17/05/16

How to Think About Validity in Effect Studies

- A case study of the effect-evidence ranking scheme of the National Board of Social Services

Master Thesis
Joachim Skanderby Johansen
MSc in Business Administration and Philosophy

Supervision by Morten Sørensen Thaning
Number of characters / pages: 181.424 / 80

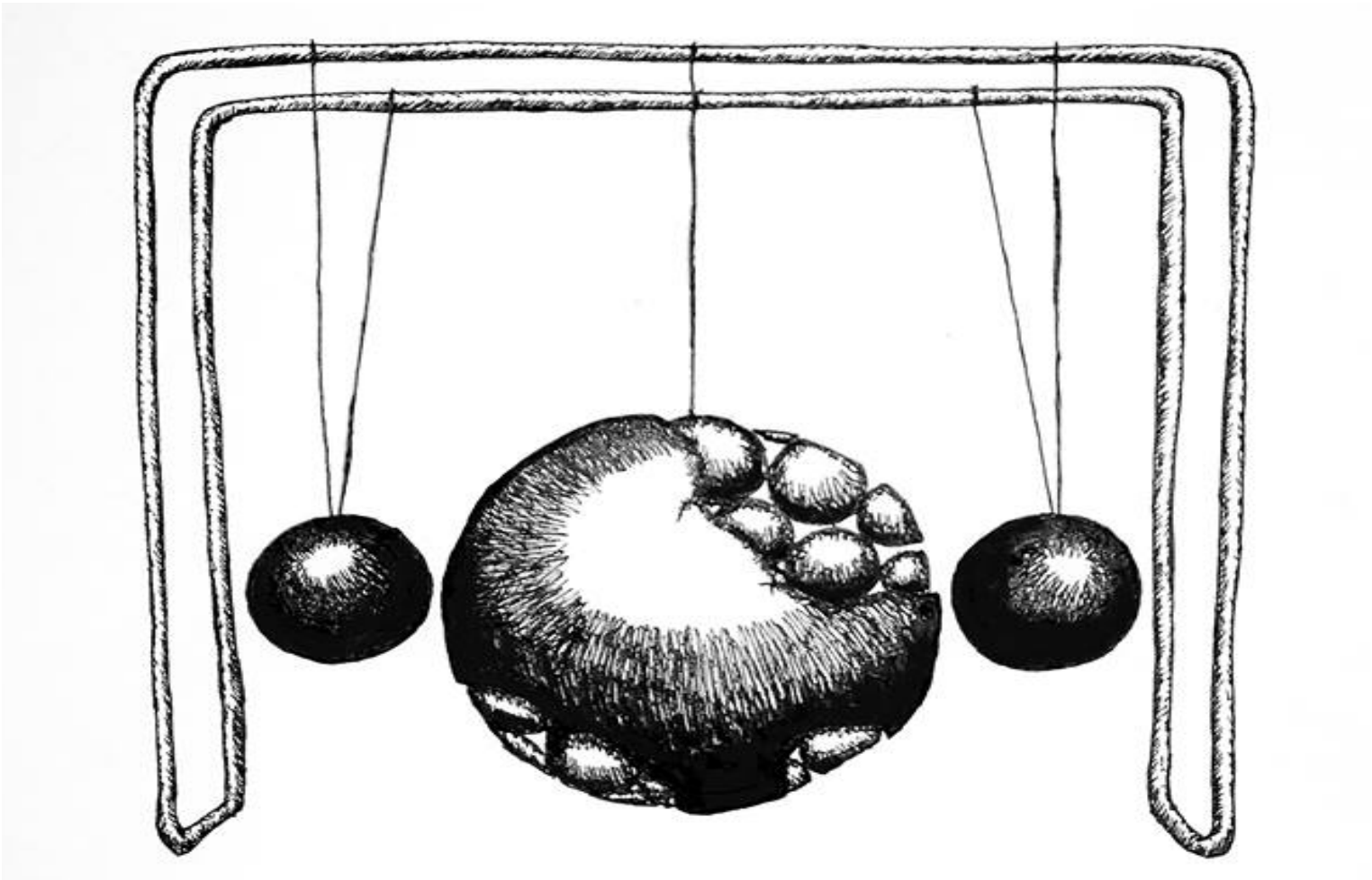


Illustration by Jonas Aahave Uhd

Abstract

This thesis evaluates the effect-evidence ranking scheme used by The National Board of Social Services (NBSS) seen in light of discussions in the philosophy of science on internal and external validity.

I start by introducing the effect-evidence scheme and its context. I then argue that we can best understand the effect-evidence ranking as an expression of how certain we are, based on conducted studies, that a social method will be effective when we employ it. Next, I analyse two central types of studies that the effect-evidence scheme ranks – variance (e.g. cohort studies, RCTs etc.) and process type studies - in regards to the possible threats to their internal and external validity.

In terms of internal validity, I argue that confounders pose a threat for both variance and process studies, and discuss when which type will do better. In terms of external validity, I argue that both types of studies face a threat from lack of support factors in the target we want to extrapolate the result of the study to, although we often have better knowledge of support factors in process studies. I also analyse the external validity threat posed by different mechanisms being present in the population that we have studied and in the target we want to extrapolate to. Here, I will discuss strategies to mitigate this threat and why we should differ between processes and mechanisms. Lastly, I consider the threat to both internal and external validity of multiple causal paths, arguing that this is a relevant concern for both variance and process studies.

In the conclusion, I argue that the NBSS effect-evidence ranking scheme does not provide a good guide for effect-evidence ranking and I outline an alternative scheme which ranks studies through probability assessments of three parameters; confounders, support factors and causal paths.

Keywords: Internal validity, External Validity, Evidence, Mechanisms, Processes.

Table of Contents

1. Introduction

In the Danish public service, the National Board of Social Services (NBSS) has the important task of assessing what evidence of effect exists for various social interventions. The social interventions, or social methods as they are also called, range widely; from group meetings for families with vulnerable children to outreach work with gang members. The assessments of the NBSS are made to be used by municipalities and other social policy intuitions in order to determine which social methods to deploy. Thus, the way that the NBSS goes about these assessments have a potentially big impact on what social methods are used – underlining the importance of the assessments being made the right way.

To give a good assessment of whether social methods are effective, the NBSS have developed guidelines for what constitutes good effect-evidence. A central element of the guidelines is a ranking scheme. This scheme determines how high an effect assessment, or ranking, a method can obtain, depending on the type of study that has been made to measure its effect. Thus, a method can never get more than a certain ranking (which spans from A to D) unless a certain type of study has been made. The two highest ranked types of studies are studies with a variance oriented understanding of causality and studies with a process oriented understanding of causality. The scheme dictates that if a social method has been tested by a study with a variance oriented understanding of causality, this allows for the highest ranking (A), while a social method that has been tested by a study with a process oriented understanding of causality can only achieve the second-highest ranking (B).

The goal of this thesis is to evaluate whether this effect-evidence ranking scheme provides a good guideline for assessing whether a social method will work. More specifically, it poses the question of whether it is justified that variance type studies allow for the highest effect-evidence ranking, while process type studies only allow for the second-highest.

This evaluation will be done from a philosophy of science perspective. Recent work in the philosophy of science has paid much attention to the fact that, for a study to work as evidence for the effectiveness of a social method or policy, it must not only have internal validity but also external validity. While internal validity denotes that the causal conclusion made in a study is

correct, external validity denotes that this conclusion will also hold outside of the study. If we want to use studies to underpin that a social method will work elsewhere than where it has been studied, we need both internal and external validity. This is almost always the situation municipalities and other users of the NBSS's effect ranking will find themselves in.

Thus, the strategy of the thesis will be to evaluate how these two types of studies fair in terms of internal and external validity. More specifically, I will look at what threats to validity have been identified in the philosophy of science literature, and discuss how these apply to the two types of studies. By this discussion, it is also the goal of the thesis to make a theoretical contribution to discussion in the philosophy of science on internal and external validity.

In light of the discussion of threats to internal and external validity of the two types of studies, the conclusion of the thesis will comment on whether the hierarchy in the scheme is justified and, finding that it is not so, I will outline an alternative ranking scheme.

The thesis will progress in the following way:

Section 1, *Introduction,* is the current section, also containing a sub-section of the scope of the thesis below.

Section 2, *The Danish Policy For The Use of Evidence*, will describe the context in which the effect-evidence ranking scheme appears. More specifically, I will analyse the content of the two documents produced by the NBSS through which the scheme is made public. Furthermore, I will discuss what exactly it is that the scheme ranks and how we may conceptualize this.

Section 3, *Different Needs to Know*, argues that what we want the studies on social method to supply us with are prediction. It is furthermore discussed what kinds of predictions we are interested in.

Section 4*, External and Internal validity – The concepts and their usefulness,* presents the concepts of internal and external validity through analysing how they are discussed in the philosophy of science literature.

Section 5, *Defining the Difference between Variance and Process studies,* analyses what defines these two types of studies.

Section 6, *Internal Validity Threats to Variance Studies,* discusses the threat of confounders and the degree to which these can be mitigated in variance type studies. Special attention is given to the prominent type of variance studies, the random control trial.

Section 7, *External Validity Threats to Variance studies,* discusses the concepts of support factors and mechanisms and how these work to secure external validity. Special attention is given to a criterion for external validity presented by Steel (2008). Process studies will be indirectly discussed here as Steel draws on these to make his argument. My main theoretical contribution to the external validity discussion is presented here, where I argue that we ought to differ between processes and mechanisms.

Section 8, *Internal Validity Threats to Process Studies,* discusses how the threat of confounders also looms of process type studies. The section also includes an analyses of a study on the effect of a social method in order to compare how the threat of confounders occurs for process and variance studies.

Section 9, *External Validity Threats to Process studies,* discusses how support factors relate to process studies, and argue these are also needed here in order to secure external validity.

Section 10, *The Threats Aggregated and Conclusive Remarks*, summarizes the discussion on the different threats to validity and evaluates the effect-evidence ranking scheme is this light. An alternative effect-evidence ranking scheme is presented.

Lastly, a note on terminology:

I will use the terms *object of study* to denote the population that has been considered in a study. Thus, if the study shows that X causes Y, it shows that this is the case for the object of study.

I will use term *target* to denote the population that we wish to extrapolate our finding in a study to. Thus, if we have found that that X causes Y in a study, and we wish to extrapolate this finding, we wish to extrapolate it to a certain target.

I will use *treatment variable* to denote what we are testing the effect of in a study (thus, a social method) and *effect variable* to denote what this social method is supposed to affect.

The Scope of the Thesis

Given the theme of this thesis, the range of relevant perspectives and indeed entire academic disciplines that could have contributed to enlighten the issue is substantial. As described in the introduction, I aim to analyze and assess the difference in terms of validity between variance and process studies with some of the latest thinking in philosophy of science, with NBSS ranking scheme as the case. Given this scope, I here wish to shortly discuss some of the perspectives on the theme that I have not employed.

First and foremost, this thesis does not go into depth with the statistics proper of process or variance studies. I bring this in to light where it seems relevant for the methodological or philosophical points I wish to make (such as the discussion on RCTs in a later section) but beyond this, it is not discussed. I have chosen to do so for two reasons: First, a thesis that would have considered both the philosophy of science and the statistics part of the theme would have to be significantly more extensive; both in terms of time and length. Secondly, being trained in philosophy, this is where I can make the most valuable contribution to the discussion.

Secondly, I do not go much into the academic field of social work or social policies, though the thesis is concerned with how social methods are ranked. I do make an effort to show the relation between the discussions in the thesis and the social methods that are ranked by the NBSS. However, the analyses of the validity of variance and process studies is independent of the concrete social methods– it is a general analysis of the two types of studies with the NBSS ranking scheme as its 'case study'. I think that there are most certainly some very interesting insights to be gained by going much deeper into the specifics of social work and social policy. However, in deciding whether focusing on these specifics and the more general philosophy of science analysis, I chose the

former. A benefit of this is that due to the generality of the analysis, the point brought up in the thesis should be applicable beyond the area of social work and policy.

2. The Danish Policy for the Use of Evidence

The National Board of Social Services has presented two documents regarding evidence-based methods[1]; "Viden til Gavn - Politik for udvikling og anvendelse af evidens" (VP) and "Vidensdeklaration - Socialstyrelsens Vidensdeklaration af sociale indsatser og metoder"[2] (VS). As I explain in the introduction, the motivation for looking into the questions of validity is the effect-evidence ranking scheme by the NBSS. This ranking is found in these two documents and I consider it important for the thesis to explore them to understand how the ranking is meant to be used and the context in which it appears. As a help for the reader, I should mention that it will not be extremely important for the rest of the thesis to remember which paper is the VS and which is the VP – they are both introduced with the intention of giving context to the ranking scheme.

The VS

The VS was published in 2012; a year before the VP. Also, the VP use the VS as a reference point. Thus, I consider it natural to start with a description of the central elements in VS before going over the VP. I will be analysing the document called VS, but there is also the 'Vidensdekleration' database over the different methods that the NBSS have evaluated (I will use VS whenever I refer to the document). I will not refer much to this database here, but I will be drawing on some cases from this database in the thesis. In the VS, methods are defined as, "en indsats der følger en særlig systematik"[3] (Socialstyrelsen, 2012, p.3). Thus, methods are concrete social interventions. For example, U18, a method that has been evaluated in the Vidensdeklaration database, consists of a series of meetings, therapy seasons and talks with young people and their families where the young people involved have a problem with substance abuse. Methods should therefore not be understood

[1] Plus at least one other document that treats the subject indirectly; see "Anbefalinger til samfundsøkonomisk evaluaring på socialområdet" (2015)
[2] Own translation: 'Utilizable Knowledge – Policy for the Development and Application of Evidence' and 'The Knowledge Declaration – the National Board of Services Services' Declaration of Interventions and Methods.'
[3] Own translation: 'An intervention that follows a certain system.'

in the more abstract sense of the word; as guiding principles for some (scientific) inquiry, but as concrete social policy interventions.

The VS document is a description of the guiding principles by which each method is evaluated in the database. The VS document's goals is thus, "at vejlede om vidensgrundlaget for indsats og metoder på det sociale området." (Socialstyrelsen, 2012, p.3)[4] The VS describes how the Vidensdeklaration database rank the different social methods along five different dimensions; targeted group, method, implementation, effect and economy.[5] Along these dimensions, each method can be rated from D to A, with A being the top score. The scores show the NBSS's evaluation of how much is known about the method, along each dimension. The majority the VS document consists of the sub-questions that the NBSS answers when they evaluate which rank to ascribe to method, within each dimension. This thesis will be concern with the effect dimension. This dimension is concerned with what we know about the effect of the social methods with regards to the problem it is supposed to mitigate. Under the effect dimension, we find the evidence ranking scheme below. It is this scheme that the thesis will specifically be concerned with.

[4] Own translation:'To give advice regarding the knowledge base for the interventions and methods in the social policy domain.'
[5] Translated from: 'målgruppe, metode, implementing, effekt og økonomi.'

Effektviden	Design og metoder	Beslutnings-kontekst	Hvem kan vurdere	Højest mulige vurdering
Specialiserede effektstudier	Fx Metareview, eksperimenter, statistisk analyse (variansorienteret effektforståelse)	1) Udrulning/ ikke udrulning 2) Anbefaling om, at alle kommuner benytter metoden	Socialstyrelsen evt. i samarbejde med relevant forskningsinst. (fx SFI)	A
Procesorienterede effektstudier (Kausale mekanismer)	Fx Teoribaseret evaluering, proces tracing, contribution analysis (procesorienteret effektforståelse)	Kommuner bør vurdere metoden sammenholdt med eksisterende praksis og træffe et metodeansvarligt valg	Socialstyrelsen	B
Før- og eftermåling (ikke-eksperimentelle)	Fx Før og eftermåling, simpel monitorering af centrale variable i indsatsens forandrings-teori (uden kausalanalyse)	Kommuner bør lære af velbeskrevet praksis fra andre kommuner	Socialstyrelsen	C
Ekspert-vurdering	En eller flere eksperter inden for metoden vurderer indsatsen har positiv effekt	Kommuner bør lære af velbeskrevet praksis fra andre kommuner	Socialstyrelsen	C
Eftermålinger	Kvalitativ eller kvantitativ måling af fx brugeres og fagprofessionelles vurdering af effekt	Kommuner bør lære af velbeskrevet praksis fra andre kommuner	Socialstyrelsen	C

Figure 1, (Socialstyrelsen, 2012, p.16)

As can be seen in the upper right corner, the type of study done determines the highest rank that the method can obtain in the 'effect' dimension. The VS explains that a method may have multiple types of evidence to underpin it, but the highest ranked type of evidence determines the highest rank the method can obtain. Thus, a method underpinned by 'process oriented effect studies' can never score more than a B, but may score less. This is also the reason I choose to analyse the scheme in particular – it has an overwriting function by putting a cap on the rank that the methods can obtain. The focus of this thesis will thus on whether it is always the case that process studies are inferior to variance studies, in the way that this scheme would indicate. The difference between these types of

studies will be analysed in the Defining the Difference Between Variance and Process Type Studies section later.

It is important to mention that the effect-evidence ranking also, in addition to the evidence-ranking scheme, include some questions that the NBSS look at when giving an effect-evidence rank. Thus, the scheme is not the only thing the NBSS looks at. This is also why the lowest score on the evidence-ranking scheme is a C while the lowest score possible is a D. I will however consider it justified to focus on the evidence ranking scheme as this, as explained, sets a cap for how high a score a study can receive, no matter other questions the NBSS pose in the ranking all in all.

The VS also clearly states that its ranking is supposed to help agents in the social area utilize methods that are well supported the right kind of knowledge, within the five different dimensions. In practice, the most central agents are the municipalities. Thus, the effect-evidence dimension, including the evidence ranking scheme, is supposed to give the municipalities knowledge that they can use in deciding what social methods to employ.

The VP

The VP also includes the evidence-ranking scheme but puts it in a slightly different context. The VP is meant as a strategy for the application and development of evidence-based methods. Thus, instead of describing the principles by which the individual methods in the Vidensdeklaration database are evaluated, the VP describes the principles that should guide how to test methods used and what kind of evidence should be developed.[6] Though there seem to be some overlap in the intention of the two documents, I think a good way of differing between the VS and the VP is that the VS is concerned with evaluating our knowledge of the methods we currently have, while the VP is concerned with setting up some principle testing new methods. The VS is retrospective, while the VP is prospective. It is worth noting that the directions in the VP are also supposed to determine what kinds of research receive funding.

[6] The paper is very clear on what this means in practice, but it could be steering the research of institutions like SFI (www.sfi.dk), which is the National Research Center for Welfare.

The VP divides the theme of the paper into three subcategories: Creation of new methods, developing existing methods and testing of methods. The document declares that its purpose is to inform decision-makers in social policy, whether these are placed in the municipalities or in the state. Thus, like the VS, it puts the ranking of the social methods in a decision-making context. The main body of the VP is the formulation of eight steps to guide the testing and development of social methods. Not all of the steps are supposed to be followed for every method test or development project, even though greenfield methods creation project should start at step one. For every testing or method development project, there should be an effect evaluation, however.

It is in the third phase that we find 'effect evaluation' where the evidence-ranking scheme is introduced. Under this phase, it is stated that only variance oriented studies and process studies are suitable for the purpose of evaluating new methods; before-and-after measurements are only seen as indications in the preliminary trial of the method. Also expert assessments are not deemed suitable for testing methods. Actually, the VP explicitly advises *against* investing in expert judgements.

For the theoretical focus of this thesis it is important to mention that the VP explicitly states that the methods that are developed should be used *generally* across the municipalities, independently of local conditions and special priorities. This motivates my research question of not only the internal validity of the methods ranked in the evidence ranking scheme, but also the external validity - how applicable evidence obtain by these methods are to domain outside of the study.

Besides the use and development of evidence-based methods, the VP's guidelines are supposed to cover three other, less well-described, categories, which I mention for context but that I will not go into depth with. These are "Projekttyper med klare forandringsmål", "Andre projekttyper" and "Aktivitetsforøgelse, driftsbevilling, puljer med brede, vagt formulerede formål"[7]. For the first of the three other categories, the purpose of following the VP is to use it as a guideline for performance management. For the second category the goal of the VP is to guide possible evaluations of these other project types, if this is required. For the last category, the VP is supposed to be used to guide administrative follow-ups.

[7] Own translation: 'Project types with clear, intented goals', 'Other projecttypes' and 'Increased focus, operational funding, funds with vague, broadly defined purposes'.

To summarise, the goal of the VS and VP sections was to introduce the evidence-ranking scheme and to present the context that it appears in; as part of the overall effect-evidence ranking. I have explained how the evidence-ranking scheme is not the only thing the NBSS takes into consideration when ranking a social method, but as it 'caps' what rank a method might receive, I think this justifies focusing on the scheme. Also, we have seen that the scheme is both supposed to be used to evaluate what effect-evidence rank the methods that are evaluated in the Vidensdeklaration database receive, but also to guide how to test methods in the future. Finally, we saw that both documents put the ranking of the methods in a decision-making context; it is supposed to inform decision-makers. In practice these might well be municipalities that need to decide which methods to use.

I should mention that I have focus rather narrowly on the parts of the VP and the VS that deals with the *effect*-evidence ranking. The reason I have not included the other dimensions (e.g. implementation, method etc.) for context is that 1), it provides me with more space for the actual analysis and discussion and 2) the ranking on the different dimensions are supposed to be read and used by the municipalities and other public employees. These readers can rightfully expect that all consideration relevant for effect are put under the 'effect' rank. Therefore, I will expect the same.

A Comment on Objectivity

There is a curious statement in the VS that I will comment on before moving on. It is not specifically connected to the effect dimension but to the VS in general. On page 7, it is stated that the ranking is based on a professional assessment of the methods, not on objective criteria. The paper continuous by stating that the evaluation of the methods is not to be considered scientific assessments. However, in case they are not supposed to be interpreted as scientific and objective, it is hard to see what authority the ranking is supposed to have in guiding what social methods that municipalities are supposed to use, or what studies we use to test methods in future. We may think that as experts, the employees of NBSS are worth taking advice from, but we should only do so if their judgments are based on proper and sound principles. Therefore, I will treat the ranking of the NBSS to be at least as 'scientific' and 'objective' as it would need to be underpin its function and authority. Also, even if we accept that current rankings are not to be taken as a scientific or objective, it certainly seems worth consider whether they *could* be said to be so, and if not, what a ranking that is made with some ambition of being 'scientific' (I take it that the NBSS by this means

systematic or principled) would look like. Therefore, I do not consider it a misunderstanding to attempt analyze whether the ranking of the NBSS is a justified in the light of state of the art philosophy of science. Furthermore, I speculate that the disclaimer in the VS is made more as a cover against (political) criticism than as something that is supposed to be kept in mind when actually looking over the rankings. This would also explain the apparent disconnection between making a guide in order to secure that social methods are well underpinned by scientific research, and then state the very guide itself is not objective or scientific. This explanation of the disclaimer as cover against criticism rather than an integrated principle in the guide, justifies an analysis of the ranking based on the assumption that it is a serious attempt at formulating objective guiding principles, even if the NBSS will not admit that this is what they are trying to do.

Analysing the Effect Dimension

Now that we have introduced the evidence ranking scheme and its context, I think it is necessary to conceptualise exactly what it is the ranking in the effect-evidence dimension is supposed to reflect. The section will thus clarify some of the vagueness in the formulations of the VS and VP on this. In the VS it is stated that;

"Effektdimensionen afdækker viden om hvorvidt, I hvilken grad og hvordan indsatsen/ metoden har ført til de ønskede effekter, herunder også om indsatsen eller metoden har haft uforudsete konsekvenser…" (a)[8] (Socialstyrelsen, 2012, p.14)

"Forudsætningen for en høj score på effektdimensionen er, at en indsats/metode er tilstrækkeligt velbeskrevet til, at den egner sig til udrulning i kommunerne…" (b)[9] (Socialstyrelsen, 2012, p.14)

The first sentence (a) states what the effect-dimension is supposed to describe. It seems to concern knowledge of whether the method has led to the effect that it supposed to, how it has done so and to what degree. When reading this, two interpretations seems possible; firstly, that the dimension is

[8] Own translation: 'The effect dimension is concerned with knowledge about whether, to what degree, and how the intervention/method has led to the desired effects, also covering whether the intervention or method has had any unwanted effects'
[9] Own translation: 'The requirement for a high score on the effect dimension is that the intervention/method is sufficiently well-described for use in the municipalities.'

concerned with rating the actual effect of the method and its magnitude. Then we would suppose that a high score was given to methods that studies have shown to be very effective. However, the first sentence (a) actually reads that the dimension regards the *knowledge* of these aspects of the method's effect. This would lead us to the second interpretation which is that the effect dimension is concerned how well we *know* the effect, not how large or small the effect is. This interpretation is supported by the second sentence, where it is stated that a high score on the effect dimension requires that the method is sufficiently *well described*. This second interpretation would, taken on face value, suggest that a method with a low but very well described effect could score high on the evidence ranking, which would run counter to the first interpretation. This second interpretation is further supported by what the NBSS writes in the VP on the effect dimension. In the VP, the NBSS writes:

> *"Højest udsagnskraft og stærkest belæg i forhold til national metodeudvikling har veludførte kvantitative analyser… der bygger på sammenligning mellem effekten for borgeren og den forandring, der (teoretisk set) ville være sket, hvis borgeren ikke havde fået indsatsen."*[10]
> *(Socialstyrelsen, 2013, p.12)*

However, what does the word "udsagnskraft" mean? It does not figure in the Danish Dictionary, but is used by some academics such as Busck (2008) in connection with evaluating sources for historical accounts. Here, the word is taken to denote how much weight one should give a particular historical source. It is contrasted with "udsagnsevne", which is defined as the relevance the source has for the subject.[11] Other academics use the term the exact opposite way though; Ankersborg (2007) use udsagnskraft to denote the relevance of the source and explicitly states that this is disconnected from how much we should trust the source.

While being interesting in itself that the NBSS uses a concept that seems to be so little consensus on, I believe that the most reasonable interpretation of what the NBSS mean by "udsagnkraft" is how Busch et al. uses by the word. It would be odd that NBSS should use the word in the sense that

[10] Own translation: 'Quantitative studies, which build on comparing the effect for the citizen of the method and the change that (theoretically speaking) would have happened if the citizen had not received the intervention, gives the highest certainty and the strongest justification in relation to the national development of methods'

[11] A lying investment banker may therefore not have a lot of "udsagskraft", according to Busck et al., but could have a great deal of "udsagsevne", if the subject was the moral fiber of investment bankers.

Ankersborg uses it, where it has to do with relevance. One would think that all studies regarding a method surely must be relevant in the sense of trying to say something about that particular method. Therefore, I think that "udsagnskraft", in the context of the VP comes down to how much weight we can put on the claims of some study. This seems to fit well with what the second interpretation of the description on the effect dimension in the VS states; that the effect ranking concerns how much we know about the effect and not the magnitude of the effect.

As noted, this would seem to imply that a study which has a *negative* effect on the citizens could receive a high effect rank, as long as this bad effect is underpinned by really good studies. However, in practice the NBSS of course only conveys methods that have a positive effect on the citizens. I will therefore assume that methods that receive high effect-evidence ranks are those where we have high quality knowledge of the positive effects of the method.

A Question of Certainty

Given this clarification; that the ranking of effect-evidence is concerned with the quality of knowledge we have about the effect, rather than its magnitude, what concept would seem to capture this? I believe that when inquiring about the weight we should we should put on an effect study, it seems plausible that we are asking *how certain* we are that some effect will occur. This would also make sense when we think about what the effect-evidence ranking is supposed to be used for; as guidance for the municipalities when deciding what methods to use and what methods to develop.

Thus, the question then turns out to be: how we should think about certainty and, in continuation, *uncertainty*. The most prominent way of doing this is in terms of *probabilities.* This is also what I will do in this thesis, mainly in the conclusive part. I should mention that probabilities are not the only way to deal with uncertainty. Other approaches include uses of fuzzy logic or other logical approaches to uncertainty (Colyvan, 2008). Therefore, I will shortly argue why I use probabilities as a way of modelling uncertainty, to underpin my decision.

First, classical probability theory is simpler and more intuitive than for example fuzzy logic. This is a merit, as the last part of the thesis will be concerned with setting up an alternative evidence-ranking scheme that the employees of the NBSS can use.

Secondly, as mentioned before, the ranking is supposed to be used to guide decision-making, and by conceptualizing uncertainty[12] through classical probability theory, we could draw on the arsenal of Decision Theory. The usual way of modelling a rational decision maker in decision theory is as a utility maximize with a defined risk-preference (also called risk attitude). An expected utility maximizer should always (trivially), if facing a set of prospect choices, choose the prospect that has the high expected utility, where this is simply calculated multiplying the utility of the prospect with the *probability* of that prospect. I will set aside how we should define the utility function of the employees that choose what social method to use. The takeaway here is that by conceptualizing uncertainty in terms of probabilities, we would allow for utility calculations to be done[13]. It may seem absurd that the people working with social policy in the municipalities would do utility calculations on each prospective method to decide which to use. However, since funds and time are limited, they must do something like this, even if do so in a more informal manner currently. Also, since the effect ranking of studies is also supposed to decide how to test and develop new methods, and the Ministry of Finance is an interested part on such funding issues[14] , I do not consider it improbable that *they* would try to do something like a utility calculation. Due to this, I consider it an advantage to think about this question in terms of probabilities.

A final note on probabilities; the interpretation of what probabilities are is no trivial matter at all. One may distinguish between two interpretations of probability; on one side the objectivist, or frequentist interpretation and on the other side the subjectivist (Morgan, Henrion, & Small, 1990; Resnik, 1987). The debate on over subjective vs. objective probabilities is tied to the debate between classical frequentist statistics vs. Bayesian methods of inference (see (Mayo, 1996)) and is both extensive and complicated. Therefore, I will not consider this further. However, for this thesis, I will here subscribe to the subjective notion of probability. According to this, probability is an expression of a given person's belief in how likely it is that some outcome will occur. This also means that the different probabilities can be ascribed to the same event by different persons and that

[12] In the terminology of decision theory, the question is actually what be what kind of *risky* is important for the decision-maker. When we can assign probabilities to the outcomes of some action, the agent in question is facing a risky choice, whereas uncertainty is reserved for choices where we assign no probabilities to the outcomes.
[13] This would of course require that any model of the uncertainty of a study would have to take in all the sources of uncertainty. I would not claim such completeness for the model I present in the final section.
[14] I know this from a personal conversation with a friend of mine working at the Ministry of Finance.

the same person can ascribe different probabilities to an event if there are changes in that person's level of information. This can lead to some quite radical subjectivist notions, but it will be beyond the scope of the thesis to discuss them in depth. The subjectivist interpretation of probabilities seems advantageous in this thesis as the alternative evidence ranking scheme that I present in the end will exactly ask the NBSS employs to ascribe probabilities to different risk factors for validity. I am not certain that these risk factors could be said the to be expressions of frequencies, like the objectivist notion would demand.

To summarise this section, I have interpreted the purpose of the evidence ranking as being a measurement of *certainty;* a measurement of how certain we are about some causal conclusion of the study. I have then briefly discussed my choice to model uncertainty as probabilities. Lastly, I have mentioned that I will interpret probabilities in the subjectivist sense, though I cannot go much into the discussion regarding this.

3. Different Needs to Know

Another matter that we need to clarify in order to discuss what kinds of studies have the most certain conclusions is what conclusions we are looking for when it comes to social methods. I will argue that what we are looking for to be valid, is a special type of predictions; predictions that 1) shows that some policy will make a positive *contribution* and 2) prediction about what this policy will do in isolation, not what other factors will lead to.

First, I will shortly discuss the notion of a positive contribution. The idea here is that we may have a multitude of degrees of precisions in our predictions (indeed, indefinitely many). Say, for example, that I have evidence from a study of 129 institutions in four municipalities, showing that the VIDA programme has a statically significant positive impact on emotional and behaviour problems (two of the five parameters that the programme takes into consideration). When we wish to use this study to be predict what will happen if we introduce the VIDA programme in some municipality, we may want to know the *precise magnitude* with which the programme will improve the emotional and behavioural problems of the children under consideration. Or we may be interested in knowing the magnitude of the effect within some interval. The assumption in this thesis is that we will often be

content with knowing that the policy has some positive contribution, i.e. that the impact of the programme will be a positive one. Thus, I will assume, we will be happy to use social methods as long as there are valid arguments that it will contribute positively. In our example from before, we would apply the VIDA programme if we knew that would be *some* positive impact on emotional and behaviour problems, without being certain that the *magnitude* of this impact would be the same as in the VIDA studies. Of course, we could think up examples were using social methods based merely on knowledge that they will have *some* positive contribution will misguide us; for example, if the positive impact is very small, and the social method very expensive. However, I think the criteria of knowledge of positive contribution strikes the right balance between being informative and yet feasible. Of course we would prefer studies were we could extrapolate the precise magnitude with which a programme will affect some variable of interest. However, in the social realm, this hardly seems feasible.

Secondly, I wish to discuss not the precision with which we want predict the outcome of social methods, but exactly what it is that we wish to predict. I will distinguish between kinds of predictions in this regard. The first type of inference that I wish to describe is what I will call a future state prediction. When we talk about making a prediction in common conversion, this is usually what we mean. Next, I will describe a ceteris paribus prediction. Such a prediction deals with the effect of a particular casual factor, but is silent on the behaviour of the rest of causal factors in our target.

A future state prediction regards the actual state of affairs in our target at some point in the future; it regards what will be the outcome of all the causes at work at some point in the future. The prediction that macro economists and that meteorologists would seek to make are good examples of what I consider future state predictions; both macro economists and meteorologists try to take all the important causal factor within their domain into consideration and then draw some conclusion about future state of affairs. These predictions are indeed important, but when it comes to social policy we can often do with less. We often just wish to know the effect of *one* causal factor, namely the one that we wish to manipulate or introduce in our target. Imagine that we implement the VIDA programme in our target and we know that the studies done on VIDA have external (and internal, for that matter) validity. What we then wish to predict is what the effect of the VIDA programme is

in itself, isolated from other causal factors. Thus, we wish to know that the VIDA programme, ceteris paribus, will reduce emotional and behavioural problems.

Of course, like with the positive outcome criteria from before, we may sometimes make wrong decisions if we only base our decisions on ceteris paribus predictions, where we only take the effect of one factor (or social methods, in this case) into consideration. First of all, we may doubt that we can make any law-like ceteris paribus predictions – will the effect of one factor not always depend on the configuration of other factors? Later in the analysis, we will however see that we can take this into consideration by the notion of support factors. Secondly, and more unredeemable, just like the criteria before of a positive contribution, we can sometimes make bad decisions when we only base our considerations upon a ceteris paribus consideration. For example, if war breaks of out and as a result, the SQD goes through the roof as a result of this, we may think that money is better spend on other initiatives than the VIDA programme, even if it may raise the SQD slightly.

To summarise, I have specified what causal conclusions we want to be able to validly draw from our studies; ceteris paribus predictions regarding positive contributions. This will be the assumption throughout the rest of thesis.

4. External and Internal Validity – The Concepts and Their Usefulness

In discussing the degrees of certainty with which we can believe in some causal conclusion, I will use two methodological concepts, namely internal validity (IV) and external validity (EV). The exact definitions of internal and external validity vary slightly in the literature. However, usually internal validity is defined along the following lines: a study is said to have internal validity in case the causal conclusions it draws (i.e. what causes what) *in the study* are valid (Cook, 1979; Guala, 2003; Jiménez-Buedo, 2010). Thus, if the conclusion in a study is that smoking causes cancer, then the study has internal validity if this was indeed the for the test population *in the study* (or the object of study, as I will refer to it by).

External validity, on the other hand, refers to how generalizable the causal relationship in study is across other domains (Cook, 1979; Guala, 2003; Jiménez-Buedo, 2010) or in other words, how valid it is to extrapolate from some study or experiment (Steel, 2008). Thus, a study on smoking has external validity if its conclusion (say, that smoking causes cancer) can validly be extrapolated to domains outside of the object of study. External validity is therefore not a property of the study itself, but of the relation between the object of study and a target that we wish to extrapolate to. Thus, a study may have high external validly in relation to one target, but poor external validity to another. I will discuss the threats to external validity later in the thesis.

There are some discussions on how exactly to define the two terms, and the definitions can vary according to the account of causality that the author subscribes to ((Cook, 1979) have some interesting remarks on this). In this thesis, the two concept will serve to denote two distinct criteria that the studies of the social methods should fulfil in order to be useful: they should both have internal validity by showing that in the study, the treatment actually was the cause of the effect that was measured (and did not spuriously correlate with the effect), and have external validity in that the causal relationship that they expose in the study should be generalizable to municipalities where the social methods are supposed to be used. I believe the best way to understand these concepts is to take on some of the most common debates concerning internal and external validity. Also, these debates are relevant for the theme of this thesis. Firstly, I will discuss the often-heard claim that there is a trade-off between IV and EV and secondly, how IV seems to be a requirement for EV.

Both the question of IV as a requirement for EV and the trade-off between the two are discussed by Jiménez-Buedo & Miller (2010) and my discussion will centre on this paper. I have chosen this because Jiménez-Buedo & Miller's paper brings together the discussions of some of the most prominent writers on this field in philosophy of science, such as Nancy Cartwright, Francesco Guala, and Donald Campbell.

First, why is IV a requirement for EV? IV is ensured when it can justify that the causal factor under study is actually the cause of the effect of interest; it ensures the causal conclusion of the study is valid. The claim that has been advanced by many methodologists (Cook, 1979; Guala, 2003; Thye, 2000) is basically that it is very difficult to extrapolate the causal conclusion of the study if we do

not know what this causal conclusion is! Therefore, if we wish to have EV we must first have IV; to be able to justify extrapolation of a causal claim, we must first know what the claim is.

Overall, I agree with this line of though, however, I do have two comments, one that is purely of principle and one which will be important for the later discussion. Firstly, I think it is still possible to draw some conclusion about a domain outside of the domain of the study (thus, external to that domain) from the data obtained in the study, even if this data does not ensure any sound causal conclusion within the study. I will illustrate this with a scenario. Imagine that a study has shown the correlation of A and B but lacks the proper control for spurious correlation to conclude that A causes B, within some domain, D1. Thus, the study has no internal validity. Now imagine that within some other domain, D2, we know that either X or Q (and no other causal factors) can cause Y. Thus, if Y is present (or is 'true', if Y is a sentence) we know that either X or Q must be present (or 'true'). Now imagine that we have a theory which states that if there is correlation between A and B in the domain of the study, D1, then Q cannot be present in D2. Also imagine that we know that Y is present. Then we can conclude that it was X that caused Y, and not Q. In this case, we use non- causal knowledge from a study (which thus does not have internal validity) to make a causal conclusion in another domain.

It is still true that we cannot say that we extrapolate or generalize some the causal conclusion in the study (there is none in this case), and thus, the thought-up example above does not show that IV is not needed for EV. It does however show that IV in a study is not necessary for the study to give causal information about some domain external to the study. However, in practice I believe that situations like the one above are very rare, if ever occurring and that we would usually need to establish a causal connection within the study in order for us to extrapolate to other domain. My argument is therefore more one of principle than of practical importance.

My second comment is that we need to be more precise regarding how much IV is demanded for EV, as we never actually achieve complete IV. As I will argue later in the thesis, a central problem with regards to IV is unknown confounders; the fact that we can never be completely certain (I will discuss this with relation to RCTs) that what we take is be a causal relation between A and B is not really the product of an unknown confounder. This means that we can never have complete IV. To know how much IV is required for EV, I think of this in terms of uncertainty and probabilities, as I

have argued above. If we do not have IV, we run the risk of introducing a causal factor that does not have the effect in our target that we wish. How much IV we wish for EV thus depends on how willing we are to accept this risk of introducing a causal factor with no effect. Thus, it essentially comes down to the risk-willingness of the agent and utility that lies in manipulating the causal factor. I do not believe that the notion that IV comes in the degrees in new, it is also hinted at in Campbell & Cook's (1979) definition of the term of IV, where IV "refers to the *approximate* (my italics) validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause." The notion of approximation points at IV as being a matter of degree.

Thus to summarise, for practical purposes I find that internal validity is indeed needed before we can have external validity. As Guala (2003) puts it: "Problems of internal validity are chronologically and epistemically antecedent to problems of external validity." However, we should note that IV comes in degrees and thus, in practice it will depend on how risk willing we are before we dare assert that we have internal validity. We should also be aware that if we want to extrapolate the result of the study, internal validity may be prior to external validity, but if we do not have external validity, we still cannot extrapolate the conclusion of the study. Thus, the fact that IV is prior to EV should not be taken to mean that one is more important than the other.

Next, I want to address the alleged trade-off between IV and EV. Jiménez-Buedo (2010) discusses this thoroughly and the analysis leads them to a critical stance on the trade-off relationship. The writers investigate a standard argument for the trade-off, namely that it stems from the artificiality of laboratory experiments, which can improve IV but losses EV in the process. On the other hand, the argument goes, in field experiments which are seen as more natural, we have more EV, but less IV. This is because field experiments are more akin to the real world that we wish to extrapolate our results to. However, it is difficult for us to be sure that he we have correctly identified a causal connection in the study as we cannot control the behaviour of other causal factors that might disturb our study. These we can control in the laboratory, but as a laboratory is not very much like the real world, this seems to impair EV. Jiménez-Buedo (2010) finds this type of argument to be present in the texts of several writers on the topic. However, they argue that artificiality is not a very well-defined concept and thus seek to find a more rigid definition within the literature. They do not find a definition of artificiality that would justify that laboratory experiments (the artificial environment)

would have more IV, and less EV and that the opposite should be the case with field experiment. Therefore, they grow sceptical of the trade-off all together: "We adopt a critical stance to the standard position on this debate by showing that problems of either external or internal validity do not necessarily nor crucially depend neither on the artificiality of experimental settings nor on the laboratory-field distinction between experiments… there seems to be no grounds to posit a general trade-off between the internal and external validity of experiments." (Jiménez-Buedo, 2010).

I agree with Jiménez-Buedo (2010) that there is no *neccesacy* link between artificiality and the proposed trade-off. The *mere* fact that a study is done in a labratory does not rob it of external validity. However, I disagree with Jiménez-Buedo (2010) when suggesting that this is enough to disprove that there is a trade-off of some kind. Therefore I will now explore this issue further.

Jiménez-Buedo (2010) actually mentions a possible, systematic trade-off, even though they seem to have omitted it when concluding that they are sceptical of the trade-off; "Depending on the degree of background knowledge of the researcher about the interaction between the variables of interest and the variables describing the degree of familiarity between agents, (which is the cause we are looking into in the example used by Jiménez-Buedo, *red*) the strategy of fixing the level of a background variables can indeed potentially cause difficulties…" Jiménez-Buedo (2010, p.12). What Jiménez-Buedo gets at here is connected to what we discussed in the previous section, namely the difference between future state predictions and ceteris paribus predictions. If we wish to make prediction regarding a future state, we ought to take all the fators into consideration (or at least all that have a non-neglitable influence). In some cases it may be easier for us to just describe the behavior of one of them, for example by isolating that causal factor from other causal factors in a laboratory. If this is easier than looking at all the factors at once, this isloation may increase the internal validy of our experiment (i.e. of our study), but will do so at the expense of the external validity – we are ommiting all the other causal factors that are at work in the real world. Thus, though I think that the authors are right in that *just* because an experiment happens in a labrotory, it does not loose internal validity. However, we instead that what we might *seek to do* (isolating) in the laboratory may lead to this trade-off.

Thus, there may actually be a systematic trade-off between internal validity and external validity when making future state predictions which Jiménez-Buedo (2010) does not pay sufficiently

attention to when doubting a systematic trade-off. However, as discussed in the Different Needs to Know section, when we dealing with social methods, we are not interested in the future state predictions but in ceteris paribus prediction. Here, it is a good thing that we only look at one factor in isolation –this is just what we want to do. For example, if wish to combat malnutrition, we may be interested in the effect of proving education and information to the exposed population. In this case we would not be interested in what else might affect malnutrition but whether education could contribute to that development, whatever the level of malnutrition would end up being. Therefore, the trade-off does not seem a danger in our case.

This theme of isolation has been discussed quite thoroughly in the philosophy of economics literature (Alexandrova, 2008; N. Cartwright, 2007a; Hausman, Kahane, & Tidman, 2013; Mäki, 2012). Cartwright has made the same argument as the one I make above in relation to economics (see for example (N. Cartwright, 2007b); that we are often interested in ceteris paribus predictions. Yet, Cartwright stills holds that there can be a trade-off, even though we may only seek to describe one causal factor in isolation. The question is then, why is this? Thus, in this last part of this section, I will discuss whether Cartwright provides us with any reasons to suppose that there is a trade-off that *will* be dangerous for the ceteris paribus prediction that we seek to make.

First, I need to briefly introduce some terminology that the argument rest on. Cartwright uses the term Galilean idealizations to denote the assumptions that has the isolating function in model or experiment. The inspiration for this term comes from Galileo's famous rolling ball experiment. Here Galileo tried to exclude the friction of the plane when modelling the decent of the ball. This is obviously unrealistic in that friction certainty exists. However, we still claim that Galileo's experiment where valuable because they supply us with knowledge of one factor, and one factor in isolation, namely gravity. This is why Cartwright chooses the term for idealizations that has the aim of isolating one cause.

In the social realm, a Galilean idealization might thus be the assumption of absence of government in an economic model, or only focusing on how a social method affects children's welfare, ignoring other factors that might influence this. In both cases, the purposes is as just discussed – we wish to isolate the effect of just a single causal factor, shielded off from the influence of other variables. Thus, that a model or a study has Galilean assumption is not a bad thing for our case.

However, writes Cartwright: “What I fear is that in a general a good number of false assumptions made with our theoretical models may not have the form of Galilean idealisations…” and “The model specific assumptions can provide a way to secure deductively-validated results where universal principles are scare. But these create their own problems. For the validity of the conclusion appears now to depend on a large number of special interconnected assumptions” (N. Cartwright, 2007a, p.12).

Thus, Cartwright’s fear is this that we will need assumptions of concrete empirical relations in order to secure our conclusion, rather than just the Galilean idealization that specify the none-interference causal factors that we are not interested in. If this is the case, then IV will come at the cost of EV; the assumptions regarding concrete empirical facts will make the model or study yield a clear causal conclusion, but these concrete empirical facts may well not be present in the target that we want to extrapolate the result of our object of study to. In other words, the conclusions of the study will not hold outside the context the study. Thus, the IV bought by these assumptions indeed does come at the cost of EV!

This is the reason Cartwright claims that there is a systematic trade-off between IV and EV. In my view, (Jiménez-Buedo, 2010) does not pays sufficient attention to this in their article. They quote Cartwright for stating that the usual reason for the trade-off is taken to be the artificiality. They argue convincingly that artificiality, if taken to refer to something like laboratories vs. field experiments, does not itself induce the trade-off (though isolation of causal factors will, if we want a future state prediction), but if artificiality it is taken to mean that we rely on very specific assumptions in our study that are artificial compared to the target of our extrapolation, there *does* indeed seem to be a trade-off.

To summarize the trade-off discussion, I have used the Jiménez-Buedo (2010) paper as a vantage point. They argue convincingly that *just* because a study is done in a laboratory rather than in the field we do not gain IV or lose EV. However, we have discussed how there does seem to be a trade-off between IV and EV, in case the model or experiment isolates the effect of one factor and in case this isolation allows us to draw more valid causal conclusions *within* the study. However, this is *only* a trade-off if our0 aim is to predict some future state and not the contribution of a single

variable. In social policy, the latter is the case. Lastly, we have discussed how there may still be a trade-off between IV and EV, even in models and studies that only seek to describe the contrition of one factor. This is, as Cartwright describes, the case if the result of our study is dependent on assumptions that will not hold outside the domain and that do not function as Galilean isolations. This is the trade-off that seems problematic for the validity of social methods. In the section on External Validity Threats for Variance Studies, I will show that we can take a potential trade-off into consideration through consider what has been coined support factors.

To take stock of the progress, at this stage we have looked into the evidence-ranking scheme and its context, I have discussed how exactly to understand what the effect dimension is supposed to rank and argued that this the certainty, which I wish to conceptualize as the probability of a method having a given effect in the target. In other words, the probability of the study being both internally and externally valid. We have now discussed whether there is a trade-off between internal and external validity. This is important for assessing the ranking of the NBSS; we want to analyse it in terms of internal and external validity, and if a gain in type of validity can lead to loss in the other, we surely want our ranking the reflect this. What we need to do now is get a better understanding of variance and process studies, before discussing their merits and weaknesses in terms of internal and external validity. In the last section, I will conclude on this discussion to see if lends credence to the ranking scheme of the NBSS.

5. Defining the Difference Between Variance and Process Type Studies

Of central importance is the definition of two terms used in the evidence ranking scheme, namely the difference between what is called “variansorienteret effektforståelse” and “procesorienteret effektforståelse”, that is, variation oriented understanding of effects and process oriented understanding of effects. For brevity, I shall simply call these variance studies and process studies. It is these two types of studies that I will focus on, when discussing whether the evidence ranking scheme of the NBSS is correct. As mention in the section before, the scheme dictates that a method where there has been made a process study at most get a score of B, while a method that has been testing in a variance study can at most get an A, the highest rank.

From correspondence with the Deputy Head of the NBSS, Mr. Carsten Strømbæk Pedersen, I have learned that the distinction stems from a paper that he wrote for what is now called the Agency for Modernisation, within the resort of the Ministry of Finance. In order to discuss the evidence-ranking scheme, I will now analyse the paper written by Pedersen to get a clear understanding of these two types of studies.

In the paper, (Pedersen, 2010) classifies the variance studies as studies that focuses on whether a variance in a variable of interest is due to a variance in some other variable. Pedersen wishes to classify studies that use some element of control as a variance study. I explain further under the Internal Validity Threats to Variance Studies section, but 'control' essentially means that we seek to make sure that other causal factors do not disturb our attempt to measure the effect of some variable of interest. This control, as Pedersen mentions, can be in the form of randomization and control group. Pedersen also mentions statistical control, where a control variable is included in the analysis so as to isolate the effect of the variable under study. He describes how we cannot be certain that we have controlled for all relevant variables, but that theory and background knowledge should allow us to identify at least some of the relevant variables.

A variance study might look the following way: We may wish to measure the effect of Tripe P, a programme that is designed to help parents tackle the emotional and behavioural problems of their children. One way to do this is to test the programme on a sample of families, and then compare the emotional and behavioural problems of the test group (or 'treatment group') with the the control group's's. To properly function as a 'control', the control group should have had the same exposure to other causal factors that might affect the emotional and behavioural problems of the children. If this is the case, the difference between the group of families that went through the Tripe P programme and the control group is supposed to be the effect of the Triple P. The two variables of interest here are 1) exposure to the Triple P programme and 2) the emotional and behavioural problems of the children.

Pedersen also describes some of the weaknesses of variance oriented strategies: if the study is carried out as an experiment, it can be hard to control for the influence of other causal factor. Furthermore, writes Pedersen, the effect in the experiment is often dependent upon the context that

the experiment happened within. Lastly, he mentions that we sometimes need more than the average-effect of a certain factor which is what variance studies provide us with. Sometimes we need, either as a substitute or as a supporting explanatory model, more 'process oriented' strategies.

Process studies are defined as studies that seek to determine *how* some causal factor can give rise to some effect. These studies look into process or mechanism, as Pedersen writes, that connects the cause and effect (Pedersen, 2010, p.14). Often, these strategies are qualitative. An instantiation of a process type study is called process tracing, which Pedersen mentions as useful for the evaluation of a policy effect. He connects this to the formulation of what he calls a programme theory. This is a theory that explicates how a given policy or social method would lead to its intended effect. The advantages of using a process oriented strategy are said to be that they expose the 'components´ of some social methods. It is worth noticing that two things at play here; process studies both seem to be concerned with the components that makes some factor X causes Y *and* the process through which X causes Y. Later in the thesis, I will argue that there is actually a difference between these two.

Pedersen's description generally fits well with other academics' description of process studies or process tracing: as an exploration of two variables by mapping out 'what connects them' (Maxwell, 2004), or 'a series of smaller causal steps in between' (Cartwright & Hardie, 2012). Others have used the metaphor of dominos to describe process studies; if you imagine that you look upon a row of dominos, where every brick besides the first and the last one are behind a veil, and the first and the last domino has fallen. If you want to see whether the first domino caused the last one to fall, you might be interested in lifting the veil to see if the other dominos have fallen too (George & Bennett, 2005).

As a concluding remark, Pedersen states that one does not have to choose between process and variance studies but can combine them. This can according to Pedersen give a better understanding of the causal relations. He is describes this the following way:

"På den made (with a mixed-study approach, red) vil det være muligt at få bade en mere ekstensiv og intensiv viden om kausalsammenhænge: Viden om sammenhænge mellem få variable udstrakt i

tid og rum, samtidig med mere lokale, komplekse og kontekstnære kausalforståelser."[15] *(Pedersen, 2010, p.18)*

I will discuss this claim later. To summarise, Pedersen defines variance studies as studies that seek to measure some relations between two variables and gain causal information from this by controlling for the influence of other factors. Thus, variance studies seem a broad category that would contain most of what is found under the label of statistics, quantitative method, or econometrics. Process studies are on the other hand methods that seek to map out the components of the mechanisms that gives rise to causal relations, also described as the process through which X causes Y.

The aim of this section has been to gain a better grasp of the difference between the two types of studies that are ranked in the ranking scheme. The more critical analysis of the ranking will be carried out in the following sections.

6. Internal Validity Threats to Variance Type Studies

As described in the Introduction, my approach to discussing the ranking of the effect-evidence ranking scheme is to analyze some of the main threats to internal and external validity, and consider how the two types of studies fair in this light. Therefore, there will be a section first on the internal validity and external validity threats to variance studies, followed by the same for process studies. I will summarize my findings in the conclusion.

Confounders

In variance studies, one of the main threats to internal validity is that of confounders, especially *unknown* ones as these can be tricky to control for statistically (Steel, 2008). A confounder is a variable that is correlated with the treatment and has a causal relationship with the effect variable. The problem of confounders may arise in a study when a correlation between two variables of

[15] Own translation: 'This way, it will be possible to get both a more extensive and intensive knowledge about causal connections: Knowledge about few variables, extended in time along with more local, complex and close-to-context understandings of causality.'

interest has been measured, and we want to infer whether or not there is a causal relationship between them, in the study. Even if the variable that we take to be the cause (this is also called the treatment variable) correlates with the effect variable, this does not, as the old saying goes, imply causation. This is because the effect we see may be the result of the confounder, and not what we take to be the effect variable. This problem is thus closely connected to the problem of spurious correlations.

The practical problem here is of course that if we erroneously believe X to be causing Y, while they are in fact just effects of a confounder, Z, the introduction of X will not cause Y. If X is a social method and Y is a desired social outcome, we see why this is problematic. Naturally, confounders are not the only concern when doing variance studies. For example, the sample size is also important, but in line with the Introduction section, I will focus on the philosophical aspect of the issue and leave considerations of the right sample size aside, which I consider part of statistics proper.

It is worth noting how many of the traditional methodological problems of internal validity are actually problems of confounders. For example, according to Mayo (1996), one of the most central problems in science is how to arrive at solid experimental knowledge. To Mayo, the best way of doing this is what she calls learning from error:

> *"It is learned that an error is absent when… a procedure of inquiry... having a high probability of detecting the error if (and only if) it exists nevertheless fails to do so, but instead produces results that accord well with the absence of the error."* (Mayo, 1996, p. 64)

When Mayo speaks of errors, she refers to the possibility of not being able to distinguish real from artificial effects in an experiment (Mayo, 1996). This is just the danger when it comes to internal validity; that we draw mistaken causal conclusions. What Mayo prescribes for our studies is a procedure in which we are highly likely to detect error, i.e. discover that the effect we are observing is 'artificial' rather than real, if this is the case.

How does this relate to confounders? Because artificial effects are here brought about by confounders. In the case of experiments in the natural sciences, these confounders may be the instruments used for measurement, which I think is what leads Mayo to talk about 'artificial' effects. The fact that the problem of artificial effects can be thought of as a problem of confounders becomes more clear when Mayo refers to Hacking's and Galison's descriptions of experimental practice. Both describe how scientists vary methods of observation (triangulation) and instruments in order to ensure that the effect of the study is real. Why would they do this? Because by alternating the methods and instruments of observation we seek to make sure that the effect that the experiment implies is not caused by the methods or instruments that we use. If we get the same correlation between two variables no matter what instruments we use, it would seem more and more plausible that the instruments we use do not give rise to the correlation, but that there is a genuine causal relationship between the variables[16]. Thus, we seek to secure ourselves against a form of confounders here –those that can be brought about by our very tools of conducting studies.

The above may also be why Steel (2008, chapter 9) considers confounders the main problem in statistical inference in social science – confounders seem a potential problem whether they come in the form of causes at work in the world outside the laboratory or from our instruments of observation. As Steel notes, there are certain statistical tools for mitigating the threat of confounders. As the focus of the thesis is not statistics proper, I will leave the more technical issue.

However, if we believe that the threat of confounders is a central issue to variance type studies, it is of course of central interest to discuss whether some types of variance-oriented studies are somehow guarded from this threat to their internal validity. Here, I am thinking of a study type that has gained tremendous popularity within recent years, especially when it comes to policy making; the random control trial (RCT). The exciting prospect that some academics and practitioners claim for this type of study is that it can *rule out* the interference of confounding factors by their very design. Even more excitingly, it is sometimes claimed that RCTs can rule out the threat of *unknown confounders.* These are the causal factors that have an influence on the effect variable we seek to measure, but that we do not know exist and which are therefore difficult to control for via other

[16] Of course, the purported effect may also be the result of a confounder that has nothing to do with the instruments.

means of control. In what follows I will discuss this claim to see whether variance type studies do not suffer from the threat to their internal validity that confounders present.

Are Random Control Trials the Solution to the Problem of Confounders?

Random control trials (RCTs) are seen by many as the gold standard of evidence (Nancy Cartwright, 2007; Worrall, 2002, 2007), ranking above other variance type studies, such cohort studies. However, the NBSS does not have a special category for them in their effect-evidence ranking scheme. Is this sound when other institutions[17] seem to think so highly of them? This chapter will be concerned with explaining what RCTs are, and what their special virtues are supposed to be. Lastly, I will evaluate whether NBSS are right not to have a special effect evaluation category for these, but instead have an overall category for variance studies and what this tells us about threats to IV from confounders.

How does an RCT work, then? I have already given an outline of this in the section that defined variance and process studies, but I will go more into depth here. A random control trial works by splitting the participants in the experiment into two groups through a random selection (through a draw, for example). One group is then exposed to a treatment, which can be everything from a drug to a social method, while the other group is left untreated. The untreated group is often referred to as the control group. The effect of the treatment is then found by looking at the difference, if any, between the control group and the treatment group after the treatment, with regards to the effect variable we are interested in. For example, one might be interested in finding out what effect a new, special job programme has, in terms of the number of people in jobs. In this case, one would randomly select a control group and a group to receive treatment from a job-seeking population. The control group would not receive any treatment (they would most likely simply participate in the ordinary job programme) while the treatment group would be enrolled in the new, special job programme. The effect of the new programme would then be measured by subtracting the number of people that found a job in the treatment group, from the number of people that found a job in the control group. Thereby, we should see the isolated effect of the job programme.

[17] See, for an example, The UK Department of Education's User Friendly Guide on https://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

RCTs are often proclaimed to have some epistemic virtues when it comes to dealing with the problems of confounders. I will consider these below.

First, with a RCT we can allegedly infer the effect of our treatment "undisturbed" by confounding factors that we know would affect our result. So if we wish to see the effect of the job programme from before, we know that other factors besides our programme can affect whether people get jobs. The RCT takes care of this by supposedly having such confounders represented in equal proportions in the control group as in the treatment group, and then subtracting the effect we see in the control group. The trick that allows us to suppose this equal representation of confounders is randomization. Because the participants in the control and the treatment group are chosen randomly from the population, we will, with enough participants, have reasons to suppose that the control and treatment group are equally influenced by the confounders, or so the argument goes (Worrall, 2002).

Secondly, this is supposed to be the case not only for known confounders but also *unknown* confounders that can influence the effect variable. The reason we can supposedly assume this is the same as with known confounders; if we randomize properly, even the factors that affect the result and which we do not know about will allegedly be represented equally in the control and the treatment group (Worrall, 2002).

These two claims have been contested in the philosophy of science literature. I wish to focus on the criticism provided by Worrall as he taken much of the debate into consideration. I will also briefly refer to Urbach (Howson & Urbach, 2005; Urbach, 1985), who brings a very concise discussion of the subject.

Worrall argues that RCTs do not necessarily lead to an equal distribution of confounding factors among the control and the treatment group, be they know or unknown. We may, for example, be interested in what effect a particular, new reading programme in school has on the pupils reading scores. Following the RCT guidelines, we would gather a sample of pupils for the test, and then randomly assign them to the control and treatment group, subjecting the pupils of the treatment group to the new reading programme, while the control group would continue the regular reading

programme. The difference between the new reading scores of the control group and the treatment group should then be the effect of the programme relative to the old programme. However, we also know that other factors influence how well pupils read, for example the intelligence of the pupil. And there might, by pure coincidence, just happen to be more intelligent pupils in our control group than in the test group. This might result in the control group having a higher average reading score at the end of the experiment than the treatment group, *even if* the new reading programme actually improved the reading scores of the pupils. Now this example only shows that randomization does not *necessarily* leads to an equal distribution of confounders, which Worrall takes to be a slightly dull and obvious point (Worrall, 2002, 2007). The more interesting question is whether it will *most likely* lead to this. However, Worrall finds no reason to suppose this either, and believe it is the result of a of misconception. This is especially a issue for the *unknown* confounders, as we cannot control for these by the means of statistical control that we might use to control for the known confounders. Therefore, I will focus on *unknown confounders* in the following.

Worrall's discussion of why RCTs cannot control for unknown and know confounder consider three accounts of RCTs one from Papineua, Cartwright and Pearl each (Worrall, 2007). In all of these accounts, Worrall finds that the ability of the RCT to rule out confounders in based on what some 'ideal RCT' could do, but which is not practically feasible. The accounts all appear to contain some version of the argument that, if we randomizations over and over again, in the sense that we re-randomize the whole sample again and again, the average difference between the control group and the treatment group should be the effect of the treatment under study. However, as Worrall points out, in practice it is often not feasible to repeat the experiment even once, and for the re-randomization to guarantee that that we can isolate the effect of the treatment variable (say, the new reading programme's effect on the reading scores), we need to re-randomize much more than that.

Urbach (Howson & Urbach, 1993; Urbach, 1985) also come to the same conclusion as Worrall (Worrall also mentions being inspired by Urbach) though he does so as part of arguing for the superiority of Bayesian statistics over the classical, frequentist approach. Urbach provides two arguments against the claim that RCTs shield is from confounders. First, Urbach (Howson & Urbach, 1993, p. 194-198) asks us to suppose that for each existing confounder, there is a chance, albeit very small if we believe the proponents of RCTs, that it is distributed unequally among the treatment and the control group. However, even if there is only a small chance that this is the case,

there may be an enormous amount of confounders, and thus the total probability that *at least one of them* is unequally distributed among the control and treatment group may be substantial. Note that, since the confounders here are *unknown* we do not, trivially, know how many of them there are. Secondly, Urbach points out that it is only the sample of the group (for example, the pupils) that is randomized, not the other conditions that may lead to a different distribution of confounders. For example, if the facilities that the control and the treatment group are placed in are not randomized, and if there are some unknown confounders relating to this, then these will not be randomized away.

Not being a statistician myself, I have relied on the debate in the literature to weigh whether Worrall, Urbach and others are right in their criticism. I have found two article of interest, where Dr. Adam La Caze is either author or co-author. The interesting part of this is that, though I cannot say so with certainty, La Caze appears to have changed his perception on Worrall and Urbach's arguments; from what looks like an outright rejection of their relevance to accepting them. This may give some credence to Worrall's argument. La Caze is interesting because his academic work is done with the context of medicine, where RCTs also appears to enjoy great support. Thus, if La Caze reflects this opinion in the medicine community, we may expect him inclined to be in favour of RCTs.

Indeed, in the first article in which La Caze is a co-author (La Caze, Djulbegovic, & Senn, 2011) it is argued that:

> "*The crucial point is that it is not necessary that randomisation control for all known and unknown confounders at baseline to make valid statistical inferences, it is sufficient to know their distribution in probability, which randomisation is designed to provide."* (La Caze et al., 2011, p.1)

This seem to simply fly in the face of Worrall and Urbach's arguments; their point is exactly that we cannot know the distribution of probabilities of *unknown* confounders, exactly because they are unknown. However, in a later article La Caze (2013) appears to accept Worrall's critique and

instead seek to defend RCTs because they rule out selection bias[18]. In the conclusion he writes, "Randomization does not provide the guarantee that all possible confounders are evenly distributed in experimental groups and therefore does not provide some irrefutable epistemic good"(La Caze, 2013, p.365). He also gives a more lengthy discussion of why, which basically underlines Worrall's point that RCTs only equally distribute confounders if we repeat randomization ad infinitum: "Randomization provides categorical assurances on the distribution of known and unknown confounders but only in the indefinite sequence of trials (La Caze, 2013, p.362)."

Summary on Internal Validity Threats for Variance Studies

Due to Worrall and Urbach's arguments, and La Caze apparently being persuaded by these, I conclude that confounders remains a persist threat to validity, especially unknown ones, even though we have RCTs at our disposal. Thus, it seems that the best we can do is to use the various forms of control available in order to ensure that the *known* confounders do not infect our studies.

7. External Validity Threats to Variance Type Studies

The chapter before addressed the threat of confounders to the internal validity of a study. However, when it comes to studies that are to be used to judge whether some social method will have a given effect, we need not only secure that the causal conclusion holds for the object of study but that it can be 'extrapolated' (Steel, 2008) or, in other words, that the causal conclusion 'travels' (Cartwright & Hardie, 2012). This is because we want to use the finding in municipalities where our studies where not conducted. This is the problem of 'external validity'.

Support factors and Properly Defined Causes

The first threat to external validity that I wish to explore is that of A major theme in Cartwright and Hardie (2012) (C&H) book is that even if we have conducted a perfect random control trial where all the confounding factors have been distributed equally, we may still be erroneous in extrapolating the causal conclusion, thus, we have still not secured external validity. This is because the domain

[18] I do not have room for a lenghty discussion of this, but I side with Worrall in holding that selection bias can be avoided without an RCT and that this thus provide no special status for RCTs.

that we wish to extrapolate the casual conclusion to might lack what C&H have named *support factors.* These are factors that need to be present in the target before the treatment that was found effective in a study will hold outside that study's domain. C&H have an excellent example of this and since it also encompasses social method, like the NBSS case does, I consider it worthwhile to describe it in some length. The work of C&H is generally interesting for this thesis as Cartwright is a leading figure in the philosophy of science and the book deals directly with social policy.

In the mid-1990s, the public schools in California had issues with the academic performance of its school pupils. There was a widespread belief that a remedy for this problem would be a reduction of class sizes, so that the pupil per teacher ratio would be lowered. In addition to being supported by popular opinion, this initiative was support by evidence from a RCT conducted in Tennessee in 1985. The study from Tennessee showed that the younger pupils did indeed perform better in smaller classes. Thus, the programme of smaller class sizes was adopted on a large scale in California.

Alas, when an evaluation was done in 2002, there could be found no link between class-size reduction and academic achievement among the students (Cartwright & Hardie, 2012, p.4-5).

How could this be? One reason was that Tennessee had both the availability of decent sized classrooms and good teachers when the RCT was conducted. When the class sizes were reduced, there was naturally a need for more teachers and for more classrooms (smaller classes meant more classes) and in Tennessee there was no problem with providing this. The contrary was the case in California, where many unqualified teachers had to be hired and sub-par class rooms had to be used (Cartwright & Hardie, 2012). Decent classrooms and qualified teachers thus turn out to be *support factors* for smaller classes to be able to bring about better academic performances. That is, it appears that the effect of smaller classrooms, found in the RCT, does not appear unless certain other factors are in place. The factors that need to be place before the effect can take place are support factors. Note that support factors are not the same as confounders. What C&H want to illustrate is that qualified teachers and good classrooms are *conditions* for smaller classroom to cause better academic performance. This does not mean the influence of these obscured the study made in Tennessee; it means that because they were present in Tennessee but not in California smaller classes had no effect.

We should note, as Cartwright and Hardie (2012, p.70-73) are quite clear on, that there is no hierarchy between the support factors and the factor that we describe as being the cause of some effect; both what we call the cause and what we call the support factors are equally important to get the effect we are after. Thus, it would be more accurate to say that it was the *combination* of smaller classrooms, qualified teachers, available classrooms etc. in Tennessee that caused better academic performance among the students.

The problem with this in practice is of course that we cannot implement the full list[19] of circumstances that was present under an RCT in another domain. Due to this, one could argue that the threat of lacking support factors in the target is actually an internal validity problem. As the consideration above indicate, if the conclusion of the RCT in Tennessee was that smaller classrooms caused better academic performances, this could be argued to be false; smaller classrooms *and* qualified teachers *and* decent class room etc. leads to better academic performances, not just smaller classroom.

However, I will follow C&H and view the problem of support factors as a problem of external validity, as it only becomes a practical problem when we wish to extrapolate the result of the study and cannot replicate all the circumstances that were present in the study.

Luckily, there is always a huge amount of features that we know will be causally irrelevant if we wish to extrapolate the result of the study. For example, common sense may tell us that the greater frequency of cowboy boot wearing teachers that presumably participated in the RCT in Tennessee, compared to California, was probably not a support factor in making smaller class sizes cause better academic achievement. Though less often so with social sciences, in many cases we will not just have common sense but also scientifically mature theories to tell us which variables will be causally effective and which will not. Like Mayo describes, we can achieve such knowledge through manipulating a single variable at a time in our experiments in order to see whether changes in these variables are causally relevant for what we are studying (Mayo, 1996).

[19] It is a philosophical curiosity whether such an exhaustive list of circumstances at some event is possible. It seems that it would require something like to logical atomism which is largely considered unfashionable today.

The upshot of this is that even if we effectively rule out confounders, there is still an uncertainty present in extrapolating variance studies – the uncertainty that we have not actually located the full set of support factors that are responsible for bringing about the effect that we inferred through the RCT. The remedy for this, as C&H describes, is that we need some kind of theory of what brings about the effect; what elements are causally relevant.

Having described support factors, we can tie a knot on the loose end in our section on the possible trade-off between internal and external validity. Here, I argued that, for our case, there was only one reason to fear a trade-off, namely Cartwright's claim that the result of studies or models may rely on specifications in the study that does not hold in the target. However, we actually already take this into consideration by checking for support factors; these are exactly the factors that the target need to have in order for the cause to do the same work as it did in our object of study. If the conclusion of a study is dependent upon certain conditions or assumptions in that study, which grants it internal validity, these will be the support factors that are needed to extrapolate the study. Thus, if we take support factors into consideration when giving an effect-evidence ranking, our ranking will reflect this potential problem.

Different Mechanisms

I now wish to turn to another threat to the external validity of variance studies that has been discussed in the philosophy of science literature. More often than support factors, the debate on external validity has been framed around possible differences between the object of study and the target in terms of the *mechanisms* that lead to the effect. For example, is it legitimate to extrapolate the effect a drug will have on mice to the effect this drug will have on humans, despite the difference in the biological mechanisms in humans and mice? To understand this issue, we should first understand what ontological commitments that support this concept, as well as the role of the concept of mechanisms in the philosophy of science literature.

An excellent reason for why we should care about underlying mechanisms is Cartwright's ontology of nomonological machines and chance-setups. A basic tenet of Cartwright's philosophy is that we come to a better understanding of the success and workings of science by asking where laws (i.e. regularities in the world) come from. Such a question may strike philosophers of a Humean view as

strange; to them, there is nothing 'behind' regularities. Cartwright argues extensively for this ontology, especially in (Cartwright, 1983). Due to the limits of my thesis I cannot analyze this debate, but I will use Cartwright's ontology to explain the problem of different mechanisms in the object of study and in the target. Cartwright's answer to the question of what gives rise to laws is *nomonological machines*. These are configurations of entities that give rise to stable regularities – laws. In using this concept, Cartwright seeks to argue that regularities only arise arise under certain conditions – when the entities that make for regularities are set up in the right way Since these laws will often be probabilistic, Cartwright states that they give rise to chance-setups (Cartwright, 1999a, 1999b). I will here give one of her examples, which illustrates the threat of different mechanisms to external validity nicely.

In her example, Cartwright draws on economics where she criticizes the work of two economists on how social expenditure affects the level of public welfare in general ((Cartwright, 1999b). The economists that she criticizes seek to make a statistical inference (a variance study in our terminology) by testing whether social expenditure is large in countries with high welfare when also trying to measure a range of other variables that may have influence on this, such as income pr. capita and technological development (thus, controlling for confounders). Cartwright writes that:

> *"Now what strikes me is that this methodology is crazy. That is because what this equation represents is a 'free-standing association': there is no good reason to think there is a chance set-up that generates it… to suppose that there really is some probability measure over welfare expenditure and welfare like that presupposed in the equation, you need a lot of good arguments" (Cartwright, 1999b, p.324)*

What Cartwright argues here is that it is meaningless to ask whether social expenditure causes higher levels of welfare for developing countries tout court. Social expenditure may well cause higher levels of welfare, but only under certain conditions, that is, when certain other entities of the economy are configured the right way (say, the banking system, the level of unemployment etc.) Thus, to build a nomonological machine, where social expenditure causes higher levels of welfare we need many other *mechanism* components.

How does this relate to any threat to external validity? Let us say that social expenditure actually causes higher welfare in Sri Lanka which the example suggests (Cartwright, 1999b). Following Cartwright's line of argument, this does not automatically imply that we can extrapolate that higher social expenditure will do so in other developing countries; these may lack the other parts of the nomonological machine that makes it possible for social expenditure to cause higher welfare in a law-like way. Thus, the threat is that there may be a different mechanism in the target, compared to the object of study, why the cause that was found to be effective in the study will not be effective in the target. Another way saying this is that the target may *lack* the mechanism that is present in the object of study.

The example above illustrates what mechanisms are in Cartwright's account; they are configurations of entities that will lead to a regular behavior. I will draw on this account of mechanisms in the rest of the discussion.

As the reader may have noticed, this problem looks curiously much like the problem of support factors, and indeed, I will argue that these are basically the same problems in slightly different framings. This is no trivial claim however, as we shall see when I go over Daniel Steel's (2008) solution to the problem of different mechanisms where he tries to outline a strategy for dealing with this threat. Few others in the philosophy of science (for instance Francis Guala and Nancy Cartwright) have made systematic attempts to set up conditions for extrapolation, why I consider Steel's argument worth discussing.

Below, I will clarify why mechanisms in Steel's view cannot be reducible to support factors. This, I will argue is due to an unwise mix-up of support factors and causal processes. The discussion of this is relevant to the NBSS effect ranking, as we ought to know whether there are two different threats to validity for variance studies, or really just one, that we should take into consideration when judging the validity of process studies or variance studies.

Comparative Process Tracing to the Rescue

In this section, I will discuss Steel's solution to the threat of different mechanisms. Steel writes that he wishes to use a mechanism-based approach to provide an account of how we can extrapolate

from one population to another when these populations are not homogenous. Thus, he seeks to deal with the problem of different mechanisms, also citing Cartwright's account of nomonological machines as an example of an account of mechanisms (Steel, 2008, p.83). Steel argues against what he calls the 'simple induction' criteria for extrapolation: "Assume that the causal generalization true of the base population (the population in the study, *red*) also holds approximately in related populations, unless there is a specific reason to think otherwise" (Steel, 2008, p.80).

This way of going about extrapolation is obviously crude and we will end up extrapolating results that cannot be extrapolated. Thus, Steel suggests a strategy that he terms 'comparative process tracing'. Steel's notion of process tracing appear synonymous to how Pedersen (2010) describes process oriented strategies (that I refer to simply as process studies); both are about exposing a mechanisms that leads from the cause to the effect (Steel, 2008, chapter 9). The comparative part of the method lies in comparing the process that leads to the causal relation in the study to the process that would be needed in the *target*, in order for this causal relation to hold true (Steel, 2008, chapter 5).

Steel presents his strategy the following way. First, we should learn the mechanism in our object of study. Secondly, we should compare the stages of this mechanism with that in the target at the most downstream point where they are likely to differ. The different stages here refer to parts of the process that connect the treatment, say a social policy, and the effect variable, say lower crime rates. Thus, we could imagine that a social policy would work through the mechanism of promoting employment opportunities, thereby reducing crime rate. The process here would be 'treatment → more employment opportunities → lower crime rates'. If we have a study showing that this process occurs somewhere, Steel's proposal urges us to check that the mechanism is also present in our target before extrapolating the effect of the treatment. If the reader thinks of this as somewhat vague, I will agree. This will be part of my criticism of Steel in the next section.

What does down and upstream stages refer to then, when Steel's criterion states that we do not need to check that all the stages of a causal process are the same in the target and our object of study, but only the downstream stage?

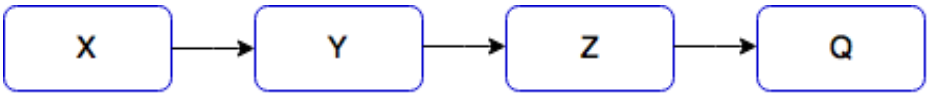


Figure 2

Figure 2 illustrates a causal process, where the mechanism that connects a treatment, X, are linked to the effect variable, Q, through Y and Z. Thus, X causes Y, which causes Z, which in turn causes Q, the effect variable. Steel writes that, "If differences in X or Y must result in differences in Z, then it is necessary only to compare the model and the target at Z." (Steel, 2008, p.91) Steel's claim is that when we are doing comparative process tracing, we have the advantage of only needing to compare the downstream part of the processes of our object of study to the downstream part of the processes of our target. The idea behind this seems to be that if we can verify that both the object of study and the target have Z, and if Z somehow shows that Y is also present, then we do not need to separately check for Y to be present.

He does mention two exceptions for this, however: if there is a causal connection between an upstream process and a downstream process that bypasses the downstream part of the process that we compare, then only looking at the downstream part of the process might lead us to wrong extrapolations. Secondly, the downstream part of the process must not have been the result of some independent cause; Steel writes that the upstream must 'leave a mark' on the downstream process (Steel, 2008, p.90). In the next section where I discuss Steel's proposal, I will argue that these exceptions will almost always be present in a social science context.

The last specification of Steel's criterion that I wish to present is his argument against a criterion for extrapolation provided by LaFollette & Shark. They suggest that for external validity to be ensured, "there must be no causally relevant disanalogies between the model and the thing being modelled" (Steel, 2008, p.93). That is, if we want to extrapolate our finding in a study that X causes Y, our target for extrapolation must not be different for our object of study in any way that would impact X's causal relation to Y in that target. Steel argues that contrary to the criterion of simple induction, this criterion is too strict. Also, argues Steel, if we were to take this criterion seriously, not only would we often not be able to use the results of experiments on animals, we would rarely be able to extrapolate from one human population to another.

Steel argues against LaFollette & Shark's criteria. He argues that there only needs to be no relevant disanalogies in case we want to extrapolate the precise, quantitative result of the study unto our target. For example, if we want to use our finding that some social method lowers committed crime rate by 10 % to conclude that the same method will also lower crime rate by 10 % in our target. However, writes Steel, often we are quite happy to just be able to infer from our studies that a social method will have *some* positive impact, thus, will lower crime in our target by some rate. In this case, we can accept some disanalogies that influence the treatment's effect. As discussed under the section Different Needs to Know, this is the kind of predictions that I will focus on in this thesis and that I think we could reasonably ask for in social policy. Thus, I agree with Steel here.

It is worth taking a moment to recapture and relate this to our case. If Steel is right in how we should deal with the threat of different mechanism in the target and the object of study, variance studies need to be underpinned by a kind of process study, that compares the process in the target and the object of study. This is important when we want to consider how variance and process studies should be ranked, and I will summarise this in the conclusion. I will now discuss whether Steel is actually right in his criteria for extrapolation. After this, we can wrap up the discussion on external validity threats to variance studies.

Discussion of Comparative Process Tracing as the Solution to Different Mechanisms

In this section I will discuss Steel's strategy for solving the problem of different mechanisms in the target. First, I will use a point made by Cartwright and Hardie (2012) as a criticism of one of Steel's claims. Secondly, I will argue, as I have already hinted at before, that Steel's account makes an unfortunate mix of mechanisms and processes. This leads to two things; that his own strategy of comparative process tracing becomes blurry, and, as I will argue in the section The Danger of an Infinite Regress, it opens up for an infinite regress. Relating this to our case, the conclusion of my argument will be that we do not necessarily need a process study to have external validly for variance studies, which seems to be the outcome of Steel's solution. This has important consequences for how we should think about the ranking of variance and process studies.

The comparative element of Steel's account strongly suggests that an element of similarity in the processes between the target and the object of study is necessary for extrapolation, albeit a less strict similarity than LaFollette & Shark's version. However, Cartwright and Hardie (2012) argue against similarity as a guiding principle for external validity in the following way: "Should you be looking for similarity between your population and the study population in these cases. No. You should be looking for what matters to getting the prediction you have in view right" (Cartwright & Hardie, 2012, p.47). What C&H points at here is that what matters is whether our manipulation, say a social method or policy, will have the effect we wish, not how this is realized. Thus, it does not matter whether the process that connects our treatment to the wanted effect in the target is the same, or is similar, to the process that did this in our object of study. As long as there *is* a process, the causal relationship between the cause and effect is secured. Thus, while similarity of the right kind between the process in an object of study and the target may be *sufficient* to secure external validity of the study, it is not *necessary.*

In Steel's defence, it may be said that we are not actually extrapolating result if we are able to establish that some cause will have an effect through another process than the one studied. In that case, we would seem that we already know a process by which our treatment can affect our effect variable, independently of the study. However, I still think that what C&H warrant is important in practice; it makes sure that we do not ignore the possibility of multiple realizations for a given effect.

Having considered C&H's argument and how it relates to Steel, I will now turn to my claim that Steel confuses processes and mechanisms in an unfortunate way.

On Steels account, mechanisms appear to be 'successive.' Thus, a mechanism seems to consist of a series of variables that affect each other in turn and finally affect the variable that we are interested in, like a row of dominos or Newton's cradle. In defining mechanism, Steel draws Machamer, Darden and Craver, who state that: "Mechanisms are entities and activities organized in such a way that they are productive of regular changes from *start or set-up to finish or terminal conditions* (own italics, red)*"* (Steel, 2008, p.61). Also, without succession, Steel's idea of an up or downstream part of a mechanism that I have laid out above does not make sense. For there to be an

up or downstream part of a mechanism, there needs to be a chain of variables causing each other. Thus, processes and mechanism become one in Steel's account.

However, just like in (Pedersen, 2010), mechanisms in Steel's work *also* denotes a set of components that lead to some outcome. Here, he may have been inspired by Cartwright's notion of nomonological machines, where succession does not appear integrated in the concept. I will argue that we should strike a difference between these two concepts of mechanisms: as causal processes and as components that enable some cause to have its effect. I will discuss this by considering an example that Steel gives of process tracing in social science:

"Let us examine a case of process tracing in social science. For example, consider Malinowski's hypothesis that the possession of many wives was a cause of wealth and influence among Trobriand chiefs (1935)... First, there is a custom whereby brothers contribute substantial gifts of yams to the households of their married sisters—gifts that are larger than usual when the sister is married to a chief. Second, political endeavours and public projects undertaken by chiefs are financed primarily with yams. As this case illustrates, process tracing in social science often provides evidence for the existence of several prevalent social practices that, when linked together, constitute a mechanism. Supposing that Malinowski was right about the two features of Trobriand society just described, the conclusion that the number of wives had an influence upon wealth among Trobriand chiefs is unavoidable" (Steel, 2008, p.189).

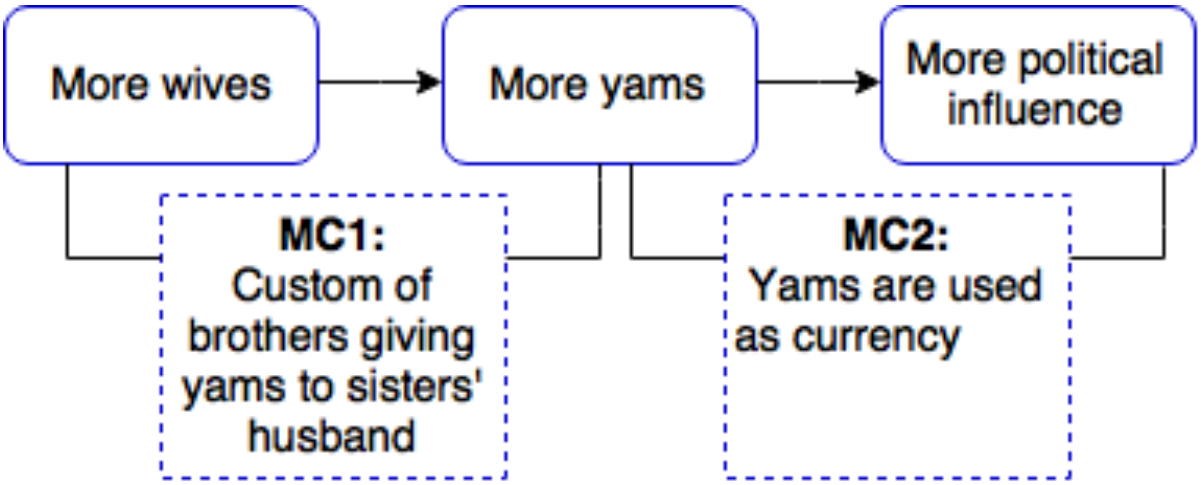


Figure 3

I have tried to illustrate Steel's example in figure 3 above. Here, we see how, in the Trobriand society, more wives will lead to more yams, which in turn will lead to more political influence. MC stands for mechanism components, and I will elaborate on why I have named them so below.

If we want to follow Steel's advice, we should do comparative process tracing before extrapolating. Thus, if we had concluded that in the Trobriand society, more wives leads to more political influence, we should look into the process through which this happens (through 'more yams' in this case) and check for differences in terms of this process in the target.

Now, how does this example of a study give us basis for extrapolation? It does this very much in virtue of MC1 and MC2; these specify what needs to be in order in a target before we could extrapolate the finding that more wives lead to more political influence. The study does *not* provide us with a basis for extrapolation *merely* by showing us that more wives lead to more yams, which leads to more political influence. Thus, 'the successive' element, which Steel and Pedersen both seem to employ when describing mechanisms (because mechanisms and processes are not properly separated) is not *sufficient* to give us basis for extrapolation. If we just knew that more yams lead to more wives and this lead to more political influence in the Trobriand society, and wished to use this as a basis for extrapolation, we would seem to be begging the question: How do we know that this happens in the target? Under which conditions does more wives lead to more yams and more yams lead to more political influence, and what mechanism would need to be constituted for this to happen? Another way of putting this, and the reason I have used the term mechanism components

in the figure, is that is it only when we have MC1 that we can build a nomonological machine, where more wives causes more yams, and it is only when we have MC2 that we can build a nomonological machine, where more wives yams causes more political influence.

This is why we ought to differ between causal processes and mechanisms, where I will define processes as a succession of causal factor affecting each other, and mechanisms as the set of factors that make some cause effective (properly speaking, the 'cause' would be a component in the mechanism that causes this effect). If we do not differ between these two things, we end up with a wrong criterion for external validity, which would invalidate our effect ranking.

With the distinction I argue for, we can also make better sense of what Steel's comparative process tracing would look like in practice. Consider the Trobriand society again. It does not seem possible to check that the causal *process* in the target is similar to the one in the object of study – if we were able to check in a study that in the target, more wives does indeed cause more yams which in turn does cause more political influence, we would already be home safe. However, the very point of extrapolating is that we will often not be able to make these studies! What we need to look for before we can extrapolate are whether the two mechanism components are present, MC1 and MC2. If this is the case, we can feel reasonably sure in our extrapolation.

Thus, we should not mix up the notions of process tracing and mechanism. Somewhat contrary to intuition, we can show a process through which a chain of variables affect each other, without having exposed the mechanisms, if we want mechanisms to denote what we need for extrapolation. This means process tracing, understood as a successive chain of variables affecting each other, is not *sufficient* for extrapolation. In the next section, I will argue that it is not even *necessary* for extrapolation! This is fortunate, I will argue, as the requirement of process tracing would lead to an infinite regress.

The Danger of an Infinite Regress

One feature that may have struck the reader is that there seems to be a regress involved in Steel's comparative process tracing solution to the problem of external validity. I will argue that this can be taken care of by the distinction between mechanisms and processes that I introduce above.

The regress in comparative process tracing is due to the following: If we take it as a requirement that we need to locate a process, meaning at least one 'intermediate' variable between the treatment and the effect variable, in order to secure extrapolation, what about locating the variable between treatment and this intermediate variable or between this intermediate variable and the effect variable? And why stop there, would we not need to trace processes between processes ad infinitum?

Steel discusses a regress in relation to an argument made by Kincaid (1990). Kincaid is sceptical about the requirement that we should be able to locate a mechanism before a correlation between X and Y allows us to infer that X indeed causes Y. Thus, Kincaid comments on the regress involved when we try to use mechanisms as a tactic for dealing with confounders. Besides arguing that this will lead to a regress, Kincaid argues that we can just use the various statistical means for dealing with confounders. This is a different problem that I will take up in the section on the internal validity of process studies as it has to do with confounders, but the infinite regress rears its ugly head for the same reasons when it comes to comparative process tracing.

I will explain Steel's response to Kincaid and show that it will not quite do for tackling the problem of a regress with regards to comparative process tracing. Again, I should say that this is perhaps to be expected since the argument that Steel gives is in relation to mechanisms as the solution to the problem of confounders, not as providing external validity. However, for Steel's account to be successful, we should also develop an answer to the danger of an infinite regress relating to comparative process tracing.

Steel is sympathetic to Kincaid's critique of mechanisms as the only way of dealing with confounders. However, Steel mentions that we cannot always use statistical techniques for dealing with confounders; it requires certain fortunate conditions that Steel describes in some detail but that I will not go through here (Steel, 2008, chapter 9). Thus, locating a mechanism through which the cause works may be the best way of making sure that the correlation between X and Y is not just produced by a confounder, writes Steel. However, this does not entail that we *must* be able to locate a mechanism or give a process account to secure us against confounders. Sometimes we may be able to make statistical control. It just entails that *if* we can find a mechanism between X and Y,

then X can cause Y. Another way of putting this is that Steel denies that locating a mechanism (not being distinguished from locating a process in Steel's terminology) is a *necessary* condition for excluding confounders. Steel takes the regress to be blocked by this argument.

While this argument may serve to secure the use of mechanism to justify internal validity and avoid the regress that Kincaid sets up, it does so by *not* requiring that mechanism or processes be accounted for – it just presents mechanisms as *one* tactic for avoiding the problem of confounders. However, when it comes to dealing with the problem of external validity, Steel exactly *requires* that we do process tracing in order to make justified extrapolations. Therefore, his defence against Kincaid's argument of a starting regress will not support his account of how to archive external validity.

I will argue that the solution lies in the distinction between processes and mechanisms that I made before. What we should be looking for when doing comparative process tracing, as argued in the Trobriand example, are mechanism components. Whether we need to find the process of the process (going one step down the regress, so to speak) depends on at which 'level' we have good knowledge of what mechanism components are needed to bring about a certain effect. Take the Trobriand example again. Here, it does seem that we would need at process study that shows an intermediate variable between more wives and more political influence. This is because we (or at least I as a non-anthropologist) have no good idea of what mechanisms would be in place for more wives to cause more political influence without looking at the intermediate variable of 'more yams'. However, once we have done a process study, finding the intermediate variable of 'more yams', we may better be able to find the mechanism components linking these variables. On this 'level' the anthropologist can use the knowledge of MC1 and MC2, (the custom whereby brothers give yams to their sister's husband and that yams are used as currency for political endeavours) to link 'more wives' with 'more yams' and 'more 'yams' with 'more political influence.'

The big question is of course, why would we stop here? If we wish to extrapolate the findings of the Trobriand society, why would we believe that we are safe in doing so when MC1 and MC2 are present in the target? Why not do an additional process study? Because we think we have good evidence that people generally follow the customs in their society (thus brothers will usually give yams to their sister's husbands and that if someone is in possession of something that is issued to

finance political endeavours, this will lead them to more political influence). When I say good evidence here, I mean that we have much inductive knowledge about these matters; for example, we have seen that in a wide variety of cultures people usually follow the customs of their cultures and that people with funds usually have more political influence.[20] This serves to justify our belief that this will also hold for societies where MC1 and MC2 are present, and that more wives should then cause more political influence.

To recapture, we worried that there might be an infinite regress involved in Steel's comparative process tracing. This is not the case, I have argued, when we distinguish between mechanisms and processes. What we need for extrapolation is mechanism (an account of the mechanism components that are necessary for X to cause Y). We have found a good mechanism when we have inductive knowledge telling us that certain components will lead to the same outcome over many different settings. If we cannot find a mechanism by only considering the treatment and effect variable we may have to do a process study. However, what this argument implies is that not only are process studies not *sufficient* for extrapolation, they are also not *necessary.* If we have good inductive knowledge on the circumstances under which X will cause Y, we do not need a process study in order for us to justify extrapolation.

How could this last claim be true? Imagine that X causes Y through some intermediate variables, thus, through a process. If the intermediate variable(s) will lead to the effect variable under a long range of possible mechanism components, we may not need to take this components into consideration. To exemplify, there is certainly a long list of intermediate variables between (I apologize for the dark example here) being misused as a child and experiencing certain problems later in life[21]. However, we *may* be able to specify certain conditions (i.e. mechanism components) under which misuse in childhood will always raise the probability of experiencing these problems, without going into the process through which this happens (i.e. all of the factors that will be involved between being misused and experiencing the problems later in life). If the intermediate variables we have are 'robust' in the sense of always leading to some conclusion, we may not need process studies for us to see the relevant mechanism components.

[20] This should calm the more hardcore empiricist; after all this talk about mechanisms and process (this may smell like metaphysics to some) we are at last back to a good, inductive base.
[21] See http://socialstyrelsen.dk/voksne/Senfolger-af-seksuelle-overgreb/definitioner/senfolger-af-seksuelle-overgreb

In order for us to *know* that the intermediate variables are robust we will usually need to test whether X causes Y in a wide variety of circumstances; without having made a process study of the intermediate variables, we cannot rely on knowledge of how these work. As this is often not feasible, a process study may still be the best way to go about this, *if* we have better knowledge of the mechanisms at the level of these intermediate variables, like we have in the Trobriand example.

I will now summarise what I have found. I have argued that process studies, as I have defined them above, are neither sufficient nor necessary for us to have knowledge of mechanisms, which is what we need to extrapolate our findings from variance studies. However, in case we have more knowledge of support factors and mechanism components at the level of the intermediate variables, a process study might help to specify the support factors needed for extrapolation. This is important for the case as it specifies what exactly the threats of external validity are to variance type studies, and how to handle them. Thus, it informs what should be taken into consideration by the NBSS when making their effect-evidence ranking scheme. I will go over whether this threat applies to process studies as well (it may well do, as the reader have already suspected from the argument above) in the section on External Validity Threats to Process Studies.

Due to this analysis where we have distinguished between mechanisms and processes, we can reduce support factors to mechanism components. To use Cartwright's terminology, they are the parts of the nomonological machine that need to be in place in order for the cause to have its effect. Therefore, the problem of support factors can be reduced to the problem of different mechanisms. This may not seem surprising as Cartwright has (partly) developed both concepts, but as I have sought to show, this reduction cannot be made for all the ways in which mechanisms are used; Steel's account being the case in point. In Steel's writings, mechanisms include a successive element, otherwise he could not differ between the up and downstream part of mechanism. However, support factors do not cause each other in a chain-like fashion like a causal process does. Instead, they are conditions for a cause to be effective.

Downstream Tracing

The last thing to discuss regarding Steel's proposal is his notion of up and downstream process tracing. Because mechanisms and processes are now separated, I think that we should think of Steels proposal of up- and downstream process tracing in the following way: if we compare the last stage of a process (for example, the 'more yams' causing 'more political influence' stage) and find that all the support factors in the object of study are present in the target we wish to extrapolate to (in the Trobriand example, this mean that MC2 is present in the target), is this sufficient to justify extrapolation to the target?

For Steels proposal to hold the upstream processes must leave a mark on the downstream processes. Still does not explain this very thoroughly, but I believe the requirement is made to avoid running into the following problem. If we just affirm that a downstream stage of the target and of an object of study are similar, this is not enough to justify that the treatment variable will lead to this downstream stage. To exemplify, if we wish to extrapolate our findings in the Trobriand society, and can confirm that in our target MC2 is present and thus that 'more yams' will lead to 'more political influence', this in no way guarantees that the support factors for 'more wives' to cause 'more yams' (MC1) will be there. However, if we can somehow infer from the fact that MC2 is present in the target to MC1 being present as well our extrapolation is secured. Therefore, Steel requires that MC1 leaves a 'mark' on MC2. The problem with this seems to be that it has very little applicability in social science. Steel seeks to make a general account of extrapolation, drawing both on biology and social science. However, in discussing comparative process tracing, his examples (which are quite technical) only consist of those present in biology where there apparently are cases where an upstream process leaves a 'mark' on the downstream part of the process. However, in the social realm, I remain sceptical that this happens.

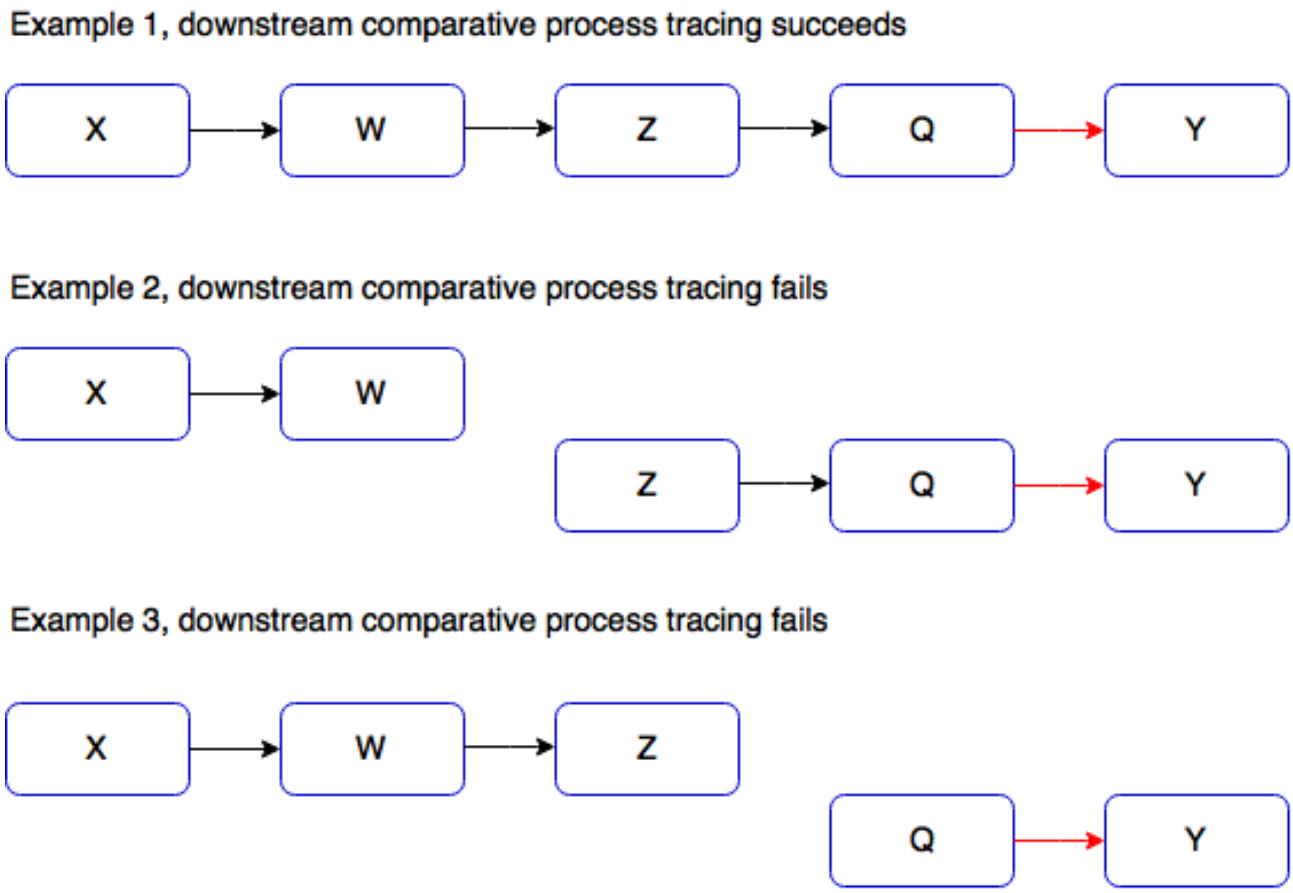


Figure 4

In figure 4, I have illustrated why downstream process tracing fails if there is no 'mark' on the downstream stage that shows that all the other stages are present. The example illustrates possible configurations of processes in the target. In the object of study, we have found that X lead to Y through the same process as illustrated in Example 1. We want to know whether we can extrapolate this claim to a target. Focusing on the downstream stage, we compare the stages or connection between Q and Y (marked red in the illustration) between the object of study and the target. If this stage is also present in the target, we would be justified in the inferring that we can extrapolate that X will cause Y in our target, if we follow the downstream process tracing guideline. We should therefore treat it as unknown whether all intermediate stages are alike. In all examples, comparative process tracing would justify an extrapolation, because Q indeed causes Y through unknown stages of W and Z. However, in Example 2 and 3 the inference will not hold because there is no

connection between X and Q. Only if there is a 'mark' from the presence of the other stages will the extrapolation hold.

Summary on External Validity Threats to Variance Studies

To take stock, we have now discussed two identified threats to the external validity of variance studies; different mechanisms in the target and the object of study and the lack of support factors in the target. I have argued that we should reduce this to one problem. Also we have discussed Steel's strategy for dealing with the problem of different mechanisms. I have argued that Steel is not quite right in stating that we must make a comparison between the process in the object of study and the target. We ought to differ between a process, as a successive chain of variable affecting each other, and mechanisms, where mechanisms consist of mechanism components that bring about an effect. We are after mechanism components for external validity; the conditions that make the cause effective, or constitute a nomonological machine, to use Cartwright's terminology. Process studies, properly defined as a successive chain of variables affecting each other are not *sufficient* to give us this, neither are they *necessary.* However, I have claimed, they *may* help us if we have more knowledge on the level of these intermediate variables in terms of support factors.

The consequence of this for the NBSS case is that we need to consider support factors when extrapolating the results of variance. However, they do necessarily need a process study in order for us to do this. I will expand upon this in the conclusion of the thesis.

8. Internal Validity Threats to Process Type Studies

The section above concludes the discussion on validity threats to variance studies. Here, we identified two threats; the internal validity threat of confounders and the external validity threat related to support factors, or mechanisms. Process studies have already been mentioned much in this discussion, due to the discussion of Steel's criterion for extrapolation of variance studies. However, in this section, I will discuss whether these internal validity threats also apply to process studies, and if so in which way.

In relation to the discussion above on the difference between processes and mechanisms, I will define process studies as studies that seek to explore the relation between two variables by showing how they are connected through one or more intermediate variables. In line with the discussion above, I will not take process studies to necessarily be exploring support factors/mechanism components.

Confounders Again

In this section, I will explore whether the internal validity (IV) threat of confounders that was present for variance studies is also a danger for process studies. This is interesting, as some authors have argued that process studies can help to eliminate the threat of confounders. For example Little (1991) argues that mechanisms can guard against confounders, and by mechanism he means a sequel of event that connects two variables; "a causal mechanism, then, is a series of events that lead from the explanans to the explanandum" (Little, 1991, p. 15), thus, a process in my definition. Little's idea is that if we can show the series of events by which X can cause Y, and we have a correlation between the variables, we can be rather confident[22] that X has a causal relation to Y. If this is the case, process studies are shielded from the threat of confounders and thus have an advantage over variance studies in this regard. However, Steel convincingly argues why this is not quite the case.

The problem is that for each 'step', or for each intermediate variable we locate between our two variables of interest, there may be a confounder. Thus, the problem has just been moved to the 'intermediate steps' of the causal relation. For example, if we wish to show that, in the short run, inflation causes lower unemployment, through the intermediate variable of lowering the real wages, we need to be able show that inflation lowers real wages and that lowered real wages leads to lower unemployment. In establishing *these* causal relations, we again face the problem of confounders.

This leads Steel to conclude that there is no fundamental difference between statistical inference (variance type studies in our terminology) and process studies. What type of study is more valid depends on, states Steel, whether we have better knowledge of the 'micro components' (the

[22] I have not been able to determine exactly how confident Little thinks we can be.

intermediate variables) that the 'macro features' (the variables of interest), and what ethical and technological constrains there are on obtaining this knowledge.

Steel does not go much more into explaining how we ought to weigh the fact that we may have more or less knowledge of the 'micro components' compared to knowledge of the relation between the variables of interest themselves. However, I will try to flesh out his proposal in some more detail, related to the threat of confounders for variance or process respectively. As I outline in the section A Question of Certainty, I will be conceptualizing this in terms of probabilities. I have tried to illustrate the situation in figure 5 below.

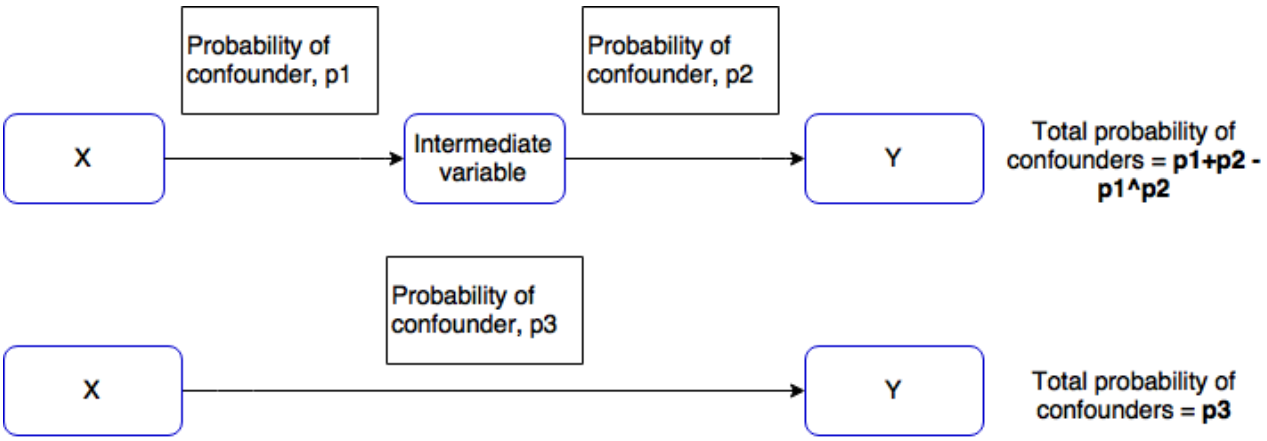


Figure 5

I have illustrated a process type study and a variance study above. p1 to p3 express the probability of our study being infected by a confounder for each causal relation the study seeks to establish. Per definition, the process study will involve more causal relations than the variance type study, thus, we would need to aggregate these probabilities. This shows us that, rather than solving the problem of confounders, including more causal relations will ceteris paribus lead to a higher probability of confounders, and thus a higher risk that the study will be lacking IV. However, as Steel argues, it may well be that we have a better understanding of the confounders at the micro level that process tracing operates on. In which case, each proposed causal relation (between X and the intermediate variable and between the intermediate variable and Y) may have a lower probability of being the result of a confounder than in a variance study, which only explores the causal relation between the two variables of interest. For a process study to be superior to a variance-oriented study, the probability of confounders in each causal step need to be sufficiently small so as to outmatch the

fact that there are more ‘steps’ than in a variance study. It is worth noting that process studies may well not include controls for confounders (this would imply a variance study inside the process study) and that the threat of confounders may for this reason be more prominent.

In this section I have argued, in line with Steel, that process studies face the same potential problem of confounders as variance studies. Furthermore, I have tried to flesh out his argument more clearly by reformulating it in probabilities. This shows us that we must have quite good reasons to think that there are no confounders in process studies if they are to fair better than variance studies. This is because 1) the added causal relations that are involved in process studies also leads to added possibilities for confounders and 2) process studies may not include the proper control for these confounders.

The Difference between Variance and Process Studies Illustrated

I have now discussed how variance and process studies differ and relate with regards to the threat of confounders to internal validity. Before this we analysed the threats to external validity of variance studies. In this section, I will illustrate what we have discussed so far with a case study on a social method (not to be confused with the effect-evidence ranking scheme, which is still the overall case in the thesis). This way, the propositions on how to think about validity will become clearer and some more particular concerns about how this applies to social method and social work will be discussed. After the case, I will discuss external validity threats to process studies. I have chosen this order of the sections because the case study can illuminate some of the later discussion on the external validity of process studies. Also, we will see that, just like with internal validity, the threats that pop up are quite similar (if not the same) as those we have already discussed with regards to variance studies.

The social method I will analyse is called ‘Familierådslagning’, or Family Group Conferences (FGC) as it is termed in the Anglo-Saxon literature and by which I shall refer to it. I will specifically look into one study done on this method. I should mention that in the effect-evidence ranking, this method has received the mark of B, despite the fact that its effect has been supported by evidence from a couple of RCTs. This underlines the fact that the effect ranking scheme is not

the only thing that determines what rank a method gets in the overall ranking, but does put a cap on this rank.

The Family Group Conferences method targets children and youngsters for whom there is serious concern regarding their well-being. These children and young people may show behavioural problems such as violent behaviour and may themselves be victims of violence in the family. The idea behind the method is to utilize the knowledge of the extended network of the child, in combination with a social worker. (Mortensen, 2014), thus attempting a more holistic approach than what has usually been the case in child protection services (J. G. Pennell, Burford 2000).

Under the effect dimension of the method by the NBSS, we find two survey articles of the studies of the method that have been conducted (Frost, Abram, & Burgess, 2014a, 2014b). Here we find reference to the study of J. G. Pennell, Burford (2000) as a process study; "Pernell and Burford's study makes a crucial contribution to the debate about FGCs and particularly the discussion about process and outcomes" (Frost et al., 2014b, p.502). The study of Pennell however, includes both what I believe should qualify as a variance study and a process study.

Pennell. et al conduct a study on the effect of using the Family Group Conference method in relation to problems with violence in families against children and adults. The project was conducted in three distinct and culturally different locations in the Canadian province of Newfoundland & Labrador (J. G. Pennell, Burford 2000). Labrador is situated on the east coast of the mainland, whereas Newfoundland is a large island in the Atlantic Ocean. The province also includes a myriad of small islands off the east coast of the mainland. 32 families where chosen to participate in the programme and the study included 384 family members. These families where guided through a five step programme where they would discuss and try to commit to solutions to solve the problems in the family.

The study has two different dimensions: "The study aims to measure outcomes of the project by using follow-up interviews with family group members (progress reports) and by reviewing the child protection services (CPSs) files for the presence of indicators of child abuse (children protection events.)"(Frost et al., 2014b, p.502). First, I should mention that I choose to interpret the 'output' as effect here. This is not a trivial assumption, as output may refer to just comparing before-and-after in the study. Without the assumptions of no confounders of any relevant

magnitude, such a study says nothing about the effect of the method. However, as the researchers have gone to some length in order to control for confounders, I will interpret their outcome study as an attempt to measure the effect of the method.

I will argue that these two different dimensions of the study are tokens of the variance and process type of study, respectively. The variance study is done by comparing the child protection service files of the families that went through the Family Group Conference with a comparison group, consisting of the 31 families. The comparison group just receives the normal services of the child protection agency. The 'treatment group' that went through the Family Group Conference programme was compared to this comparison group for differences in performance with regards to indicators of child abuse. The comparison group thereby functioned as a control for confounding factors, even though this is not explicitly stated, which is trademark of variance studies. A further indicator of this is that the comparison group is composed of families with the same type of problems, children in the same age and same length of involvement with the children protection services as the group that went through the Family Group Conferences programme (J. G. Pennell, Burford 2000). These are exactly the kind of variables that we might expect to affect the families' score in terms of indicators of child abuse, and it would therefore make sense to make sure that the comparison group and the treatment group are similar in these regards. The conclusion of the variance part of the study is that there is a better outcome for the group that went through the Family Group Conference programme in terms of indicators of child abuse than the comparison group. Thus, the study would seem to show that the variance in the effect variable, the level of child abuse (operationalized by child protection events recorded), is due to the variance in the treatment variables: ordinary children protection services compared to the Family Group Conferences method.

The second dimension of the study, which I argue is a process study, consists of a series of interviews conducted with the families that participated in the programme. Of the conclusive remarks concerning the results of the interviews is the following: "A qualitative analysis of the Progress Reports found one overriding reason as to why the conferences left the families better off — they promoted family unity. Irrespective of their home community or their role at the conference, the family group members spoke at length of how the conferences strengthened positive ties among the participants, removed some negative ties, and enhanced their sense of being family." (J. G. Pennell, Burford 2000, p.144).

What the interview helps expose is thus what leads the programme to its supposed positive effect on the families. This is exactly the purpose of the process studies according to (Maxwell, 2004; Pedersen, 2010). In this study, the process is exposed through interviews which I think is quite a common version of process studies within the field of social work, social policy making etc.

In this part of the section, I will discuss the internal validity of the variance and process part of the study to apply the considerations that we have had on these so far.

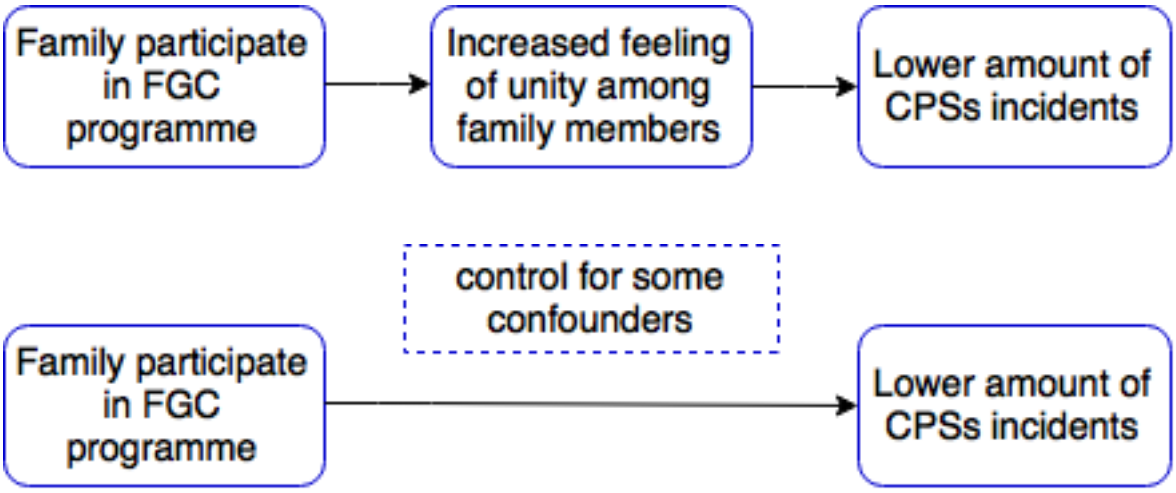


Figure 6

Figure 6 illustrates how the variance and the process type study infer that participation in the Family Group Conferences programme leads to lower amounts of Child Protection Services (CPSs) incidents. I will compare the two parts of the study with regards to confounders and thus internal validity. In the variance study, we have found a correlation between families participating in the FGC programme and lower amounts of CPSs incidents. We have improved the chances of internal validity by controlling for some important confounders via our control group. In practice, the assessment of the threats from confounders must be estimated by our knowledge of what factors typically play a role when it comes to the amount CPSs. Thus, an expert in the area might, from other studies, know that besides controlled for age, type of problems and length of involvement with the children protection services, we should also control for say, what school the child attends, as this may be an important confounder.[23] Therefore, I would argue that to assess the probability of

[23] This is merely an example to illustrate the case, I know of no evidence that this should be true.

a confounder ruining the internal validity of the study, we need the background knowledge of an expert in the field. However, this stands at a somewhat odd relation with a feature of the NBSS evidence ranking. Here, the fact that an expert recommends some method only makes it able to achieve the rank of C. Our case is different in that an expert is not the basis of the study, but is needed to assess its validity. However, this does raise the point that there seems to be no way of getting around expert assessments; the validity of a study is not merely a function of the design and execution of the study itself but also the background knowledge from other studies. This is the knowledge that an expert would have.

There is also the question of whether the internal validity is ruined by *unknown* confounders. This could be the case in the FGC study, even if an expert assesses that we have controlled for all the *known* confounders via the control group. It may seem a bizarre idea that we should be able to assess the probability of confounders that we do not know exist. However, I will claim that 1) we can do this indirectly if we posses good background knowledge, 2) there is no way around doing this when we make decisions. I will explore the last point in more detail in the final section, but I will also briefly discuss this with regards to our case. How could we use background knowledge to assess the probability of an unknown confounder? If we wish to assess the probability of an unknown confounder being the reason we see a correlation between the participation in the FGC programme and lower CPSs, we could look into whether we have experience with other programmes (where we have controlled for *known* confounders) that should have helped to lower CPSs or related variables, but which failed to do so. This would provide prima facie evidence that there is an unknown causal factor at work in the domain, and that we need to take this into consideration when we assess internal validity.

If we now proceed to look at the process study part, this shows a correlation between participating in the programme and experiencing increased family unity and also a correlation between increased family unity and benefitting from the programme. Given these two separate causal claims, the threats of confounders rise for each separate arrow in figure 6, i.e. the two causal relations that the process study rests on. Once again, I argue that the best way of controlling for confounders is through background knowledge on what factors could typically lead to experiencing feeling of unity while correlating with the FGC programme, as well as what could cause lower CPSs and correlates with a feeling of unity.

This is of course made no less complicated by the fact that it seems process studies in social methods often refer to mental states – the process has the form of a treatment, a mental state, and then some behavioural output. In the FGC programme, the metal state is the experience of unity. It is beyond the scope of this thesis to go into a deep discussion of this, but it should be noted that taking mental states as causes or producers of lawful relations is a matter of controversy (see for example Barrett (2006) on whether emotions can be natural kinds or Davidson (1963) for a discussion on reasons as causes).

As discussed before, process studies do not typically include controls for confounders and this is also the case with the process study of the FGC programme. However, it seems intuitively possible that the FGC programme could lead to a feeling of more family unity and that this is turn could lead to less CPSs incidents. Therefore, this study seems to show a trade-off that I mentioned previously, where the additional causal relations that are involved in a process study, compared to a variance study, would ceteris paribus make the threat of confounders more severe for process studies, but background knowledge would suggest that the causal relations of the process studies are likely to be genuine. In assessing whether this is actually the case, expert knowledge is relevant again; if we have knowledge from other studies showing that unity in the family leads to fewer CPSs, and these studies have done a good job of controlling for confounders, this strengthens the internal validity of the process study we are concerned with here. While it is often supposed that process studies can support the causal conclusions made in variance studies, this analysis shows how variance studies can support the conclusions of process studies; by justifying that each causal connection in the process study is genuinely causal and not the result of a confounder.

Summary on Internal Validity Threat to Process Studies

I will now summarize on the discussion of internal validity threats to process studies and the case used to illustrate this. We have discussed how process studies have ceteris paribus a higher risk of the internal validity being spoiled by a confounder because 1) they do not by themselves control for confounders and 2) they involve more causal connections to establish that X causes Y, and each connection will suffer from the threat of confounders.

However, we have also discussed how process studies will often involve variables where we have some background knowledge to justify that there is a causal connection between each link, though it of course varies how certain we are regarding this.

I have used the Family Group Conferences programme to illustrate these considerations. Here, we saw how the variance type study that was done includes control for confounders via a comparison group. In the process study, this was not the case, though it does seem plausible that the FGC could cause more unity in the family, and that this in turn could cause fewer CPSs incidents. I have argued that the best we can do is to use expert knowledge to assess the probability of confounders, both in the variance and the process study.

In relation to the NBSS ranking scheme, the discussion shows that confounders are a threat for both variance and process studies. They each have their own advantages in relation to this threat, as discussed above.

9. External Validity Threats to Process Type Studies

Now that we have found how the internal validity threat of confounders also applies to process studies, I will consider the possible external validity threats to this type of study. As mentioned, many of the points have already been covered indirectly when discussing external validity threats to variance studies. However, now is the time to describe exactly what consequences our previous discussion has for how we should assess the external validity of process studies.

Support Factors for Process Studies

The first possible threat I wish to consider is whether process studies may also fail to have external validity due to the lack of support factors. To analyse this, I will once again cement the distinction I draw between processes study and exploring mechanism components. Process studies are studies that use one of more intermediate variables to show a connection between two variables of interests. This should not be confused with exploring support factors or mechanism components and I have argued previously why this would lead to an infinite regress.

Given this, I will argue that we should pay just as much attention as we would for variance type studies to the threat that support factors might not be present in the target we wish to extrapolate to. I will illustrate this using the FGC programme process study as an example.

Here, the process was described as FGC causing an experience of more family unity, which caused lower amount of CPSs incidents. The threat applies in the following way; perhaps it is only when certain other factors are in place that the FGC will cause an experience of more family unity. And also, perhaps certain factors need to be in place before an experience of unity can cause a lowering of the CPSs incidents. With regards to the FGC causing an experience of unity in the study, perhaps the social workers who executed the programme may have received certain education which was needed for successfully managing this. Or perhaps the meeting took place in close proximity to where the family lived, making it feasible for them to fit the meeting into their schedule without too much trouble. These *may* be the support factors that one should consider before extrapolating the effect of the study to a target, just like we should consider support factors for variance studies.

As discussed before, a process study may have an advantage in terms of external validity, if we have better knowledge of what support factors are needed for the causal relations to hold in the process, than we have by just considering the treatment and the effect variable. Thus, by specifying that the FGC leads to lower CPSs incidents by causing an experience of more unity in the family, we may have a better idea of what support factors are needed. This is the case if we have more inductive knowledge with regards to the conditions under which such a programme can influence an experience of unity, and the conditions under which an experience of unity can influence the problems in a family, than we have on the condition for the FGC programme to influence the problems in a family directly. Thus, it is a matter of our background knowledge. However, as mentioned under the section on External Validity Threats, it seems very likely that we have more knowledge of the support factors needed for the intermediate variable and their relation to the variables of interest (i.e. the treatment and the effect variable) to hold, than we will have if we just consider the variables of interest.

I wish to mention a second reason why processes might fair better in terms of having specified support factors. So far in this thesis, I discussed knowledge of support factors like background knowledge; knowledge that we already have which can help make sure that we do not make wrong

extrapolations. However, by looking into the process through which the treatment variable can affect the effect variable, we may fall upon the relevant, otherwise unknown support factors. Thus, in the example of the Trobriand society, it seems that Malinowski would most likely not have found the relevant support factors, MC1 and MC2, if he had not looked into the process through which more wives can lead to more political influence for the Trobriand chieftains.

I believe this is the reason that Steel confuses a process study for a mechanism (or support factor) account. In practice, we may often come to know about support factors through investigating processes. However I still maintain that we should differ between the two for conceptual clarity, unless we are to fall in pitfalls, like the infinite regress discussed earlier.

I will now summarize what we have found regarding the need for support factors in process studies. Process studies, just like variance studies, need an account of support factors to secure external validity. However, if we know more about the support factors that are needed for the links in the process to hold than we do by only considering the two variables of interest, like we would in a variance study, process study may have an external validity advantage. Also, it may well be through the study of a process that we come to learn what support factors are needed.

More Deep, Less Broad?

The last threat I wish to consider for the external validity of process studies is a claim that Pedersen (2010) makes. He states that process studies may give us a more ‘deep’ but less ‘broad understanding of causal connections, compared to variance studies (Pedersen, 2010, p.13). Pedersen does not describe this thoroughly, but I will now discuss what reasons we have for supposing that this claim holds tight.

One reason for this may be the group (a focus group etc.) used in process studies will be small compared to the group used in a variance study, because of the additional resources it would take to conduct interviews with a larger group. This may mean that the sample studied in a process study is too small to draw any conclusions regarding the population with an acceptable degree of certainty. However, this seems a problem of internal, rather than external validly. Thus, process studies would not be less broad than variance studies, but in danger of stating nothing at all. However, in the FGC

programme study, 115 interviews where conducted (J. Pennell & Anderson, 2005), so it is certainly not impossible to have rather large samples for a process study.

Another reason why writers like Pedersen may think that process studies are less broad is that they do not, by themselves, control for confounders. This is however again an internal validity problem that does not concern how 'broad' the study can be applied; if the result of the study turns out to be due to a confounder, while we think it is the result of the causal factor we study, this implies zero breadth; the conclusion of the study is not justified in the object of study, let only any target.

One way in which Petersen's claim can be justified is that a process study is usually only concerned with one 'causal path' connecting the variables of interest. Steel (2008, chapter 9) has made this observation that I will be drawing on and explain in relation to the case.

I will use the example of how lowering the pupil/teacher ratio might lead to better academic performances that we discussed with relation to Cartwright. Lowering the pupil/teacher ratio may increase the reading skills of the pupil because the teacher has more time to take care of each of the pupil's academic problems. But it may also happen through the increased attention the teacher is able to give the students, which might also raise their motivation, causing them to do more home-work and this in turn might increase their reading skills. Here, we see two causal paths connecting a lower pupil/teacher ratio to better reading skills among the students. However, a process study may only take one of these paths into consideration. If we had made a good variance study of the connection between the two variables, controlling for confounders, this would show the effect of the lower ratio on reading scores for *all* the causal paths connecting them. However, if we make a process study that only focuses on one of these paths, we may underestimate the influence of lowering the ratio.

Note that due to our outline in the Different Needs to Know section, the example I mentioned above is actually not that bad for the validity of the study. This is because we are interested in whether or not the study allows us to infer that the social method will make any *positive* contribution. If this happens through more causal paths than we considered initially though, than this is no threat to the conclusion that the method will give a positive contribution. However, consider the figure below.

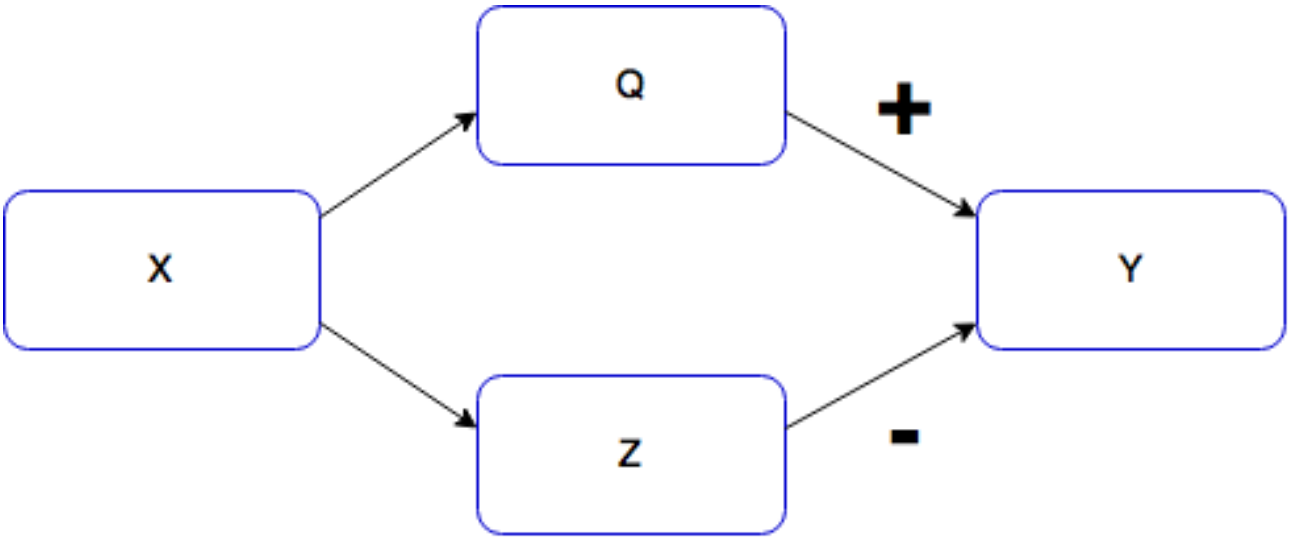


Figure 7

The situation in figure 7 is direr. Here, X causes both Q and Z, with Z having a negative influence upon Y. If we only look at the process from X to Y through Q we may wrongly infer that X will make a positive contribution to Y, while it may not, due to X also causing Z. However, if we had made a variance study of X and Y, we would have caught this double influence. In this way, things are actually quite contrary to what Pedersen suggest: "Fordelen ved en procesorienteret strategi er, at evalueringen giver bedre viden om virksomme komponenter i indsatsen, men kan potentielt også give viden om bi-effekter (positive såvel som negative utilsigtede effekter), hvilket kan give et bedre grundlag at rådgive ud fra, når det handler om at forbedre en indsats"[24] (Pedersen, 2010). Pedersen is right that process studies may reveal what intermediate variables that lead to negative outcomes, however, when it comes to validity, process studies seem disadvantaged compared to variance studies, if they only consider one causal path.

Based on this, I believe that that we should be considering whether there are other causal paths that might lead to negate the positive contribution that the process study might suggest. This problem seems not to concern variance study in the same way. However, in practice I believe that variance studies may suffer from a very similar problem! Consider the illustration below.

[24] Own translation: "The advantage of a process oriented strategy is that the evaluation gives better knowledge with regards to active components in the treatment but can also potentially give knowledge regarding the side effects of the treatment (positive as well as negative), which can provide a better foundation for concealing when it comes to improving the treatment."

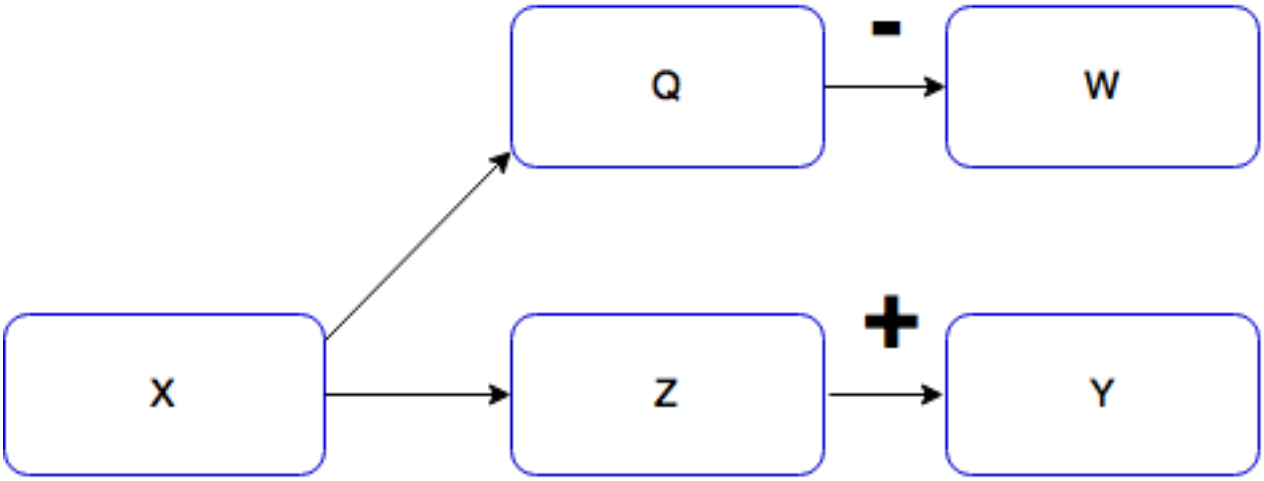


Figure 8

Here, there is only one causal path from X to Z, but X also has the effect Q which has a negative contribution to W. If W is a variable that we do not want to be exposed to a negative influence, then this is problematic. Also, a variance study of X and Y will not catch this side effect. This leads credence to an argument made by Cartwright and Hardie (2012, p.68-69), urging us to look into the causal role a treatment can have for other things that we care about, besides the effect variable. Variance studies are no exception here.

The conclusive remark on this is that we should make sure to look into other causal pathways than the ones our studies are concerned with, both when it comes to variance and process studies. Steel's concern about the fact that we only look into one causal path through process studies is warranted, as we may wrongly infer that a positive contribution will be made by the treatment, when it will not. This is not the case with variance studies, but here there are negative side effects on other variables that we care about. This does neither strictly speak against the external nor the internal validity of the variance study, but appears just as important as a practical problem.

Thus, to return to our original question, whether Pedersen was right in asserting that process studies are more 'deep' than variance studies which on the other hand are more 'broad.' The only defendable interpretation I find of this is, as Steel have noted and as I have tried to explain above, that variance studies will encapsulate all the causal paths that leads from X to Y, whereas process

studies only encapsulate one. Though Pedersen's remark certainly seemed a threat to external validity to start with, I believe that what we have now found out is that this may actually be a threat to both internal validity and/or external validity. It can be an internal validity threat if we conclude that in the object of study, X made a positive contribution to Y through Q, while in fact this positive contribution was negated by X causing Z, which made a negative contribution to Y (Figure 7). However, it may be problem of external validity if, in the target, there will be a casual path between the treatment variable and the effect variable, which means that the positive contribution we found in the study will be negated in the target. This may happen even if there was no causal pathway that negated the positive contribution in the study. This is because there may be support factors in the target that were not present in the object of study, which makes the negative causal path possible.

Summary on External Validity Threats to Process Studies

This concludes the discussion on external validity threats to process studies. As found in the discussion on the internal validity of process studies, the same threats apply to process and variance studies, though presenting in a slightly different form. Thus, I have discussed why process studies also do need to take support factors into consideration, even though they, under certain circumstances, will be less threatened by this than variance studies. Secondly, I have discussed how multiple causal paths may create validity problems for process studies that variance studies do not appear to suffer from. However, variance studies are not safe from the grim threat that the treatment variable may influence other variables that we care about, besides the effect variable.

10. The Threats Aggregated and Conclusive Remarks

In this final section, I will summarize on the threats to internal and external validity that we discussed and evaluate the effect-evidence ranking scheme that the NBSS has made. As mentioned, in the scheme a social method can be given the rank of A only if a variance study has been conducted while a process study only makes it possible for a social method to score a B.

One of our findings is that the same three types of threats are present both in variance and in process studies. I will here summarize these:

Confounders were discussed as a threat to internal validity. Variance studies can through various means of control try to rule out the influence of confounders – through the discussion of random control trials, I concluded that the debate seemed to show that it was unlikely that variance studies can control for unknown confounders, however. We also concluded that because process studies do not by themselves control for confounders, and because they involve more 'joints' to establish their conclusion, they seem more at risk regarding confounders. However, we also discussed how process studies, by involving an intermediate variable, might postulate relations where we can better utilize our background knowledge to check for these confounders. We seem, for example, to have more knowledge about what social customs lead people to do, than the relation between wives and political power directly, as we explored in Trobriand society example.

The threat of lack of support factors, or the danger regarding different mechanisms in the target and in the object of study, was discussed as a threat to external validity. This presents a threat to external validity, as the target may not have the support factors, which I have argued is same as mechanism components, needed for the conclusion that we found in the object of study to hold. Also, I argued that contrary to Steel's claim, accounting for a process by which X causes Y is neither sufficient nor necessary for an account of mechanism components. Also, if we confuse the two together we run into various problems such as the danger of an infinite regress. We did conclude however, that if studying the intermediate factors between the treatment and the outcome makes us better able to specify the support factors, then process studies might have an edge over variance studies with regards to external validity.

Lastly we looked into the threat of multiple causal paths. This is a threat to process studies, because even if such a study shows a treatment to have a positive contribution on the effect variable, there may be other causal paths from the treatment variable leading to a negative contribution to the effect variable. This does not in the same way appear to be a threat to variance studies, as they try to measure the relation between the treatment and the outcome directly, and should thus 'pick up' all the paths through which the treatment affects the effect variable. However, even though this is not strictly a threat to validity for variance studies, we are in practice only interested in how the treatment will affect other variables which we care about. That is, the treatment variable may affect other things in a negative way, even if it makes a positive contribution to the effect variable.

Going back to the start of the thesis, I argued that the effect ranking dimension was concerned with certainty – and that the most forward approach to operationalizing certainty and uncertainty was through probabilities. Given this, it seems that the most general formulation of our findings in the thesis is that what should be ranked the highest depends on what gives the lowest probability of failure, when assessing this probability for these three threats. Or, the reverse way of putting this, whether or not a study implies a low or high degree of certainty depends on how high the aggregated probability of validity is, when taking these three threats into consideration.

I will express this summarization in the equation below. This will form the basis of my recommendation for how the NBSS ought to rank studies. When ranking studies, we should maximize P, the probability of validity. In the equation, P(¬c) is the probability that no confounders have spoiled the internal validity, P(sf) is the probability that there are the right support factors in the target, while P(¬cp) is the probability that there are no alternate causal paths that would spoil any positive contribution found in the study. As discussed, this is only relevant for process studies. However, for variance studies we should still consider the probability of any causal paths that might negatively influence other variables that we are interested in, to such a degree that we would not consider it worth the positive influence on our effect variable.

$$Max(P)= P(\neg c) *P(sf) * P(\neg cp)$$

Just formulated in this general manner this generally, this equation may not seem very informative. However, I believe it does show us two things. First, what are (at least some of) the important things we should consider when judging the validity of a study. I do not believe this was obvious in the beginning of this analysis. Secondly, it shows us that process studies may well be superior to variance studies; it depends on how the study fair on these three parameters. While it may seem trivial that the scheme does not get it right every time, I think that our consideration in the analysis show that it is not only in the marginal case that process type studies may do it better. Thus, we discussed why process studies may well have a higher probability of getting the support factors

right, thus getting a higher P(sf) value. Due to this, the effect-evidence ranking appears unsatisfactory.

Before moving on to how this equation may inform how the NBSS should rank studies, I need to comment on an important implication of the equation. The equation implies that the probability of no confounders spoiling the internal validity is independent of the probabilities of external validity. Thus, it implies a stand on whether there is a systematic trade-off between internal and external validity; namely that the systematic trade-off is not there. This may strike the reader as weird, given the discussion of Cartwright's argument for a trade-off. However, what I believe this argument shows is that there *may* be a trade-off between the two and I think that it is an empirical question from case to case whether this is trade-off occurs. Thus, I would argue that the equation should be altered if we could make solid empirical argument that this trade-off is indeed systematic. In any case, the equation as it is now should stand as an *approximation* of how valid a study is.

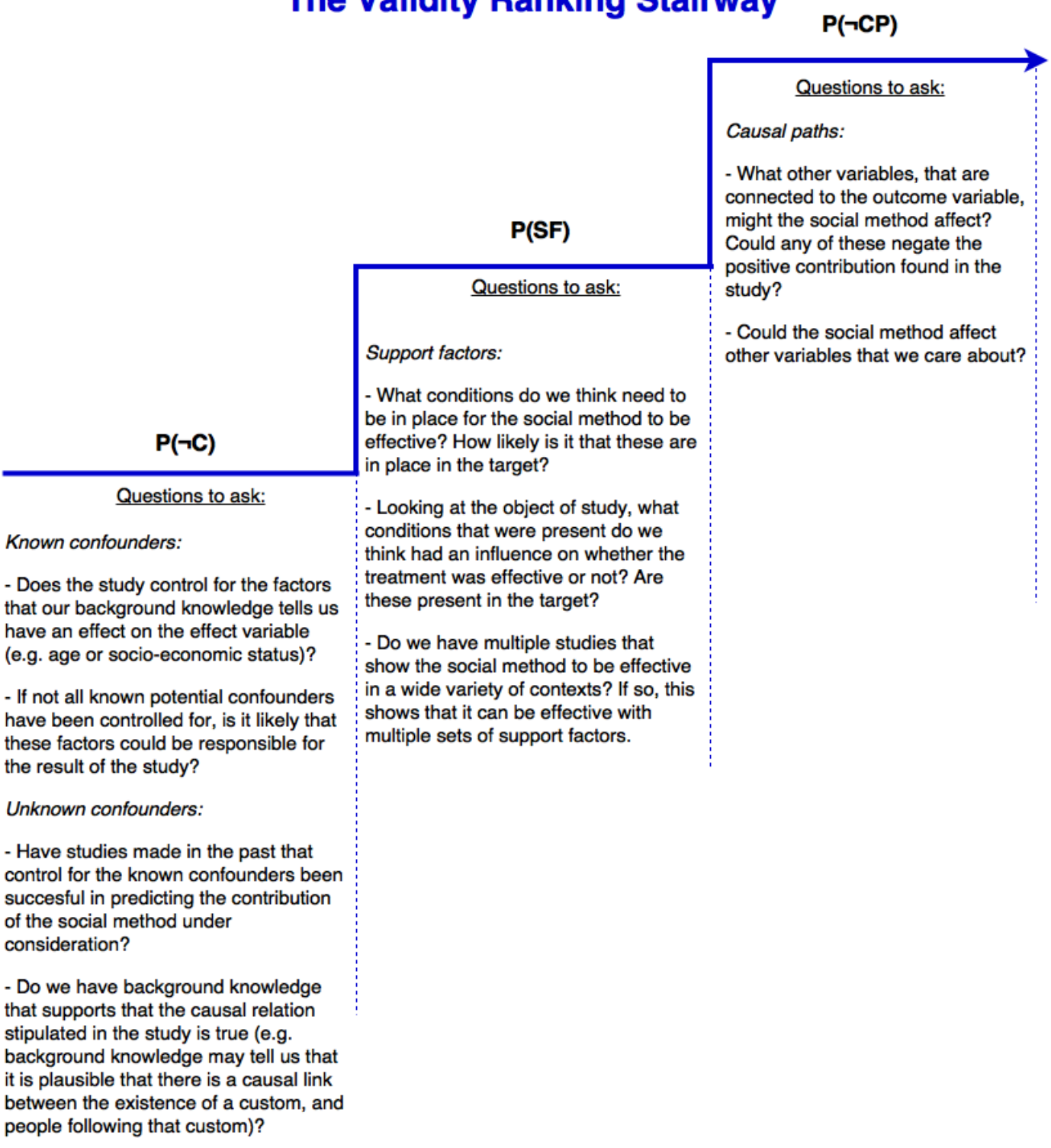
The Validity Ranking Stairway

The question then becomes how the employees of NBSS can apply the findings in this analyses that I have boiled down in the equation. Despite all its shortcomings in terms of rigor, the ranking scheme certainly provides an easy and practical tool for assessing the validity of studies.

In this last very last section of the thesis, I will try to develop an outline of a model that could be used by the NBSS to assess the validity of a study, given the analysis in the thesis. I will give a presentation of the model and thereafter give a brief defense of the fact that it requires the NBSS rankers to assign probabilities.

There are two things I need to say before going into developing the model. First, that the model, like the ranking scheme, only considers one study and not how all the different studies would add up in terms of validity. Secondly, the model will only deal with the threats to validity that I have analyzed in this paper –as I also outlined in the start of the paper, there are naturally important threats to validity that should be informed by statistical methodology (not having a large enough sample size etc.). Thus, this model should be supplemented by considerations of these possible problems as well. I have called the model The Validity Ranking Stairway and it is presented below.

The Validity Ranking Stairway



The Validity Ranking Stairway is an illustration of the ranking process that I wish to suggest as an alternative to the current one used by the NBSS. It can be used in two ways; one that is quite simple and another that is a bit more 'rigid' but may save time and resources. In the simple way, the NBSS employees go through each of the three steps and assign a probability to each of the parameters. The questions below each step are meant as a starting point for how to consider each parameter, and are inspired by the analysis done previously in thesis on the different threats and how to interpret them. After having assigned a value to each of the parameters, the total probability could simply be aggregated as shown in the equation I presented (thus, simply multiply the three parameters). Note that while probabilities are normally only restricted in the sense of having to be between 1 and 0, the NBSS might dictate that only certain values should be chosen (70 %, 80 % etc.). This may make more sense, since we will most likely not have any relevant information that would make it meaningful to differ between being 93% and 92% certain that there are no relevant confounders. Instead of using the aggregated probability as the ranking in itself, the NBSS might choose to assign each of their current rankess (A, B , C) to an interval of the aggregated probability. For example, an aggregated probability of <.6 might correspond to a C, .7-.8 to a B and >.8 to an A.

Now that we understand the basics of the Validity Ranking Stairway, I want to specify what our analysis in the chapter before says about how we should use the Evidence Ranking Stairway, before moving on to the more sophisticated way of using it.

- First, I have included unknown confounders under the first step as our discussion suggested that RCTs can provide no guarantee against unknown confounders. To access the probability of there being an unknown confounder is of course very hard, but I believe we may try to do so by looking at the predictive track record of the controlled studies we have done so far (thus, the first question under the unknown confounders box). The idea is that if it turns out some social method fails in providing the positive contribution that we thought it would, based on a study, *one* explanation of this is that there was an unknown confounder that ruined the internal validity of the study. Of course, it will still be difficult to assign a probability, but as I will defend later, there is no better alternative. We will have to assess the probability of the result of the study not being infected by a confounder both for process and variance studies, as discussed. Variance studies will often have controls for

confounders, while process studies might stipulate causal connections that our background knowledge can underpin.

- Secondly, as discussed many times, it was found that discovering an intermediate variable (thus, making a process study) is not in itself sufficient to have the support factors – we will have to assess how sure we are of what the important support factors are there anyways. As we have also discussed however, process studies may well fair better than variance studies in terms of support factors if we have better knowledge of what support factor are needed between each causal link in the chain, which we often have.

- Thirdly, the external validity of a study is always relative to the target we want to extrapolate to. Therefore, the ranking should either be target specific (in practice, specify what municipalities the ranking is related to) or it should state that it express the probability that the treatment will be effective in any given municipality.

With these specifications in mind, I want to describe the more sophisticated approach to using the Evidence Ranking Stairway. This approach utilizes the fact discussed earlier that (at least for practical purposes) before external validity is relevant, we need to have internal validity. Also, in order for us to consider causal paths we should ensure external validity; it does not seem to make much sense to consider what other variables a treatment might affect if we have not established that the effect in the study will actually pertain in the target.

Therefore, under this approach, we will set a bar of certainty (i.e. a minimum probability) for each 'step' on the stair. Thus, before considering moving on to the external validity step, we might say that we should as a minimum assess that we are 70 % sure that the internal validly of the study is not ruined by confounders. Then we might also set such a bar for the probability that the right support factors are present in the target before considering causal paths.

The advantage of this approach is that it may save us time in terms of assessing probabilities for support factors and causal paths, if the probability that the study has internal validity is so low that we find the study too uncertain anyways. It is very important that we should not interpret this

approach in a manner where each ‘step’ that the study passes gives it a notch up in the ranking (e.g. achieving a C if it passes the internal validity bar, a B if it passes both the internal and external validity bar etc.). Internal validity without external validity, when we wish to extrapolate the results of a study, does not get us halfway, it will get us no way at all. Therefore, even if we set a minimum bar, I think the best course of action will still be to aggregate the probabilities when we make the final ranking, like we would also do with the simple approach.

A Brief Defense of the use of Probabilities

To end the thesis, I will make a short note on the fact that the Validity Ranking Stairway requires that the employees of the NBSS assign probabilities. The idea of assigning probabilities, as a way of dealing with uncertainty is likely the most orthodox, but can also be criticized in a couple of ways. I will briefly go through some objections to assigning probabilities.

First, it may be claimed that it is simply not possible to assign probabilities in many cases. This is often the case if we are talking about questions where we have very little knowledge or which appear abstract (I would personally find it hard to assess the probability of an economic recession beginning tomorrow, for example). The question is, would it also be impossible to assign the probability of a study’s internal validity being ruined by a confounder, or that the target may not have the proper support factors? The questions in the model above are precisely made to help in this matter. They are meant as mental teasers for bringing out the knowledge that we have on these question.

Also, it seems to me that there is simply no way around trying to assess probabilities; it is the best we can do. The threats to validity do not disappear simply because we stay silent on them and if we need to make a decision based on the possible validity of the study, I see no alternative to probabilities. I agree that it may seem a tall order that we should assess the probability of, for example, no confounders ruining the internal validity of a study, but if we choose to just ignore this problem, we will end up acting as if the probability of this is zero, which seems more epistemically irresponsible than trying to give a probability assessment.

Another objection to requiring the NBSS to assign probabilities is that it might leave to too much room for subjective assessment; there may be very different opinions about what probabilities should be assigned. However, once again, I believe that this is better than our other opinions, which is something like the current NBSS evidence ranking scheme. This leaves no room for subjective assessment, but as I have sought to show, it is crude and also will get a lot of validity judgments wrong. Furthermore, as mentioned, the NBSS also uses other criteria to judge what effect-evidence ranking a study should have, besides the evidence ranking scheme. This leaves ample room for subjective judgment anyways. By having to assign a probability, disagreements on validity can be brought to light and discussed. A probability assessment can provide a cornerstone around which such a discussion can be based.

Lastly, some might object that a probability assignment is way too fine-grained to be meaningful, given the knowledge we have about the threat of confounders etc. However, as also mentioned above the NBSS can dictate certain values that should be assigned so that one avoids meaningless nuances like the difference between a 0.93 and 0.94 probability of a study not being ruined by a confounder.

In summary, while assigning probabilities to these parameters may be challenging, I believe the two only alternatives are to outright ignore them (which in practice is the same as assigning them a probability of zero) or make a very crude ranking system like the one that the NBSS currently employs.

These remarks conclude the thesis. Motivated by the NBSS effect-evidence ranking scheme, I have sought to cover and discuss some of the main threats to internal and external validity that have been identified in the philosophy of science literature. In the above section I have brought these considerations together and formed an outline to an alternative model.

Bibliography

Alexandrova, A. (2008). Making Models Count. *Philosophy of Science, 75*(3), 383-404.

Ankersborg, V. (2007). *Kildekritik - i et samfundsvidenskabeligt perspektiv* (Vol. 1). Gylling: Samfundslitteratur

Barrett, L. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science, 1*.

Busck, S., Rågård, J., Rasmussen, C., P. (2008). *Kildekritisk Tekstsamling*

Cartwright, N. (1983). *How the laws of physics lie*. Oxford, New York: Clarendon Press; Oxford University Press.

Cartwright, N. (1999a). *The dappled world : a study of the boundaries of science*. Cambridge, UK New York, NY: Cambridge University Press.

Cartwright, N. (1999b). The Limits of Exact Science, from Economics to Physics *Perspectives on Science, 7*(3), 318-336.

Cartwright, N. (2007). Are RCTs the Gold Standard? *BioSocieties,* (2), 11-20.

Cartwright, N. (2007a). *Hunting causes and using them : approaches in philosophy and economics*. Cambridge ; New York: Cambridge University Press.

Cartwright, N. (2007b). *The Vanity of Rigour in Economics: Theoretical Models and Galiliean Experiments* Discussion Paper. Centre for Philosophy of Natural and Social Science. London School of Economics and Political Science.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy : a practical guide to doing it better*. Oxford ; New York: Oxford University Press.

Colyvan, M. (2008). Is probability the only coherent approach to uncertainty? *Risk Anal, 28*(3), 645-652. doi:10.1111/j.1539-6924.2008.01058.x

Cook, T. D. C., D.T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally and Company.

Davidson, D. (1963). Actions, Reasons and Causes. *Journal of Philosophy, 60*, 691-703.

Frost, N., Abram, F., & Burgess, H. (2014a). Family group conferences: context, process and ways forward. *Child & Family Social Work, 19*(4), 480-490. doi:10.1111/cfs.12047

Frost, N., Abram, F., & Burgess, H. (2014b). Family group conferences: evidence, outcomes and future research. *Child & Family Social Work, 19*(4), 501-507. doi:10.1111/cfs.12049

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, Mass.: MIT Press.

Guala, F. (2003). Experimental Localism and External Validity *Philosophy of Science, 70*, 1195-1205.

Hausman, A., Kahane, H., & Tidman, P. (2013). *Logic and Philosophy : a modern introduction* (12th, instructor's ed.). Australia ; Boston, MA: Wadsworth, Cengage Learning.

Howson, C., & Urbach, P. (1993). *Scientific reasoning : the Bayesian approach* (2nd ed.). Chicago: Open Court.

Howson, C., & Urbach, P. (2005). *Scientific reasoning : the Bayesian approach* (3rd ed.). Chicago: Open Court.

Jiménez-Buedo, M., Miller, L. M. (2010). Why a Trade-off? The Relationship between the External and Internal Validity in Experiments. *THEORIA, 69*, 301-321.

Kincaid, H. (1990). Molecular Biology and the Unity of Science *Philosophy of Science*(57), 573-593.

La Caze, A. (2013). Why Randomized Interventional Studies *Journal of Medicine and Philosophy, 38*, 352-368.

La Caze, A., Djulbegovic, B., & Senn, S. (2011). What does randomization achieve? *Evidence-Based Medicine*.
Little, D. (1991). *Varieties of social explanation : an introduction to the philosophy of social science*. Boulder: Westview Press.
Maxwell, J. A. (2004). Using Qualitative Methods for Causal Explanation *Field Methods, 16*, 243-264.
Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
Morgan, M. G., Henrion, M., & Small, M. (1990). *Uncertainty : a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge ; New York: Cambridge University Press.
Mortensen, B. S., Jane. (2014). Familierådslagningen, URL: http://vidensportal.dk/temaer/styring-og-sagsbehandling/indsatser/familieradslagning, access: 07/04/16
Mäki, U. (2012). *Philosophy of Economics*. Amsterdam: North Holland is an imprint of Elsevier.
Pedersen, C. S. (2010). *Overordnede strategier til effektmåling* Økonomistyrelsen.
Pennell, J., & Anderson, G. R. (2005). *Widening the circle : the practice and evaluation of family group conferencing with children, youths, and their families*. Washington, DC: NASW Press, National Association of Social Workers.
Pennell, J. G., Burford (2000). Family Group Decision Making: Protecting Children and Women *Child Welfare, 79*(2), 131.158.
Resnik, M. D. (1987). *Choices : an introduction to decision theory*. Minneapolis: University of Minnesota Press.
Socialstyrelsen. (2012). *Vidensdeklaration*. Odense Socialstyrelsen.
Socialstyrelsen. (2013). *Viden til gavn - Politik for udvikling og anvendelse af evidens* Odense: Socialstyrelsen
Steel, D. (2008). *Across the boundaries : extrapolation in biology and social science*. Oxford ; New York: Oxford University Press.
Thye, S. R. (2000). Reliability in Experimental Sociology. *Social Forces*(78), 1277-1309.
Urbach, P. (1985). Randomization and the Design of Experiments *Philosophy of Science, 52*, 256-273.
Worrall, J. (2002). *What* Evidence in Evidence-Based Medicin *Philosophy of Science, 69*, 316-330.
Worrall, J. (2007). Why There's No Cause to Randomize. *The British Journal for the Philosophy of Science, 58*(3), 451-488. doi:10.1093/bjps/axm024