SYMPOSIUM ABSTRACT – EXTENDED VERSION:[1]

REFLECTING ON NANCY CARTWRIGHT'S CONCEPTION OF THE EXTERNAL VALIDITY OF RANDOMISED CONTROLLED TRIALS

By Mikkel Helding Vembye[2]

Introduction

This symposium paper aims to discuss Nancy Cartwright's criticism against randomised controlled trials (hereafter RCTs), and especially *the use* of the knowledge extrapolated from RCTs. The primary focus of this paper will be on Cartwright's perception of the *external validity* of RCTs. The paper aims in particular to discuss the prospects and boundaries of the generalisability of RCTs. In this regard, *external validity*, sometimes described as *extrapolation* (Reiss, 2018, p. 5), refers to the notion that "*random* selection … permits you to generalize your findings valid back to that population" (Murnane & Willett, 2010, p. 44) or in other words "whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables" (Cook, Campbell, & Shadish, 2002, p. 38). Furthermore, I will try to extend and relate these views and critiques to meta-analysis[3] and systematic reviews[4] applied in evidence-based research.

Throughout the presentation, I will show the significance of Cartwright's theorisation in relation to the evidence-based movement (Cartwright & Hardie, 2012), in particular within the field of Nordic educational research and policy-making. In that context, I will demonstrate how the theory injects shining nuances into fields which predominately focus on hunting/testing causes and less on the use of that hunting. However, I will try philosophically to challenge some of the theoretical conceptions of the theory, especially the practical use of Cartwright's conception of the INUS conditions.[5]

I want to direct my attention towards meta-analysis and systematic reviews for at least two reasons. First, because it seems to be the case that "[s]ystematic reviews are the primary method for managing knowledge in the evidence movement approach" (Oakley, 2003, p. 23), and I consider RCTs to be a key part of this movement. Second, many statisticians recognise that meta-analysis of RCTs possibly transcend the deficiencies of individual RCT studies. John Ioannidis argues, for instance, that "research findings are more likely true in confirmatory designs, such as large (…) randomized controlled trials, or meta-analyses thereof" (2005, p. 698), and, "[b]etter powered evidence, e.g., large studies or low-bias meta-analyses, may help, as it comes closer to

[1] Many thanks to Hans Siggaard Jensen, Christian Kjeldsen, Henrik Kasch, Magnus Birkemose Nordam, and Stacey Marie Cozart for valuable comments on the extended version of the abstract.

[2] Danish School of Education, Aarhus University, Denmark. Email: mihv@edu.au.dk

[3] Meta-analysis "refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the finding" (Glass, 1976, p. 3). Meta-analysis "refer specifically to statistical analysis in research synthesis and not to the entire enterprise of research synthesis" (Harris, Hedges, & Valentine, 2009, pp. 6–7).

[4] "[S]ystematic reviews is a review of existing research using explicit, accountable rigorous research methods" (Gough, Oliver, & Thomas, 2017, p. 4).

[5] "An INUS condition is an ***I***nsufficient but ***N***ecessary part of an ***U***nnecessary but ***S***ufficient condition for producing a contribution to the effect" (Nancy Cartwright & Hardie, 2012, p. 25; Mackie, 1965, p. 24).

the unknown "gold" standard" (2005, p. 700). On related matters, Rex Kline argues that "our data routinely lie, they lie through multiple types of error, and it is only when results are averaged over studies, such as in the technique of meta-analysis, that some of these errors begin to cancel out" (2015, p. 56). It seems to me that many researchers actually reflect on the shortcomings of single RCTs and single studies but believe that the generalizability/external validity increases when research findings are pooled together. As an example, the Danish Clearinghouse for Educational Research (DCER) states in one of their systematic reviews that "a big part of the included reviews is based on meta-analysis, which implies generalizability in an *absolute sense*. The findings of these reviews reach (…) beyond the context, because in a meta-analysis other conditions and factors have been controlled for" [my italics and translation] (Nielsen, Tiftikci, & Larsen, 2013, p. 85). Consequently, DCER conclude that meta-analytical findings are "*directly* applicable for practice" [italics mine] (Dyssegaard, Larsen, & Tiftikçi, 2013, p. 12). In a Danish context, the trust in the enumerations of meta-analysis is often based on the belief in Austin Bradford Hill's (1965, p. 296) criteria/notion of *consistency*: if a relation is "repeatedly observed by different persons, in different places, circumstances and times," then we have a generalizable causal claim. My claim is that this might be one of the reasons why we find an almost religious[6] trust in (meta-)meta-analytical works such as John Hattie's Visible Learning (2009).

Although I believe that well-conducted meta-analysis and systematic reviews are able to cast light on the diverse efficacy and effectiveness of given interventions and practices, I will argue that the pooled result does not bring external validity simply because the pooled results can at best show us what has worked in a past states. It cannot tell us anything about the current state of the target population. Unless we believe in simple inductive reasoning, we need additional information and background knowledge about the target population besides the findings of meta-studies when we want to intervene on a policy-level or local level. Therefore, I will argue that (and explore if) the lesson Cartwright teaches us goes beyond the critique of single RCTs. It is my impression that many researchers wrongly ignore Cartwright's theory due to the rigorous trust in meta-analysis and systematic reviews.

Finally, I intend to question Cartwright's statement that "we do not have good explicit methodologies for how to establish *tendency claims*" [my italics] (2011, p. 767). In that regard, I will probe the question whether well-conducted meta-analysis and systematic reviews are actually able to illuminate the concept of *stable tendencies* in the given causal semantics/model across different contexts, or, on the contrary, indicate whether stable tendencies are absent when we try to apply causal models widely? For my argument, I will use an extended version of Cartwright's causal principles (CP).

[6] For examples of the religious interpretation of the book, see Mansell (2008) stating that the book "reveals teaching's *Hole grail*" and Egelund & Qvortrup (2014, p. 13) saying that, "the book should be a *bible* for teachers" [my translation].

The significance of Cartwright's theorisation

In the field of (Nordic) educational science – which is my main scientific theoretical field of interest – we are currently experiencing an extensive and increased use of RCTs. Since 2011, at least 39 new large-scale RCTs have been initiated in the Nordic countries within economics, political science and educational science (Pontoppidan et al., 2018, p. 325). Most RCTs are publicly funded – simply because policy-makers believe and argue that RCTs *by design* yield the highest degree of generalisability on whether given interventions should be disseminated (for these arguments, see the Danish Ministry of Education, 2014, p. 5). Without further hesitation, policy-makers of the Nordic countries in most cases – if the findings verify the political agenda – use the findings of the above-mentioned RCTs as key evidence for amendments of the Nordic educational legislation. I, therefore, think it is reasonable to argue that in the Nordic countries we experience a quite rigorous belief in what Julian Reiss characterises as *a foundationalist methodology*, which "maintains that certain methods are *intrinsically* more reliable than others" [my italics] (2018, p. 10). This belief is intimately connected to the *liberal form of experimentalism*, where RCTs are recognised as the *golden standard* and all other methods should "be ranked with respect to reliability according to how closely they mimic experiments" (Reiss, 2018, p. 9). Cartwright et al. have tirelessly contested and opposed the widely accepted conviction that RCTs (should) work as *the* "golden standard" and are the best method to ensure *internal* and *external validity*. Therefore, Cartwright is a significant voice against this kind of *methodological foundationalism*. Furthermore, Cartwright's theories have turned out to be very powerful for illuminating the problems at stake in the current movement of evidence-based education in the Nordic countries (Kvernbekk, 2016; Vembye & Jensen, 2018).

Arguing against the external validity of RCTs, Cartwright and Hardie (C&H) show that all study populations (SP) of RCTs are governed by causal principles (CP) (1):

$$\text{(CP) SP: } y(i)c = a_1 + a_2 y_0(i) + a_3 b(i)x(i) + a_4 z(i) \qquad (1)$$

"*i's* range over the individuals in the population to which the causal principle applies, *y(i)* is the outcome, *x* is the policy variable, *a's* are constant across all individuals, $a_{2\ldots n}$ represent "boost factors", y_0 is a "base level" of *y* for *i*, *b(i)* represents all the different factors in all the support teams that work with the *x*, *z(i)* represents all other factors and their support teams that contribute additively with *x* but do not include *x*" (2012, pp. 26–27).

According to C&H, the anticipation of what knowledge RCTs provide often excludes information of the above-mentioned other causally structurally relevant factors (1) (2012, p. 27). This is often due to researchers expecting that *randomisation* assures that "*x* is probabilistically independent of y_0, *b*, and *z*" (2) (2012, p. 35), and therefore that the given cause/intervention, *x*, has worked with in many different structural factors contained in the given study population, which in turn makes the result highly generalizable. This often results in the widespread belief that "*ideal" RCTs* simply reflect the difference between receiving the treatment versus not receiving the treatment (2).

$$T = a_3 Exp(b(i)(X - X') \quad (2)$$

T signifies the treatment effect, "X is the value of the treatment in the treatment group and X' is its value in the control group" (2012, p. 34).

When one wants to apply the knowledge from an RCT, C&H argue that one cannot simply expect all these factors and especially not the support factor, *b(i)*, to be the same or to be equally distributed ($Prob_{SP}(b(i)=B = Prob_{TP}(b(i)=B)$[7] across different populations or even between the study population and the defined population. C&H argue that wanting to apply the knowledge from an RCT to a target population (TP) requires external knowledge about the structural factors that govern TP. Equation (3) reflects this condition.

$$\text{TP: } y(i)c = a'_1 + a'_2 y'_0 + a'_3 b'(i)x(i) + a'_4 z'(i) \quad (3)$$

a´ denotes that one does not know the relation between *a* and *a´*. More decisive, we are rarely in a position in the social domain to expect that TP contains the same support factors as in SP. Cartwright thereby shows us that it is much more complicated to expect that the claims *what-works-there = it-will-work-here*. What Cartwright concisely shows is that "[d]ifferent underlying structures yield different causal and probabilistic relations" (Cartwright & Munro, 2010, p. 210). Therefore, Cartwright actually thinks that the idea of external validity is the wrong idea, because for it to be valid it requires that the causal principles of SP is equal to the causal principles governing TP. According to Cartwright, this is a rare case in (the social) reality. Another way that Cartwright (Cartwright & Efstathiou, 2009) expresses her point is to show that CP of the population possibly changes every time one crosses different space/context (Σ) and times (T). This means that one needs to draw new/different conclusions when one crosses these spatiotemporal points (see table 1). Many think of external validity as the concept of generalising the results from the study population valid back to the target population across settings (Σ). But that is where Cartwright show us that it is all-important to take *time* into account as well. As Cartwright (2011, p. 751) shows:

> Even if the entire target population were enrolled in the study, predictions will be about future effectiveness where there may be no guarantee that this population stays the same over time with respect to the causally relevant factors

A different way to put this is to say that every time we move from Σ1,T1 to Σ1,T2 (see table 1), we have to be aware of structural variances that might affect the result of the intervention, although one often intuitively think that the context is similar, for instance when doing research at fourth grade students across different year groups in the same country.

[7] SP= study population, TP= target population

Table 1. – *context and time*

	Time (T)	
Space/context (Σ)	Σ1,T1	Σ2,T1
	Σ1,T2	Σ2,T2

The last point I will draw attention to in this paragraph is Cartwright's argument regarding the use of the findings of the average treatment effect[8] (ATE) deduced from RCTs. To know that an intervention has produced a positive ATE is not a deterministic concept, Cartwright argues (Cartwright, 2007a; Deaton & Cartwright, 2018, p. 15). This means that 'X causes Y in φ[9]' is consistent with 'X causes ¬Y in φ' (Cartwright, 2007a, p. 15). I think, this statement has a significant practical impact, because it makes practitioners like teachers and doctors aware of that they cannot inductively expect ATE to equal the concrete case. Raising the mean can be an important matter for policy-makers who work with larger systems/populations, but not knowing the mechanism behind given ATEs can actually 'harm' people or have no effect, when, for instance, teachers apply knowledge of ATE on the concrete class. Another important point that Cartwright stresses is that the "most difficult thing to predict, even if you are prepared to admit a reasonable degree of error, is the actual value of the outcome, individual by individual" (Cartwright & Hardie, 2012, p. 30). This makes it rather complicated to travel the way from research finding to the concrete practice.

Critiques and discussions

One of the critiques against (Deaton and) Cartwright's (D&C) (2018) notion of external validity is that it does not deliver a model for how one can predict the outcome from one (study) population to another (Pearl, 2018). Critics argue that D&C's approach does not deliver any solution of how to compare and apply the knowledge of the structure of one population with another. In relation to this critique, I find that it would be interesting more deeply to discuss:

i What implications Cartwright's approach on external validity could have on the way researchers and policy-makers extrapolate and use the knowledge from SP to TP?

If Cartwright is right, it implies that we can never build our causal models about the target population *beforehand* (Cartwright, 2011, p. 76). We always need to engage with local knowledge about TP before any prediction of what is going to happen when we intervene. Consequently, this will

[8] ATE equals $E[\delta] = E[Y^1 - Y^0]$ (Morgan & Winship, 2015), which basically is the same as equation (2).

[9] φ signifies a subpopulation of the given (study) population.

mean that policy-makers, researchers, and practitioners need to think much more carefully about their intervening than I think they do at the current state. As mentioned before, this is due to the (blind) trust in the return of meta-analysis and systematic reviews, which draw on inductive reasoning. But if we cannot simply use the knowledge extrapolated from RCTs the question often becomes:

ii How do we determine when a well-conducted RCT is relevant for TP? And what is the applications of the knowledge of well-conducted RCTs?

I basically agree with Cartwright's assumption that RCTs are only relevant for TP, when we have knowledge about the causal principles governing both SP and TP (Cartwright & Hardie, 2012, p. 23). To know that a salary bonus for teacher in India decreases days of absent does not require that the same will happen if the intervention is introduced in, for instance, the Nordic countries: simply because teachers in the Nordic countries already have a low number of days of absent and they have a decent wage. This example tries to illustrate how different support factors are in place in different places. Although it occasionally appears to me that Cartwright's advice of the use of support factors ends up in the same trouble as the advice of the US Department of Education saying that RCTs are only relevant when conducted "in school settings similar to yours". It seems to me that RCTs only become relevant when the support factors of SP equal the support factors of TP. This is a strong assumption and difficult to achieve, and solely applying RCTs when the context of TP equals SP will severely limit the use of RCTs. Nevertheless I think Cartwright teach us a good lesson showing exactly what RCTs can tell us – if there exist a positive difference between the treatment and non-treatment group – "that X causes Y in some causally homogeneous subpopulation φ of the study population" (Cartwright & Munro, 2010, p. 261). Therefore, RCTs can shows us that *X* has had the ability to work in some settings. This is a good start when wanting to transfer an effective intervention from one place to another, but it still does not tell us if the study population shares the same causal principles as the target population, or whether the intervention can be expect to work in co-operation with other support factors. Therefore, one could, ask:

iii How is it possible to obtain knowledge about the *causal principles* (structures) that govern the SP as well as the TP? Can any methods deliver this knowledge?

To the above question, I will argue in line with Cartwright (2007b, p. 39) that there exists no universal methods for obtaining both the causal principles of SP and TP. To find out whether SP and TP share the same (or have different) causal principles requires additional knowledge about TP, which scientific studies cannot deliver as such. However, I find it difficult to find out how we should obtain knowledge about TP. Should politicians apply register data to a greater extent to recognise the current causal structure of TP before they intervene, for instance, in entire school systems? Or should teachers be consulted when new interventions is introduced etc.. This is still an unsettled matter for me. I try to elaborate on these matters in my presentation.

Although, I agree with Cartwright that no methods *as such* can tell whether SP = TP, because as Cartwright among other things says, "time passes, [most] things change," I will contest Cartwright's argument that "we do not have good explicit methodologies for how to establish *tendency claims*" (2011, p. 767). Cartwright argues that we should use the conception of *stable tendencies* instead of external validity. Stable tendencies draw on the idea of whether a factor has the stable capacity to promote the effect across a range of situations under consideration (2010, p. 262). Contrary to Cartwright's stance that we do not have any methods to find stable tendencies, I will argue that well-conducted meta-analysis and systematic reviews can indicate if given interventions contain stable capacities across different structures or not. I think meta-analysis and systematic reviews allow us to get an overview of the effect of an intervention in various situations. I try to show this fact by using Cartwright's causal principles in an extended version in equation (4).

$$
\begin{aligned}
y(i_1)c &= a_{1,1} + a_{2,1}y_{0,1}(i_1) + a_{3,1}b(i_1)x(i_1) + a_{4,1}z(i_1)\\
y(i_2)c &= a_{1,2} + a_{2,2}y_{0,2}(i_2) + a_{3,2}b(i_2)x(i_2) + a_{4,2}z(i_2)\\
y(i_3)c &= a_{1,3} + a_{2,3}y_{0,3}(i_3) + a_{3,3}b(i_3)x(i_3) + a_{4,3}z(i_3)\\
y(i_4)c &= a_{1,4} + a_{2,1}y_{0,4}(i_4) + a_{3,4}b(i_4)x(i_4) + a_{4,4}z(i_4)\\
y(i_5)c &= a_{1,5} + a_{2,5}y_{0,5}(i_5) + a_{3,5}b(i_5)x(i_5) + a_{4,5}z(i_5)
\end{aligned}
$$

The second submerged numbers 1-5 and submerged numbers around *i* represent different contexts (Σ) and times (T). My argument will be that if we use this conception wisely it allows us to know when we have an intervention with stable tendencies or if the intervention just has momentary effects. I actually think that this way of thinking allows us to know when and when not it can be reasonable to draw on Pearl's graphical theory of causation (Pearl, 2009) which mostly presupposes tendencies. I will not argue that this gives us a general licence not to seek any further knowledge about given TPs, but I think this could represent a wise use of how research can inform decision-making. It can indicate whether the intervention contains a stable tendency, and thereby, whether the intervention can be expected to have a wider use. And on the contrary a well-conducted meta-analyses or systematic review might be able to show how/if the intervention has different effects across different setting. This could elucidate when it would be an unwise idea to apply the intervention widely.

However, I think to realise the above-mentioned potential of meta-analysis and systematic reviews requires a renewed idea of the conception of synthesising results across different study populations, which often forces simplified conclusions (Vembye & Jensen, 2018). Furthermore, I will argue that this way of thinking allows us to understand how interventions previously have worked in between various support factors (moderator variables) and how causes are mediated to their effect (mediator variables) in different ways.[10]

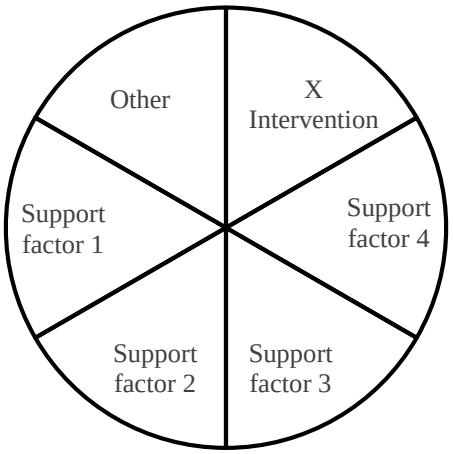
[10] For an excellent example of this kind of systematic review, see Kavanagh et al. (2011).

For another philosophical point of discussion, Cartwright describes CP as an attempt to represent the notion of INUS conditions, [11] but one of the problems about the INUS condition is that it lacks a clear distinction between causal mechanisms and circumstantial conditions (Pearl, 2009, p. 314). This means that it is difficult to distinguish between relevant causal factors from irrelevant other circumstantial present factors. Even though, C&H (2012, p. 187) argue: "that causes are INUS conditions. But not (…) all INUS conditions are causes," it is often difficult from a perspective of the INUS condition to see:

iv How to determine which factors represent conditions and which represent cause? How to avoid over-determination, when obtaining knowledge about structures?

I always have difficulties to understand when we know that we do not over- or underestimate the important factors that produce a contribution to the effect. Cartwright uses causal cakes (see figure 1) partly to illustrate that causes never work in isolation and partly to make the point that other causal cakes can contribute to an effect independently of the given intervention. Cartwright thereby escapes the argument of similarity and makes a wide space for practitioners to think themselves about their own practice, but Cartwright's use of the factor "other" in the cakes makes it difficult to find out if the cakes contain the *exact* relevant factors to make their contribution to the effect. When do we know that our cakes are satisfied with the right factors? As Cartwright states: "To partitions too finely is as bad as not to partition finely enough. Partitioning on an irrelevancy can make a genuine cause look irrelevant, or make an irrelevant factor look like a cause" (1984, p. 36). I will try to discuss these issues further in the presentation.

Figure 1. – *Example of Cartwright's causal cakes*



[11] "An INUS condition is an ***I***nsufficient but ***N***ecessary part of an ***U***nnecessary but ***S***ufficient condition for producing a contribution to the effect" (Nancy Cartwright & Hardie, 2012, p. 25; Mackie, 1965, p. 24).

References:

Bradford-Hill, A. (1965). The Enviroment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300. https://doi.org/DOI: 10.1016/j.tourman.2009.12.005

Cartwright, N. (1984). How the laws of physics lie. AAPT.

Cartwright, N. (2007a). Are RCTs the Gold Standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/s1745855207005029

Cartwright, N. (2007b). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Cartwright, N. (2011). Predicting'it will work for us':(way) beyond statistics. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 750–768). Oxford University Press.

Cartwright, N., & Efstathiou, S. (2009). Hunting Causes and Using Them: Is There No Bridge from Here to There? *The London School of Economics and Political Science Houghton*, 1–29.

Cartwright, N., & Hardie, J. (2012). *Evidence-based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *J Eval Clin Pract*, *16*(2), 260–266. https://doi.org/10.1111/j.1365-2753.2010.01382.x

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

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine*. https://doi.org/10.1016/j.socscimed.2017.12.005

Dyssegaard, C. B., Larsen, M. S., & Tiftikçi, N. (2013). Effekt og pædagogisk indsats ved inklusion af børn med særlige behov i grundskolen. *Systematisk Review. København: IUP, Aarhus Universitet*.

Egelund, N., & Qvortrup, L. (2014). Forord til bogen. In J. Hattie & G. Yates (Eds.), *Synlig læring og læringens anatomi*. Frederikshavn: Dafolo.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8.

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews*. Sage.

Harris, C., Hedges, L., & Valentine, J. (2009). Handbook of research synthesis and meta-analysis. *Russell Sage Foundation, New York*.

Hattie, J. (2009). *Visible Learning – A Synthesis of Over 800 Me-ta-analysis Relating to Achievement*. New York: Routledge.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kavanagh, J., Oliver, K., Oliver, S., Mara, A. O., Stansfield, C., Thomas, J., … Eppi-centre, J. T. (2011). Childhood obesity and educational attainment A systematic review, (1901).

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kvernbekk, T. (2016). *Evidence-based Practice in Education. Functions of evidence and causal presuppositions*. New York: Routledge.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

Mansell, W. (2008). Research reveals teaching's Holy Grail. Retrieved from https://www.tes.com/news/research-reveals-teachings-holy-grail

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Nielsen, T. K., Tiftikci, N., & Larsen, M. S. (2013). *Virkningsfulde tiltag i dagtilbud: Et systematisk review af reviews*. Danish Clearinghouse for Educational Research.

Oakley, A. (2003). Research evidence, knowledge management and educational practice: early lessons from a systematic approach. *London Review of Education*, *1*(1), 21–33.

Pearl, J. (2009). *Causality : models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Pearl, J. (2018). Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science & Medicine*. https://doi.org/https://doi.org/10.1016/j.socscimed.2018.04.024

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., & Andersen, S. C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research*, *60*(3), 311–335. https://doi.org/10.1080/00131881.2018.1493351

Reiss, J. (2018). Against External Validity. *CHESS Working Papers*, (2017-03).

Vembye, M. H., & Jensen, H. S. (2018). How the Laws of Education Lie. Retrieved from https://pure.au.dk/portal/da/activities/how-the-laws-of-education-lie(265be54c-88c7-4e46-894d-4c256fc14006).html