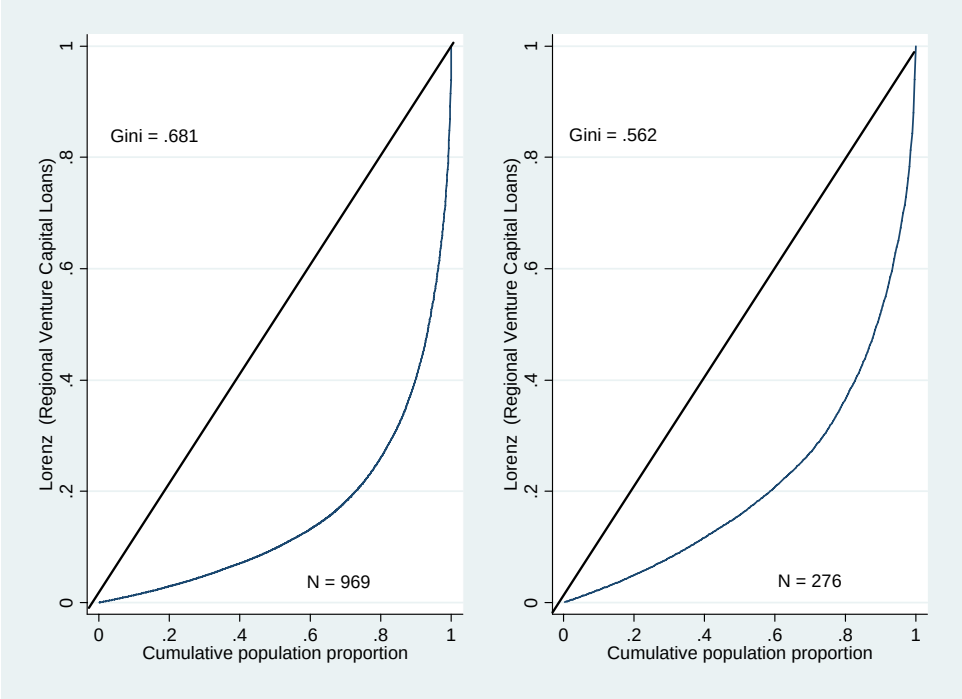


Figure C1 Regional Venture Capital Loans

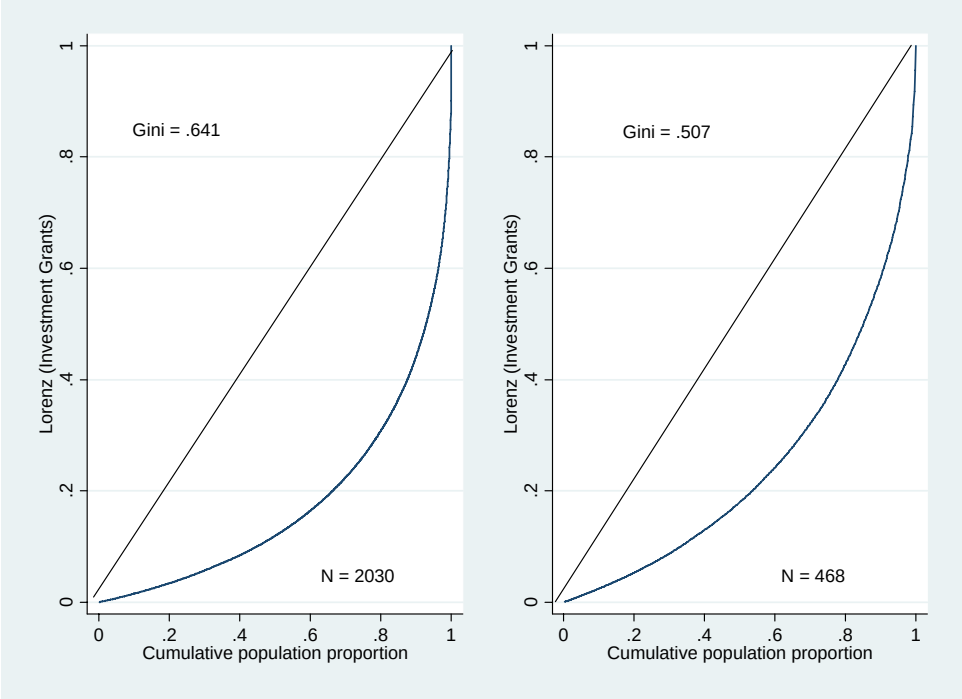


Figure C2 Investment Grants

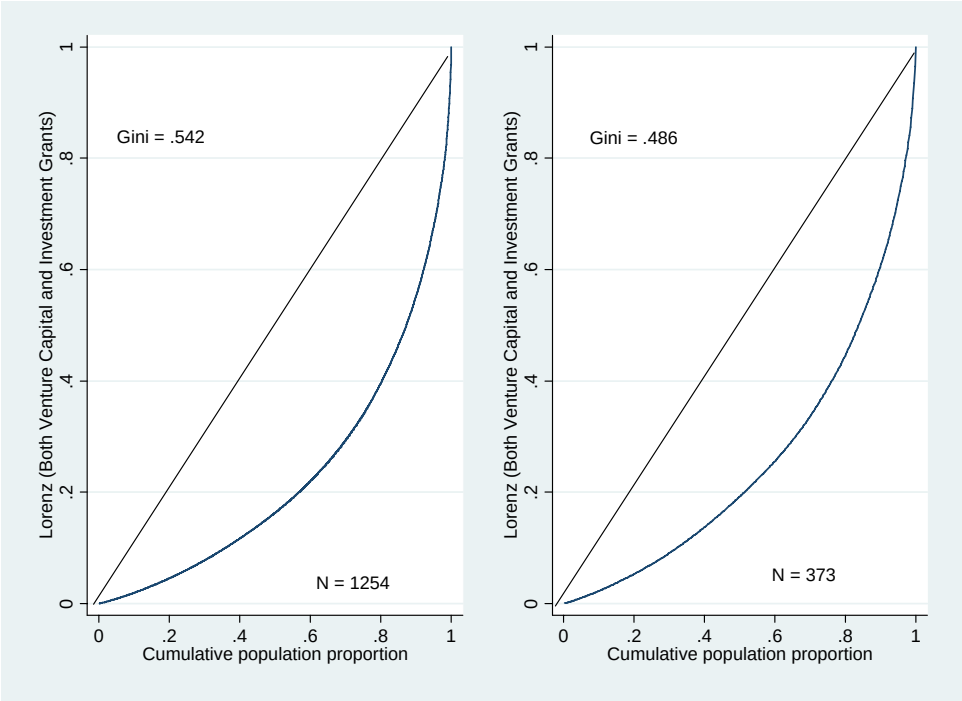


Figure C3 Venture Capital Loans and Investment Grants Combined

Appendix D

Table D1 Differences between Analyzed and Total Means – Mill. NOK

Kind of Allocation	Sample	Year	Obs.	Mean	Std. Dev.	[95% Conf.		Difference	*t* (diff).)
Regional Venture	Total	1990	354	0,6815	1,507	0,5240	0,8390	0,1298	0,89
Capital Loans	Analyzed	1990	119	0,5517	0,921	0,3845	0,7189		
	Total	1991	238	0,7112	1,389	0,5338	0,8886	0,3747	2,37
	Analyzed	1991	82	0,3365	0,583	0,2084	0,4646		
	Total	1992	228	1,2278	3,930	0,7149	1,7407	0,6864	1,17
	Analyzed	1992	45	0,5414	0,710	0,3281	0,7547		
	Total	1993	95	0,9993	2,517	0,4866	1,5120	0,0533	0,08
	Analyzed	1993	17	0,9460	2,045	-0,1054	1,9974		
	Total	1994	54	1,1483	2,793	0,3860	1,9106	0,7713	0,99
	Analyzed	1994	13	0,3770	0,350	0,1655	0,5885		
Investment Grants	Total	1990	312	1,0567	8,534	0,1061	2,0073	0,7414	0,72
	Analyzed	1990	69	0,3153	0,320	0,2384	0,3922		
	Total	1991	320	0,6267	1,188	0,4960	0,7574	0,2691	1,90
	Analyzed	1991	72	0,3576	0,331	0,2798	0,4354		
	Total	1992	325	0,7648	1,607	0,5894	0,9402	0,2751	1,33
	Analyzed	1992	63	0,4897	0,726	0,3069	0,6725		
	Total	1993	469	0,7341	2,021	0,5507	0,9175	0,3364	1,54
	Analyzed	1993	86	0,3977	0,359	0,3207	0,4747		
	Total	1994	604	0,6613	1,139	0,5703	0,7523	0,0987	1,03
	Analyzed	1994	178	0,5626	1,046	0,4079	0,7173		
Both Categories	Total	1990	359	0,8691	1,503	0,7131	1,0251	0,0507	0,32
The Loan Part	Analyzed	1990	121	0,8184	1,605	0,5295	1,1073		
	Total	1991	328	0,8228	1,146	0,6983	0,9473	0,0837	0,73
	Analyzed	1991	111	0,7391	0,693	0,6087	0,8695		
	Total	1992	269	0,9949	2,118	0,7406	1,2492	0,2566	0,96
	Analyzed	1992	71	0,7383	1,442	0,3970	1,0796		
	Total	1993	186	0,7253	0,881	0,5979	0,8527	0,0961	0,61
	Analyzed	1993	34	0,6292	0,581	0,4265	0,8319		
	Total	1994	112	1,1574	3,447	0,5120	1,8028	0,5185	0,90
	Analyzed	1994	36	0,6389	0,643	0,4213	0,8565		
Both Categories	Total	1990	359	0,6385	1,042	0,5303	0,7467	0,1309	1,31
The Investment Part	Analyzed	1990	121	0,5076	0,606	0,3985	0,6167		
	Total	1991	328	0,5980	0,999	0,4895	0,7065	0,0490	0,48
	Analyzed	1991	111	0,5490	0,681	0,4209	0,6771		
	Total	1992	269	0,6392	1,034	0,5151	0,7633	0,1247	0,91
	Analyzed	1992	71	0,5145	1,020	0,2731	0,7559		
	Total	1993	186	0,5256	0,627	0,4349	0,6163	0,1412	1,27
	Analyzed	1993	34	0,3844	0,392	0,2476	0,5212		
	Total	1994	112	0,9506	3,043	0,3808	1,5204	0,2972	0,58
	Analyzed	1994	36	0,6534	0,791	0,3858	0,9210		

Table D2 Statistics for Analyzed Sample – Financial Schemes – Mill. NOK

	Venture Capital Loans				Investment Grants				Both Regional Venture Capital Loans and Investment Grants: Capital Loans Part				Investment Grants Part		
Year	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	St.dev	Sum
1990	0,5517	119	0,921	65,65	0,3153	69	0,320	21,75	0,8184	121	1,605	99,03	0,5076	0,606	212,69
1991	0,3365	82	0,583	27,59	0,3576	72	0,331	25,75	0,7391	111	0,693	82,04	0,5490	0,681	61,42
1992	0,5414	45	0,710	24,36	0,4897	63	0,726	30,85	0,7383	71	1,442	52,42	0,5145	1,020	60,94
1993	0,9460	17	2,045	16,08	0,3977	86	0,359	34,20	0,6292	34	0,581	21,39	0,3844	0,392	36,53
1994	0,3770	13	0,350	4,90	0,5626	178	1,046	100,14	0,6389	36	0,643	23,00	0,6534	0,791	13,07
Sum	*1013,8*			*138,6*				*212,7*				*277,9*			*384,6*
N	*1117*	*276*				*468*				*373*					

Table D3 Statistics for All Available data – Financial Schemes – Mill. NOK

	Venture Capital Loans				Investment Grants				Both Regional Venture Capital Loans and Investment Grants: Capital Loans Part				Investment Grants Part		
Year	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	N	St.dev	Sum	Mean	St.dev	Sum
1990	0,6815	354	1,507	241,25	1,0567	312	8,534	329,70	0,8691	359	1,503	312,01	0,6385	1,042	229,24
1991	0,7112	238	1,389	169,26	0,6267	320	1,188	200,53	0,8228	328	1,146	269,87	0,5980	0,999	196,14
1992	1,2278	228	3,930	279,94	0,7648	325	1,607	248,55	0,9949	269	2,118	267,63	0,6392	1,034	171,95
1993	0,9993	95	2,517	94,93	0,7341	469	2,021	344,28	0,7253	186	0,881	134,91	0,5256	0,627	97,76
1994	1,1483	54	2,793	62,01	0,6613	604	1,139	399,42	1,1574	112	3,447	129,63	0,9506	3,043	106,46
Sum	*4285,5*			*847,4*				*1522,5*				*1114,0*			*801,6*
N	*4253*	*969*				*2030*				*1254*					

Table D4 Future values of initial endowments for Venture Capital Loans – Analyzed Allocations – Mill. NOK

					No Loss						**30 % loss over 10 years**						**30 % loss over 5 years**					
					Assumed discount rate 7%						Discount rate 7% - 3%=4%						Discount rate 7% - 6%=1%					
		Ad-		Cum-	Future values for different						Future values for different						Future values for different					
	Total	justed		ula-	starting poins - Mill. NOK						starting poins - Mill. NOK						starting poins - Mill. NOK					
Year	Alloca-tions	Alloca-tions *	N	tive N	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum
1990	65,7	79,4	119	119	*79,4*					79,4	*79,4*					79,4	*79,4*					79,4
1991	27,6	33,4	82	201	*85,0*	*33,4*				118,4	*82,6*	*33,4*				116,0	*80,2*	*33,4*				113,6
1992	24,4	29,5	45	246	*90,9*	*35,7*	*29,5*			156,1	*85,9*	*34,7*	*29,5*			150,1	*81,0*	*33,7*	*29,5*			144,2
1993	16,1	19,5	17	263	*97,3*	*38,2*	*31,5*	*19,5*		186,5	*89,4*	*36,1*	*30,7*	*19,5*		175,6	*81,8*	*34,1*	*29,8*	*19,5*		165,1
1994	4,9	5,9	13	276	*104,1*	*40,9*	*33,7*	*20,8*	*5,9*	205,5	*92,9*	*37,6*	*31,9*	*20,2*	*5,9*	188,5	*82,7*	*34,4*	*30,1*	*19,7*	*5,9*	172,7
1995				276	*111,4*	*43,8*	*36,1*	*22,3*	*6,3*	219,9	*96,6*	*39,1*	*33,2*	*21,0*	*6,2*	196,1	*83,5*	*34,7*	*30,4*	*19,8*	*6,0*	174,4
1996				276	*119,2*	*46,8*	*38,6*	*23,8*	*6,8*	235,3	*100,5*	*40,6*	*34,5*	*21,9*	*6,4*	203,9	*84,3*	*35,1*	*30,7*	*20,0*	*6,0*	176,2
1997				276	*127,6*	*50,1*	*41,3*	*25,5*	*7,3*	251,8	*104,5*	*42,2*	*35,9*	*22,8*	*6,7*	212,1	*85,2*	*35,4*	*31,0*	*20,2*	*6,1*	177,9
1998				276	*136,5*	*53,6*	*44,2*	*27,3*	*7,8*	269,4	*108,7*	*43,9*	*37,3*	*23,7*	*6,9*	220,5	*86,0*	*35,8*	*31,3*	*20,4*	*6,2*	179,7
1999				276	*146,0*	*57,4*	*47,3*	*29,2*	*8,3*	288,2	*113,1*	*45,7*	*38,8*	*24,6*	*7,2*	229,4	*86,9*	*36,1*	*31,6*	*20,7*	*6,2*	181,5
2000				276	*156,3*	*61,4*	*50,6*	*31,2*	*8,9*	308,4	*117,6*	*47,5*	*40,3*	*25,6*	*7,5*	238,5	*87,7*	*36,5*	*31,9*	*20,9*	*6,3*	183,3
2001				276	*167,2*	*65,7*	*54,2*	*33,4*	*9,5*	330,0	*122,3*	*49,4*	*42,0*	*26,6*	*7,8*	248,1	*88,6*	*36,9*	*32,2*	*21,1*	*6,4*	185,2
2002				276	*178,9*	*70,3*	*58,0*	*35,8*	*10,2*	353,1	*127,2*	*51,4*	*43,6*	*27,7*	*8,1*	258,0	*89,5*	*37,2*	*32,6*	*21,3*	*6,4*	187,0
2003				276	*191,4*	*75,2*	*62,0*	*38,3*	*10,9*	377,8	*132,3*	*53,4*	*45,4*	*28,8*	*8,4*	268,3	*90,4*	*37,6*	*32,9*	*21,5*	*6,5*	188,9
2004				276	*204,8*	*80,5*	*66,4*	*41,0*	*11,7*	404,3	*137,6*	*55,6*	*47,2*	*30,0*	*8,8*	279,1	*91,3*	*38,0*	*33,2*	*21,7*	*6,5*	190,8
2005				276	*219,2*	*86,1*	*71,0*	*43,8*	*12,5*	432,6	*143,1*	*57,8*	*49,1*	*31,2*	*9,1*	290,2	*92,2*	*38,4*	*33,5*	*21,9*	*6,6*	192,7
2006				276	*234,5*	*92,1*	*76,0*	*46,9*	*13,4*	462,9	*148,8*	*60,1*	*51,0*	*32,4*	*9,5*	301,8	*93,1*	*38,8*	*33,9*	*22,1*	*6,7*	194,6
Sum	**138,6**	167,7			* *Assumed Revenue Effects due to tax financing is 21%*																	

Table D5 Future values of initial endowments for Venture Capital Loans – All Allocations – Mill. NOK

					No Loss						**30 % loss over 10 years**						**30 % loss over 5 years**					
					Assumed discount rate 7%						Discount rate 7% - 3%=4%						Discount rate 7% - 6%=1%					
	Total	Ad-justed		Cum-ula-	Future values for different starting poins - Mill. NOK						Future values for different starting poins - Mill. NOK						Future values for different starting poins - Mill. NOK					
Year	Alloca-tions	Alloca-tions *	N	tive N	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum
1990	241,2	291,9	354	354	*291,9*					291,9	*291,9*					291,9	*291,9*					291,9
1991	169,3	204,8	238	592	*312,3*	*204,8*				517,1	*303,6*	*204,8*				508,4	*294,8*	*204,8*				499,6
1992	279,9	338,7	228	820	*334,2*	*219,1*	*338,7*			892,1	*315,7*	*213,0*	*338,7*			867,5	*297,8*	*206,9*	*338,7*			843,4
1993	94,9	114,9	95	915	*357,6*	*234,5*	*362,4*	*114,9*		1069,4	*328,4*	*221,5*	*352,3*	*114,9*		1017,0	*300,8*	*208,9*	*342,1*	*114,9*		966,7
1994	62,0	75,0	54	969	*382,6*	*250,9*	*387,8*	*122,9*	*75,0*	1219,3	*341,5*	*230,4*	*366,4*	*119,5*	*75,0*	1132,7	*303,8*	*211,0*	*345,5*	*116,0*	*75,0*	1051,4
1995				969	*409,4*	*268,5*	*415,0*	*131,5*	*80,3*	1304,6	*355,2*	*239,6*	*381,0*	*124,2*	*78,0*	1178,0	*306,8*	*213,1*	*349,0*	*117,2*	*75,8*	1061,9
1996				969	*438,1*	*287,2*	*444,0*	*140,7*	*85,9*	1396,0	*369,4*	*249,2*	*396,3*	*129,2*	*81,2*	1225,2	*309,9*	*215,3*	*352,5*	*118,3*	*76,5*	1072,5
1997				969	*468,7*	*307,4*	*475,1*	*150,6*	*91,9*	1493,7	*384,1*	*259,1*	*412,1*	*134,4*	*84,4*	1274,2	*313,0*	*217,4*	*356,0*	*119,5*	*77,3*	1083,2
1998				969	*501,6*	*328,9*	*508,3*	*161,1*	*98,4*	1598,2	*399,5*	*269,5*	*428,6*	*139,8*	*87,8*	1325,1	*316,1*	*219,6*	*359,6*	*120,7*	*78,1*	1094,1
1999				969	*536,7*	*351,9*	*543,9*	*172,4*	*105,2*	1710,1	*415,5*	*280,3*	*445,7*	*145,3*	*91,3*	1378,1	*319,3*	*221,8*	*363,2*	*121,9*	*78,9*	1105,0
2000				969	*574,2*	*376,5*	*582,0*	*184,4*	*112,6*	1829,8	*432,1*	*291,5*	*463,6*	*151,2*	*94,9*	1433,3	*322,5*	*224,0*	*366,8*	*123,2*	*79,6*	1116,0
2001				969	*614,4*	*402,9*	*622,7*	*197,4*	*120,5*	1957,9	*449,4*	*303,2*	*482,1*	*157,2*	*98,7*	1490,6	*325,7*	*226,2*	*370,5*	*124,4*	*80,4*	1127,2
2002				969	*657,4*	*431,1*	*666,3*	*211,2*	*128,9*	2095,0	*467,4*	*315,3*	*501,4*	*163,5*	*102,7*	1550,2	*328,9*	*228,5*	*374,2*	*125,6*	*81,2*	1138,5
2003				969	*703,5*	*461,3*	*713,0*	*226,0*	*137,9*	2241,6	*486,1*	*327,9*	*521,5*	*170,0*	*106,8*	1612,2	*332,2*	*230,8*	*377,9*	*126,9*	*82,1*	1149,9
2004				969	*752,7*	*493,5*	*762,9*	*241,8*	*147,6*	2398,5	*505,5*	*341,0*	*542,3*	*176,8*	*111,1*	1676,7	*335,5*	*233,1*	*381,7*	*128,2*	*82,9*	1161,4
2005				969	*805,4*	*528,1*	*816,3*	*258,7*	*157,9*	2566,4	*525,7*	*354,7*	*564,0*	*183,9*	*115,5*	1743,8	*338,9*	*235,4*	*385,5*	*129,4*	*83,7*	1173,0
2006				969	*861,8*	*565,1*	*873,4*	*276,8*	*169,0*	2746,1	*546,7*	*368,8*	*586,6*	*191,3*	*120,1*	1813,5	*342,3*	*237,8*	*389,4*	*130,7*	*84,5*	1184,7
Sum	847,4	1025,3			* *Assumed Revenue Effects due to tax financing is 21%*																	

Table D6 Future values of initial endowments for Investment Grants – Analyzed data & All Allocations – Mill. NOK

Analyzed data

Year	Total Allocations	Adjusted Allocations *	N	Cumulative N	Assumed discount rate 7% Future values for different starting poins - Mill. NOK 1990	1991	1992	1993	1994	Sum
1990	21,8	26,3	69	69	*26,3*					26,3
1991	25,7	31,2	72	141	*28,2*	*31,2*				59,3
1992	30,8	37,3	63	204	*30,1*	*33,3*	*37,3*			100,8
1993	34,2	41,4	86	290	*32,2*	*35,7*	*39,9*	*41,4*		149,2
1994	100,1	121,2	178	468	*34,5*	*38,2*	*42,7*	*44,3*	*121,2*	280,9
1995				468	*36,9*	*40,8*	*45,7*	*47,4*	*129,7*	300,5
1996				468	*39,5*	*43,7*	*48,9*	*50,7*	*138,7*	321,5
1997				468	*42,3*	*46,8*	*52,4*	*54,2*	*148,4*	344,1
1998				468	*45,2*	*50,0*	*56,0*	*58,0*	*158,8*	368,1
1999				468	*48,4*	*53,5*	*59,9*	*62,1*	*169,9*	393,9
2000				468	*51,8*	*57,3*	*64,1*	*66,5*	*181,8*	421,5
2001				468	*55,4*	*61,3*	*68,6*	*71,1*	*194,6*	451,0
2002				468	*59,3*	*65,6*	*73,4*	*76,1*	*208,2*	482,6
2003				468	*63,4*	*70,2*	*78,6*	*81,4*	*222,8*	516,3
2004				468	*67,9*	*75,1*	*84,1*	*87,1*	*238,4*	552,5
2005				468	*72,6*	*80,3*	*90,0*	*93,2*	*255,0*	591,2
2006				468	*77,7*	*86,0*	*96,3*	*99,7*	*272,9*	632,5
Sum	**212,7**	257,4			* *Assumed Revenue Effects due to tax financing is 21%*					

All data

Year	Total Allocations	Adjusted Allocations *	N	Cumulative N	Assumed discount rate 7% Future values for different starting poins - Mill. NOK 1990	1991	1992	1993	1994	Sum
1990	329,7	398,9	312	312	*398,9*					398,9
1991	200,5	242,6	320	632	*426,9*	*242,6*				669,5
1992	248,5	300,7	325	957	*456,7*	*259,6*	*300,7*			1017,1
1993	344,3	416,6	469	1426	*488,7*	*277,8*	*321,8*	*416,6*		1504,9
1994	399,4	483,3	604	2030	*522,9*	*297,2*	*344,3*	*445,7*	*483,3*	2093,5
1995				2030	*559,5*	*318,1*	*368,4*	*476,9*	*517,1*	2240,1
1996				2030	*598,7*	*340,3*	*394,2*	*510,3*	*553,3*	2396,9
1997				2030	*640,6*	*364,1*	*421,8*	*546,0*	*592,1*	2564,7
1998				2030	*685,4*	*389,6*	*451,3*	*584,3*	*633,5*	2744,2
1999				2030	*733,4*	*416,9*	*482,9*	*625,2*	*677,9*	2936,3
2000				2030	*784,8*	*446,1*	*516,7*	*668,9*	*725,3*	3141,8
2001				2030	*839,7*	*477,3*	*552,9*	*715,8*	*776,1*	3361,7
2002				2030	*898,5*	*510,7*	*591,6*	*765,9*	*830,4*	3597,1
2003				2030	*961,4*	*546,5*	*633,0*	*819,5*	*888,5*	3848,9
2004				2030	*1028,7*	*584,7*	*677,3*	*876,8*	*950,7*	4118,3
2005				2030	*1100,7*	*625,7*	*724,7*	*938,2*	*1017,3*	4406,6
2006				2030	*1177,7*	*669,5*	*775,5*	*1003,9*	*1088,5*	4715,0
Sum	**1522,5**	1842,2			* *Assumed Revenue Effects due to tax financing is 21%*					

Table D7 Future values of initial endowments for combined Venture Loans and Investment Grants – Analyzed data – Mill. NOK

					Regional Venture Capital Loans Part																		Investment Grant Part							
					No Loss							30 % loss over 10 years							30 % loss over 5 years							No Loss					
	Total	Ad-justed		Cum-ula-	Assumed discount rate 7% Future values for different starting poins - Mill. NOK							Discount rate 7% - 3%=4% Future values for different starting poins - Mill. NOK							Discount rate 7% - 6%=1% Future values for different starting poins - Mill. NOK							Total	Ad-justed	Assumed discount rate 7% Future values for different starting poins - Mill. NOK					
Year	Alloca-tions	Alloca-tions *	N	tive N	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	Alloca-tions	Alloca-tions *	*1990*	*1991*	*1992*	*1993*	*1994*	Sum
1990	99,0	119,8	121	121	*119,8*					119,8	*119,8*					119,8	*119,8*					119,8	212,69	257,4	*257,4*					257,4
1991	82,0	99,3	111	232	*128,2*	*99,3*				227,5	*124,6*	*99,3*				223,9	*121,0*	*99,3*				220,3	61,42	74,3	*275,4*	*74,3*				349,7
1992	52,4	63,4	71	303	*137,2*	*106,2*	*63,4*			306,8	*129,6*	*103,2*	*63,4*			296,3	*122,2*	*100,3*	*63,4*			285,9	60,94	73,7	*294,6*	*79,5*	*73,7*			447,9
1993	21,4	25,9	34	337	*146,8*	*113,7*	*67,9*	*25,9*		354,2	*134,8*	*107,4*	*66,0*	*25,9*		334,0	*123,5*	*101,3*	*64,1*	*25,9*		314,7	36,53	44,2	*315,3*	*85,1*	*78,9*	*44,2*		523,5
1994	23,0	27,8	36	373	*157,1*	*121,6*	*72,6*	*27,7*	*27,8*	406,8	*140,2*	*111,7*	*68,6*	*26,9*	*27,8*	375,2	*124,7*	*102,3*	*64,7*	*26,1*	*27,8*	345,6	13,07	15,8	*337,3*	*91,0*	*84,4*	*47,3*	*15,8*	575,9
1995				373	*168,1*	*130,1*	*77,7*	*29,6*	*29,8*	435,3	*145,8*	*116,1*	*71,3*	*28,0*	*28,9*	390,2	*125,9*	*103,3*	*65,4*	*26,4*	*28,1*	349,1			*361,0*	*97,4*	*90,3*	*50,6*	*16,9*	616,2
1996				373	*179,8*	*139,2*	*83,1*	*31,7*	*31,9*	465,8	*151,6*	*120,8*	*74,2*	*29,1*	*30,1*	405,8	*127,2*	*104,3*	*66,0*	*26,7*	*28,4*	352,6			*386,2*	*104,2*	*96,7*	*54,1*	*18,1*	659,4
1997				373	*192,4*	*149,0*	*89,0*	*33,9*	*34,1*	498,4	*157,7*	*125,6*	*77,2*	*30,3*	*31,3*	422,0	*128,5*	*105,4*	*66,7*	*26,9*	*28,7*	356,1			*413,3*	*111,5*	*103,4*	*57,9*	*19,4*	705,5
1998				373	*205,9*	*159,4*	*95,2*	*36,3*	*36,5*	533,3	*164,0*	*130,6*	*80,3*	*31,5*	*32,6*	438,9	*129,8*	*106,4*	*67,3*	*27,2*	*29,0*	359,7			*442,2*	*119,3*	*110,7*	*62,0*	*20,7*	754,9
1999				373	*220,3*	*170,6*	*101,9*	*38,8*	*39,0*	570,6	*170,6*	*135,9*	*83,5*	*32,7*	*33,9*	456,5	*131,1*	*107,5*	*68,0*	*27,5*	*29,2*	363,3			*473,1*	*127,7*	*118,4*	*66,3*	*22,2*	807,8
2000				373	*235,7*	*182,5*	*109,0*	*41,6*	*41,8*	610,5	*177,4*	*141,3*	*86,8*	*34,1*	*35,2*	474,7	*132,4*	*108,6*	*68,7*	*27,7*	*29,5*	366,9			*506,3*	*136,6*	*126,7*	*71,0*	*23,7*	864,3
2001				373	*252,2*	*195,3*	*116,6*	*44,5*	*44,7*	653,3	*184,5*	*146,9*	*90,3*	*35,4*	*36,6*	493,7	*133,7*	*109,7*	*69,4*	*28,0*	*29,8*	370,6			*541,7*	*146,2*	*135,6*	*75,9*	*25,4*	924,8
2002				373	*269,9*	*208,9*	*124,8*	*47,6*	*47,8*	699,0	*191,8*	*152,8*	*93,9*	*36,8*	*38,1*	513,5	*135,0*	*110,8*	*70,1*	*28,3*	*30,1*	374,3			*579,6*	*156,4*	*145,1*	*81,3*	*27,2*	989,5
2003				373	*288,8*	*223,6*	*133,5*	*50,9*	*51,2*	747,9	*199,5*	*158,9*	*97,6*	*38,3*	*39,6*	534,0	*136,4*	*111,9*	*70,8*	*28,6*	*30,4*	378,0			*620,2*	*167,4*	*155,2*	*87,0*	*29,1*	1058,8
2004				373	*309,0*	*239,2*	*142,9*	*54,5*	*54,7*	800,3	*207,5*	*165,3*	*101,6*	*39,8*	*41,2*	555,4	*137,7*	*113,0*	*71,5*	*28,9*	*30,7*	381,8			*663,6*	*179,1*	*166,1*	*93,0*	*31,1*	1132,9
2005				373	*330,6*	*256,0*	*152,9*	*58,3*	*58,6*	856,3	*215,8*	*171,9*	*105,6*	*41,4*	*42,8*	577,6	*139,1*	*114,1*	*72,2*	*29,2*	*31,0*	385,6			*710,1*	*191,6*	*177,7*	*99,5*	*33,3*	1212,2
2006				373	*353,7*	*273,9*	*163,6*	*62,4*	*62,7*	916,2	*224,4*	*178,8*	*109,8*	*43,1*	*44,6*	600,7	*140,5*	*115,2*	*72,9*	*29,5*	*31,4*	389,5			*759,8*	*205,0*	*190,1*	*106,5*	*35,6*	1297,1
Sum	**277,9**	336,2		** Assumed Revenue Effects due to tax financing is 21%*																			**384,6**	456,4						

Table D8 Future values of initial endowments for combined Venture Loans and Investment Grants – Analyzed data – Mill. NOK

					Regional Venture Capital Loans Part																		Investment Grant Part							
					No Loss						30 % loss over 10 years						30 % loss over 5 years							No Loss						
					Assumed discount rate 7%						Discount rate 7% - 3%=4%						Discount rate 7% - 6%=1%							Assumed discount rate 7%						
	Total	Ad-justed		Cum-ula-	Future values for different						Future values for different						Future values for different						Total	Ad-justed	Future values for different					
	Alloca-	Alloca-		tive	starting poins - Mill. NOK						starting poins - Mill. NOK						starting poins - Mill. NOK						Alloca-	Alloca-	starting poins - Mill. NOK					
Year	tions	tions *	N	N	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	*1990*	*1991*	*1992*	*1993*	*1994*	Sum	tions	tions *	*1990*	*1991*	*1992*	*1993*	*1994*	Sum
1990	312,0	377,5	359	359	*377,5*					377,5	*377,5*					377,5	*377,5*					377,5	229,2	277,4	*277,4*					277,4
1991	269,9	326,5	328	687	*404,0*	*326,5*				730,5	*392,6*	*326,5*				719,2	*381,3*	*326,5*				707,9	196,1	237,3	*296,8*	*237,3*				534,1
1992	267,6	323,8	269	956	*432,2*	*349,4*	*323,8*			1105,5	*408,3*	*339,6*	*323,8*			1071,8	*385,1*	*329,8*	*323,8*			1038,8	172,0	208,1	*317,6*	*253,9*	*208,1*			779,6
1993	134,9	163,2	186	1142	*462,5*	*373,9*	*346,5*	*163,2*		1346,1	*424,7*	*353,2*	*336,8*	*163,2*		1277,9	*389,0*	*333,1*	*327,1*	*163,2*		1212,4	97,8	118,3	*339,8*	*271,7*	*222,6*	*118,3*		952,4
1994	129,6	156,9	112	1254	*494,9*	*400,0*	*370,8*	*174,7*	*156,9*	1597,2	*441,7*	*367,3*	*350,3*	*169,8*	*156,9*	1485,9	*392,9*	*336,4*	*330,3*	*164,9*	*156,9*	1381,4	106,5	128,8	*363,6*	*290,7*	*238,2*	*126,6*	*128,8*	1147,9
1995				1254	*529,5*	*428,0*	*396,7*	*186,9*	*167,8*	1709,0	*459,3*	*382,0*	*364,3*	*176,6*	*163,1*	1545,3	*396,8*	*339,8*	*333,6*	*166,5*	*158,4*	1395,2			*389,0*	*311,1*	*254,9*	*135,4*	*137,8*	1228,3
1996				1254	*566,6*	*458,0*	*424,5*	*200,0*	*179,6*	1828,6	*477,7*	*397,3*	*378,8*	*183,6*	*169,7*	1607,1	*400,8*	*343,2*	*337,0*	*168,2*	*160,0*	1409,1			*416,3*	*332,9*	*272,7*	*144,9*	*147,5*	1314,3
1997				1254	*606,2*	*490,1*	*454,2*	*214,0*	*192,2*	1956,6	*496,8*	*413,2*	*394,0*	*191,0*	*176,4*	1671,4	*404,8*	*346,6*	*340,4*	*169,9*	*161,6*	1423,2			*445,4*	*356,2*	*291,8*	*155,1*	*157,8*	1406,3
1998				1254	*648,7*	*524,4*	*486,0*	*229,0*	*205,6*	2093,6	*516,7*	*429,7*	*409,8*	*198,6*	*183,5*	1738,2	*408,8*	*350,1*	*343,8*	*171,6*	*163,2*	1437,5			*476,6*	*381,1*	*312,2*	*165,9*	*168,9*	1504,7
1999				1254	*694,1*	*561,1*	*520,0*	*245,0*	*220,0*	2240,1	*537,3*	*446,9*	*426,1*	*206,6*	*190,8*	1807,8	*412,9*	*353,6*	*347,2*	*173,3*	*164,9*	1451,8			*510,0*	*407,8*	*334,1*	*177,5*	*180,7*	1610,0
2000				1254	*742,7*	*600,3*	*556,4*	*262,1*	*235,4*	2396,9	*558,8*	*464,8*	*443,2*	*214,8*	*198,5*	1880,1	*417,0*	*357,1*	*350,7*	*175,0*	*166,5*	1466,3			*545,6*	*436,3*	*357,5*	*189,9*	*193,3*	1722,7
2001				1254	*794,6*	*642,4*	*595,4*	*280,5*	*251,9*	2564,7	*581,2*	*483,4*	*460,9*	*223,4*	*206,4*	1955,3	*421,2*	*360,7*	*354,2*	*176,8*	*168,2*	1481,0			*583,8*	*466,9*	*382,5*	*203,2*	*206,9*	1843,3
2002				1254	*850,3*	*687,3*	*637,0*	*300,1*	*269,5*	2744,2	*604,4*	*502,7*	*479,4*	*232,3*	*214,7*	2033,5	*425,4*	*364,3*	*357,7*	*178,5*	*169,8*	1495,8			*624,7*	*499,5*	*409,3*	*217,5*	*221,3*	1972,3
2003				1254	*909,8*	*735,4*	*681,6*	*321,1*	*288,4*	2936,3	*628,6*	*522,8*	*498,5*	*241,6*	*223,2*	2114,8	*429,7*	*368,0*	*361,3*	*180,3*	*171,5*	1510,8			*668,4*	*534,5*	*437,9*	*232,7*	*236,8*	2110,4
2004				1254	*973,5*	*786,9*	*729,3*	*343,6*	*308,6*	3141,9	*653,8*	*543,7*	*518,5*	*251,3*	*232,2*	2199,4	*434,0*	*371,6*	*364,9*	*182,1*	*173,3*	1525,9			*715,2*	*571,9*	*468,6*	*249,0*	*253,4*	2258,1
2005				1254	*1041,6*	*842,0*	*780,4*	*367,7*	*330,2*	3361,8	*679,9*	*565,5*	*539,2*	*261,4*	*241,5*	2287,4	*438,3*	*375,4*	*368,6*	*183,9*	*175,0*	1541,1			*765,3*	*612,0*	*501,4*	*266,4*	*271,1*	2416,2
2006				1254	*1114,5*	*900,9*	*835,0*	*393,4*	*353,3*	3597,1	*707,1*	*588,1*	*560,8*	*271,8*	*251,1*	2378,9	*442,7*	*379,1*	*372,2*	*185,8*	*176,7*	1556,6			*818,9*	*654,8*	*536,5*	*285,1*	*290,1*	2585,3
Sum	**1114,1**	1348,0		** Assumed Revenue Effects due to tax financing is 21%*																			**801,6**	969,9						

Figure D1 Interest rate 1990 - 2006

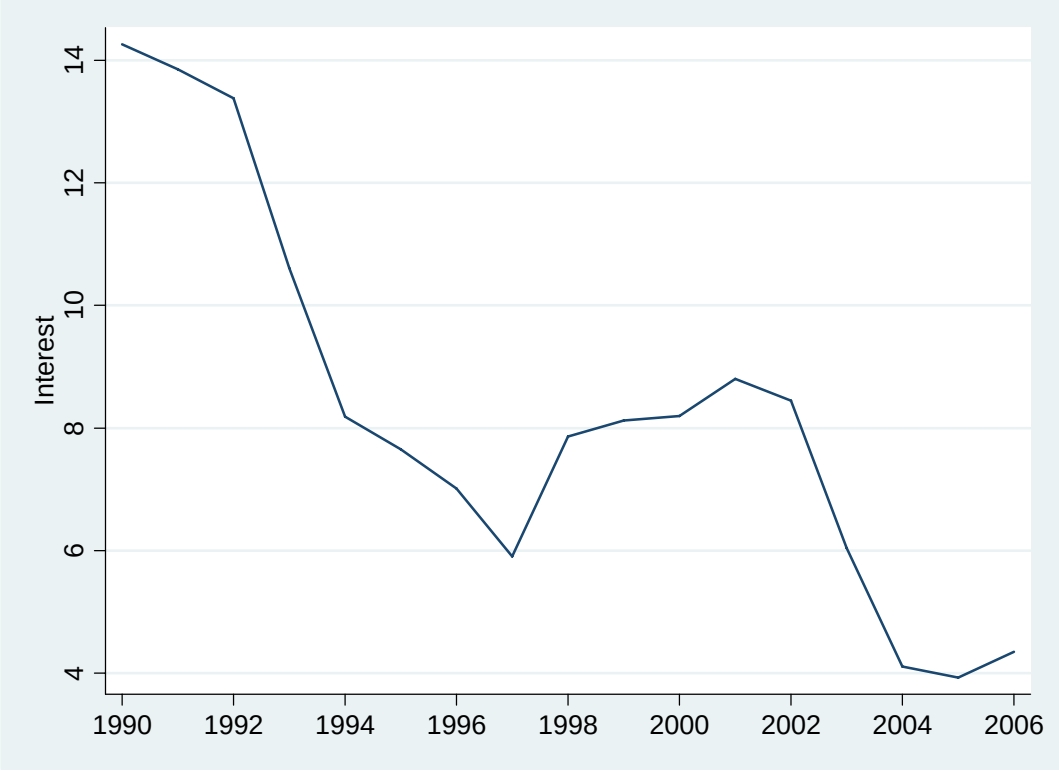


Table D9 Calculation methods for Two-period difference-in-differences – Missing by year versus missing by pair - 1000 NOK

Regional Venture Capital Loans

Gardner & Altman (missing by year)

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf.	Interval]	
1989	302	151	0.00		240.361	-473.01	473.01	*Before*
1990	338	187	134.27		242.180	-342.11	610.65	
1991	338	187	-75.96		226.004	-520.52	368.60	
1992	321	170	139.83		266.082	-383.67	663.32	*Intervention*
1993	311	160	576.53		293.880	-1.73	1154.79	
1994	309	158	959.87	*	344.172	282.64	1637.11	
1995	306	155	479.96		358.942	-226.36	1186.29	
1996	306	155	487.11		447.413	-393.31	1367.52	
1997	295	144	599.34		569.283	-521.06	1719.74	
1998	285	134	910.98		589.863	-250.10	2072.05	
1999	279	128	2037.20	*	520.143	1013.26	3061.13	
2000	275	124	1835.83	*	582.345	689.37	2982.29	*After*
2001	266	115	1359.01	*	648.526	82.07	2635.95	
2002	253	102	1356.54	*	637.585	100.85	2612.24	
2003	250	99	614.49		575.237	-518.48	1747.47	
2004	242	91	1284.53	*	641.128	21.58	2547.49	
2005	235	84	2001.74	*	598.097	823.38	3180.11	
2006	228	77	2183.59	*	725.685	753.62	3613.57	

Average 1989-2006 = *938.05*
Average 1995-2006 = *1262.53*

* *p≤.05*

Pairwise Comparisons (missing by pair by year)

Year	Obs.	**Estimate**		Std.Err.	% Conf.	Interval]	
1989	151	0.00		0.00	0.00	0.00	*Before*
1990	151	204.77		148.49	-88.64	498.19	
1991	145	-25.51		171.31	-364.12	313.10	
1992	133	136.43		228.73	-316.03	588.89	*Intervention*
1993	124	539.47	*	257.79	29.18	1049.75	
1994	123	834.15	*	381.78	78.39	1589.92	
1995	119	274.11		388.43	-495.08	1043.30	
1996	119	287.21		476.36	-656.11	1230.53	
1997	114	551.15		688.50	-812.89	1915.18	
1998	107	934.03		704.42	-462.54	2330.60	
1999	102	1955.67	*	693.12	580.70	3330.63	
2000	100	1734.67	*	793.38	160.44	3308.91	*After*
2001	93	1039.36		880.34	-709.07	2787.78	
2002	81	219.63		695.47	-1164.39	1603.65	
2003	79	25.65		638.59	-1245.69	1296.98	
2004	73	707.11		736.45	-760.99	2175.21	
2005	67	1101.24		635.68	-167.93	2370.41	
2006	61	1029.48		951.18	-873.17	2932.12	

Average 1989-2006 = *679.33*
Average 1995-2006 = *821.61*

Table D9 Continued…

Investment Grants

Gardner & Altman (missing by year)

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	424	212	0.00		303.253	-596.08	596.08	*Before*
1990	512	300	183.46		257.301	-322.04	688.96	
1991	549	337	112.18		254.361	-387.46	611.82	
1992	556	344	341.63		274.301	-197.17	880.43	*Intervention*
1993	576	364	733.69	*	305.894	132.88	1334.50	
1994	574	362	1023.92	*	315.773	403.70	1644.14	
1995	560	348	874.00	*	341.334	203.54	1544.45	
1996	564	352	936.95	*	433.616	85.25	1788.66	
1997	546	334	1081.25		575.458	-49.14	2211.65	
1998	523	311	1555.02	*	578.113	419.30	2690.74	
1999	507	295	1811.97	*	734.455	369.00	3254.93	
2000	487	275	2299.33	*	806.736	714.21	3884.46	*After*
2001	465	253	1579.54	*	788.704	29.66	3129.43	
2002	440	228	1015.54		795.883	-548.68	2579.77	
2003	420	208	1571.32	*	792.444	13.65	3128.99	
2004	407	195	1901.06	*	865.542	199.55	3602.58	
2005	394	182	1397.73		789.539	-154.53	2950.00	
2006	376	164	1340.86		919.702	-467.58	3149.29	

Average 1989-2006 = 1097.75
Average 1995-2006 = 1447.05

* *p≤.05*

Pairwise Comparisons (missing by pair by year)

Year	Obs.	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	212	0.00		0.00	0.00	0.00	*Before*
1990	207	330.42		188.31	-40.85	701.69	
1991	209	367.24		228.62	-83.47	817.96	
1992	196	257.81		261.24	-257.42	773.03	*Intervention*
1993	196	795.01	*	282.30	238.27	1351.75	
1994	195	1280.31	*	325.39	638.56	1922.07	
1995	193	1193.88	*	366.30	471.39	1916.36	
1996	195	541.07		485.40	-416.28	1498.42	
1997	188	398.43		675.40	-933.96	1730.81	
1998	179	1551.05	*	694.02	181.48	2920.62	
1999	173	1586.54		937.15	-263.26	3436.34	
2000	161	1466.23		965.25	-440.04	3372.50	*After*
2001	155	1123.95		1019.23	-889.53	3137.43	
2002	141	318.86		1005.18	-1668.44	2306.16	
2003	131	1015.83	*	1058.72	-1078.72	3110.39	
2004	123	1514.50	*	1188.69	-838.63	3867.63	
2005	117	1403.35		1161.20	-896.54	3703.25	
2006	106	2316.46		1179.30	-21.88	4654.80	

Average 1989-2006 = 1027.11
Average 1995-2006 = 1202.51

Table D9 Continued…

Both Regional Venture Capital Loans & Investment Grants Combined

Gardner & Altman (missing by year)

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	598	299	0.00		248.489	-488.02	488.02	*Before*
1990	673	374	438.27	*	220.915	4.50	872.04	
1991	690	391	596.30	*	227.598	149.43	1043.17	
1992	681	382	1049.85	*	247.518	563.86	1535.85	*Intervention*
1993	676	377	1277.25	*	274.103	739.05	1815.45	
1994	671	372	1814.07	*	309.561	1206.24	2421.90	
1995	666	367	2309.14	*	357.225	1607.71	3010.56	
1996	667	368	2013.32	*	424.515	1179.76	2846.87	
1997	641	342	2175.29	*	489.538	1213.99	3136.58	
1998	620	321	2760.49	*	569.185	1642.72	3878.26	
1999	593	294	2621.42	*	582.131	1478.12	3764.72	
2000	579	280	2942.37	*	693.571	1580.14	4304.60	*After*
2001	551	252	1972.08	*	656.114	683.28	3260.88	
2002	537	238	1522.45	*	646.891	251.70	2793.21	
2003	520	221	1279.49	*	641.793	18.66	2540.33	
2004	509	210	1420.75	*	685.671	73.64	2767.86	
2005	493	194	2113.10	*	661.747	812.90	3413.31	
2006	465	166	2444.49	*	766.326	938.58	3950.40	

Average 1989-2006 = 1708.34
Average 1995-2006 = 2131.20

* $p \leq .05$

Pairwise Comparisons (missing by pair by year)

Year	Obs.	**Estimate**		Std.Err.	% Conf. Interval	]	
1989	299	0.00		0.00	0.00	0.00	*Before*
1990	296	520.85	*	163.70	198.68	843.02	
1991	286	830.79	*	203.09	431.04	1230.54	
1992	268	1240.03	*	244.00	759.62	1720.43	*Intervention*
1993	259	1541.10	*	298.78	952.75	2129.46	
1994	255	2071.99	*	344.03	1394.48	2749.51	
1995	251	2424.55	*	400.47	1635.83	3213.28	
1996	252	2042.25	*	461.97	1132.42	2952.07	
1997	234	2003.50	*	596.06	829.15	3177.85	
1998	220	2735.07	*	716.10	1323.73	4146.41	
1999	202	2484.83	*	731.78	1041.89	3927.77	
2000	196	3083.07	*	876.33	1354.78	4811.37	*After*
2001	181	1713.79	*	800.06	135.08	3292.50	
2002	175	1166.39		818.61	-449.29	2782.07	
2003	165	961.39		863.08	-742.80	2665.57	
2004	157	822.77		960.83	-1075.14	2720.68	
2005	146	1727.30		973.93	-197.63	3652.22	
2006	124	2196.42		1235.73	-249.62	4642.46	

Average 1989-2006 = 1642.56
Average 1995-2006 = 1946.78

Table D9 Continued…

The FRAM Program

Gardner & Altman (missing by year)

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf. Interval	]	
1991	323	646	0.00		222.425	-436.77	436.77	*Before*
1992	349	672	127.77		231.808	-327.39	582.93	
1993	365	688	66.88		198.406	-322.67	456.44	
1994	374	697	358.46		191.654	-17.83	734.75	*Intervention*
1995	380	703	434.07	*	196.238	48.78	819.36	
1996	379	702	642.50	*	209.257	231.65	1053.35	
1997	369	692	675.69	*	254.292	176.40	1174.99	
1998	346	669	489.10		291.506	-83.29	1061.49	
1999	319	642	924.37	*	329.651	277.04	1571.70	
2000	297	620	871.71	*	376.273	132.78	1610.63	
2001	269	592	529.73		559.126	-568.39	1627.85	
2002	250	573	605.69		444.499	-267.36	1478.74	*After*
2003	233	556	518.75		489.980	-443.70	1481.19	
2004	220	543	836.97		534.318	-212.62	1886.56	
2005	206	529	568.49		598.684	-607.61	1744.59	
2006	188	511	443.39		657.031	-847.44	1734.21	

Average 1991-2006 = 505.85 * *p≤.05*
Average 1997-2006 = 643.13

Pairwise Comparisons (missing by pair by year)

Year	Obs.	**Estimate**		Std.Err.	% Conf. Interval	]	
1991	323	0.00		0.00	0.00	0.00	*Before*
1992	323	209.84		174.71	-133.89	553.56	
1993	323	157.68		120.22	-78.83	394.19	
1994	323	413.81	*	145.79	126.99	700.63	*Intervention*
1995	319	470.58	*	178.36	119.66	821.50	
1996	317	644.22	*	196.23	258.13	1030.31	
1997	309	535.98	*	240.18	63.37	1008.59	
1998	292	372.71		293.73	-205.39	950.82	
1999	274	593.54		330.94	-57.97	1245.05	
2000	260	495.29		398.92	-290.25	1280.82	
2001	235	-98.17		621.31	-1322.26	1125.91	
2002	221	205.01		506.95	-794.08	1204.11	*After*
2003	205	86.16		598.57	-1094.01	1266.33	
2004	193	564.25		663.85	-745.11	1873.62	
2005	182	107.07		773.81	-1419.78	1633.91	
2006	166	116.22		902.51	-1665.74	1898.19	

Average 1991-2006 = 324.95
Average 1997-2006 = 297.81

Table D9 Continued…

The Network Program

Gardner & Altman (missing by year)

Year	Pairs Obs.	Total Obs	**Estimate**		Std.Err.	% Conf.	Interval]	
1992	1364	2728	0.00		558.121	-1094.38	1094.38	*Before*
1993	1360	2724	458.18		582.239	-683.50	1599.86	
1994	1330	2694	830.25		621.703	-388.82	2049.31	*Intervention*
1995	1288	2652	1214.54		656.764	-73.28	2502.36	
1996	1245	2609	1847.26	*	698.991	476.63	3217.90	
1997	1160	2524	1930.82	*	694.840	568.30	3293.33	
1998	1082	2446	2232.43	*	744.199	773.10	3691.75	
1999	1000	2364	4008.54	*	906.567	2230.79	5786.29	
2000	958	2322	3636.47	*	971.249	1731.87	5541.08	
2001	899	2263	5611.40	*	989.927	3670.14	7552.66	*After*
2002	829	2193	4513.24	*	1031.067	2491.27	6535.22	
2003	772	2136	3397.61	*	1009.099	1418.69	5376.53	
2004	727	2091	3722.14	*	1043.948	1674.85	5769.42	
2005	670	2034	4058.36	*	1093.130	1914.59	6202.14	
2006	614	1978	4199.36	*	1150.487	1943.07	6455.66	

Average 1992-2006 = *2777.37* * *p≤.05*
Average 1997-2006 = *3731.04*

Pairwise Comparisons (missing by pair by year)

Year	Obs.	**Estimate**		Std.Err.	% Conf.	Interval]	
1992	1364	0.00		0.00	0.00	0.00	*Before*
1993	1360	463.73	*	234.52	3.67	923.79	
1994	1330	842.67	*	272.36	308.37	1376.97	*Intervention*
1995	1288	1205.26	*	363.36	492.41	1918.11	
1996	1245	1867.65	*	418.79	1046.04	2689.25	
1997	1160	2132.66	*	475.93	1198.87	3066.45	
1998	1082	2482.83	*	532.14	1438.70	3526.97	
1999	1000	4295.65	*	723.88	2875.16	5716.15	
2000	958	3823.20	*	892.78	2071.17	5575.24	
2001	899	5801.91	*	936.12	3964.68	7639.15	*After*
2002	829	4616.19	*	979.65	2693.30	6539.08	
2003	772	3394.90	*	1001.47	1428.97	5360.82	
2004	727	3516.42	*	1062.23	1431.01	5601.82	
2005	670	3871.75	*	1194.39	1526.55	6216.94	
2006	614	4139.43	*	1435.56	1320.22	6958.63	

Average 1992-2006 = *3032.45*
Average 1997-2006 = *3807.49*

Figure D2 Comparisons of calculation methods for Gardner & Altman versus Pairwise deletion

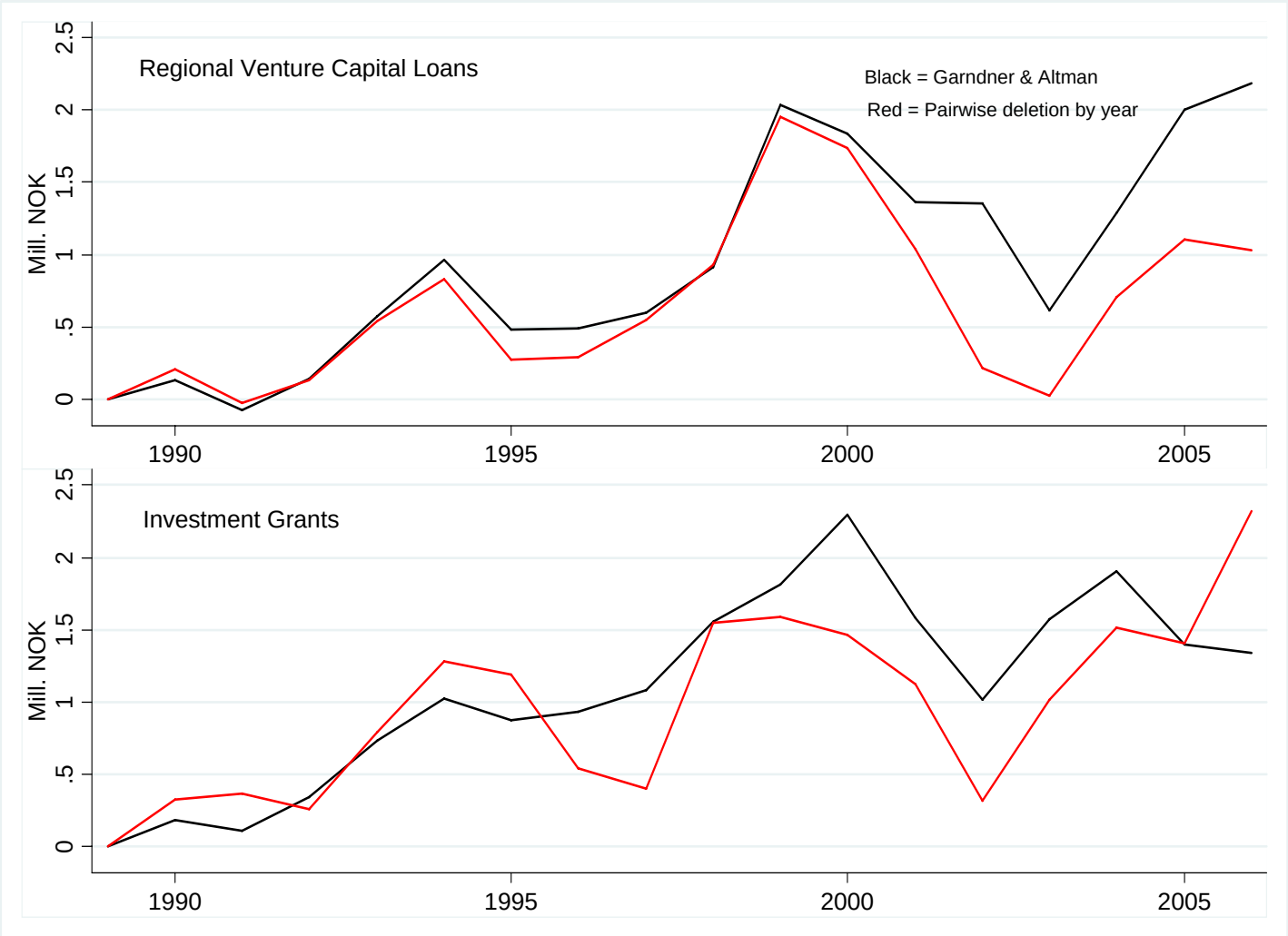


Figure D2 Comparisons of calculation methods for Gardner & Altman versus Pairwise deletion, *Continued*...

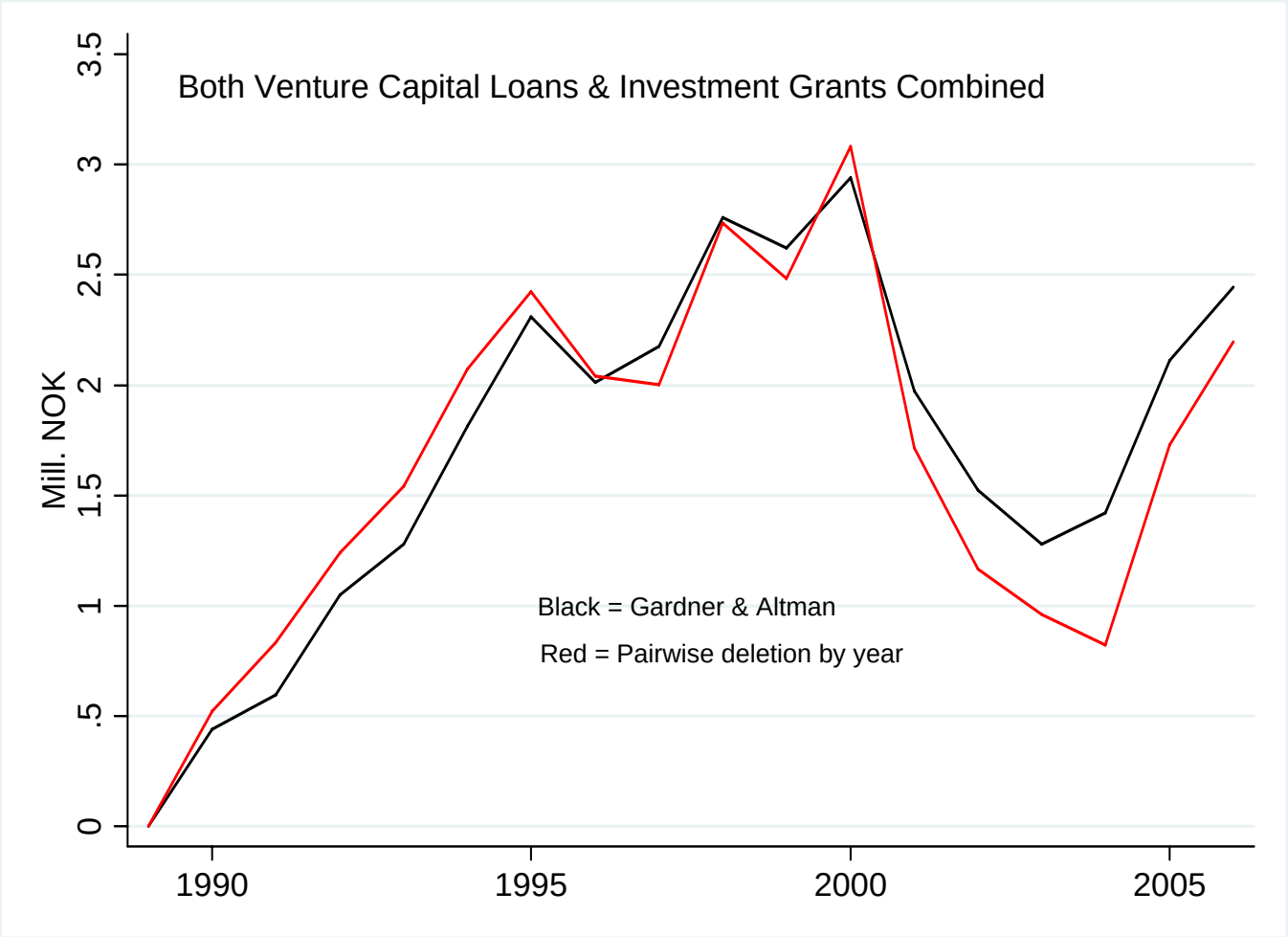
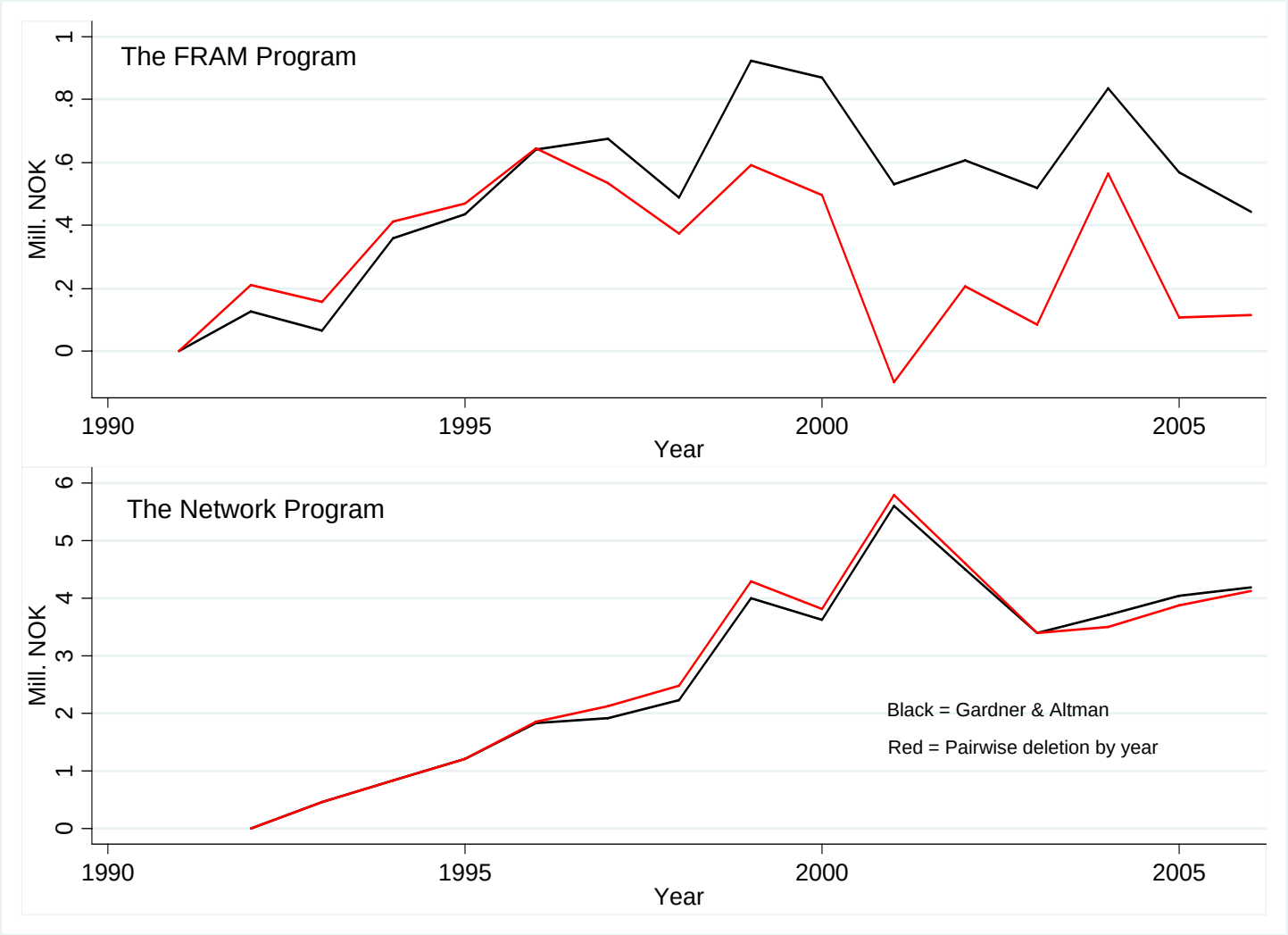


Figure D2 Comparisons of calculation methods for Gardner & Altman versus Pairwise deletion, *Continued…*



NHH

Norges Handelshøyskole

Norwegian School of Economics and Business Administration

NHH
Helleveien 30
NO-5045 Bergen
Norway

Tlf/Tel: +47 55 95 90 00
Faks/Fax: +47 55 95 91 00
nhh.postmottak@nhh.no
www.nhh.no

NHH, the Norwegian School of Economics and Business Administration, is one of the leading business schools in Scandinavia. Over 2,700 students are enrolled in our various programmes. We have recently reorganised our programmes so that they follow a 3+2+3 year sequence, leading to the awarding of Bachelor of Science, Master of Science and Ph.D. degrees respectively.

NHH has a long reputation for its high academic level and contributions to the international research community. A large number of our faculty hold a Ph.D. from institutions outside of Norway, in particular top US universities. This creates a diverse and stimulating academic environment.

The Ph.D. student body is made up of around 100 men and women working within different specialisations. The programme encourages close interaction between students and faculty in a social/academic climate where students are regarded as junior colleagues.

The Ph.D. programme offers courses over a wide range of topics within Accounting, Economics, Finance, Management Science and Strategy and Management. The programme aims at giving the graduate a solid training in performing high quality scientific research in these areas, making use of state of the art empirical and theoretical techniques. This prepares the student for employment in national and international policy institutions, within research centres, business enterprises, and for the international academic job market. The entire programme is taught in English and runs over three years, with the first year consisting primarily of course work. The next two years are then devoted to independent research and the writing of a doctoral thesis, under the supervision of an advisor appointed from the NHH faculty.

Trykk: Allkopi AS, Bergen 55 55 30 55 • www.allkopi.no