Comparing Two Models of Evidence

Tone Kvernbekk

University of Oslo

Follow this and additional works at: https://scholar.uwindsor.ca/ossaarchive

Part of the Education Commons, and the Philosophy Commons


This Paper is brought to you for free and open access by the Conferences and Conference Proceedings at Scholarship at UWindsor. It has been accepted for inclusion in OSSA Conference Archive by an authorized conference organizer of Scholarship at UWindsor. For more information, please contact scholarship@uwindsor.ca.
Comparing Two Models of Evidence

TONE KVERNBEKK
Department of Education
University of Oslo
P.O.Box 1092 Blindern, N-0317 Oslo
Norway
tone.kvernbekk@iped.uio.no

Abstract: The context for this paper is evidence-based practice (EBP). EBP is about production of desirable change. The evidence should come from randomized controlled trials (RCTs). To make sense of RCT evidence it must be placed in an argument structure. I compare two different models, Toulmin and Cartwright, and investigate whether the two models can be merged into one. I shall argue that such merging is not feasible.

Keywords: argument structure, Cartwright, EBP, evidence, holism, model, model merging, practical reasoning, RCT, Toulmin

1. Introduction

In many areas of life there are political demands for better research bases to inform and guide policy and practice. This demand is grounded in an impeccable what works logic: we wish to bring about desirable results and prevent undesirable results, and this is better achieved if our actions are guided by evidence of what actually works. Researchers respond to this demand in many ways; some by producing the kind of evidence it is assumed can serve as a base for practice, others by fleshing out practical guidelines. Some researchers respond by criticizing or even rejecting the whole enterprise of EBP, often because of worries about restrictions in the freedom of professionals to exercise their judgment.

The term ‘evidence-based’ obviously draws attention to evidence. Epistemologists seem to agree that the term ‘evidence’ refers to that which serves to confirm or disconfirm a claim (hypothesis, belief, theory) (e.g., Achinstein, 2001). The main function of evidence can thus be summed up in the word support: evidence speaks to the truth or the trustworthiness of a claim. One big question is therefore whether evidential support of a hypothesis is identical to the hypothesis being based on the evidence. That is to say, what role does the evidence actually play in EBP. Another much debated question concerns the kind of evidence EBP can or should use. EBP is often criticized for involving an evidence hierarchy which ranks quantitative data at the top and professional judgment and personal experience near the bottom (e.g., Oancea & Pring, 2008). I shall side-step this question, just note the following: in EBP the evidence is meant to be research-based and quantitative; ideally it should emanate from randomized controlled trials (RCTs). This makes good sense in EBP, given the fact that EBP is deeply causal and concerns the production of desired results (or the prevention of undesired results)—for short, we do X and we get result Y. The RCT is a research design which is well suited to distinguish between correlation and causation, which of course is of paramount interest to a what works logic.

I shall simply take as my point of departure that the evidence we have in EBP is quantitative, emanates from RCTs and speaks to the truth value of a causal hypothesis of the form “X leads to Y”. If X actually does lead to Y, which is our desired result, we say that X works. If it does not lead to Y, then we say that X does not work. But if the evidence speaks to the truth value of a causal hypothesis, what if anything does it mean to say that practice is based on this evidence? In more general terms; what is the relation between evidence and practice?
This paper is about different ways of understanding the role that evidence plays (or can play) for practice in EBP. To figure out what that role is, we have to place evidence in a model; a model in this case being an argument structure. I shall discuss and compare two such models: Stephen Toulmin’s argument model (Toulmin, 2003) and Nancy Cartwright’s evidence-in-use model (Cartwright, 2011, 2012; Cartwright & Hardie, 2012). Both models accommodate RCT evidence, but their structure is different. In Section 2 I describe the models, and in Section 3 I draw out their similarities and differences. In Section 4 I shall discuss whether it is possible to merge the two models and thus capitalize on their good features while avoiding any shortcomings they might have. It might sound paradoxical, but I shall argue that in this case the two models are better left unmerged.

2. Two models of evidence: Toulmin and Cartwright

It is important to my overall argument in this paper that EBP is practical. It is goal-directed and concerns what we should do, not what we should believe. Hence, EBP is not really about finding causes. It is about using causes in practice to produce a change of a more or less specified magnitude in some specified entity or situation; for example to make students more proficient readers during first grade, or improve certain aspects of people’s character.

How does RCT evidence help us do that? Before I dive into Toulmin and Cartwright I would like to briefly inquire into a third model which will provide a useful contrast. This is Philip Davies’ (1999) much cited rendering of EBP (or EBE, as he calls it: evidence-based education). Davies is a defender of EBP, and here is what he thinks educational practitioners should do (Davies 1999, p. 109):

- Pose an answerable question;
- Know where and how to find evidence;
- Retrieve, read, and critically appraise the evidence;
- Organize and grade the power of the evidence;
- Determine its relevance to their educational needs and environments.

This is Davies’ definition of EBP. Implicitly it also provides a model for the function of evidence. It is useful to have this model as a backdrop, for two reasons: it highlights orthodox EBP reasoning about evidence, and it provides an instructive contrast to my two selected models. However, Davies’ (1999) model shares a common feature with the Toulmin (2003) and Cartwright (2011, 2012; Cartwright & Hardie, 2012) models: they all begin with an answerable question, and the relevance of the evidence must be determined locally, by the practitioner. But there the likenesses end. There is no mention of the causal claim that the evidence is evidence for in Davies’ model, rather the model centers on evidence which is to be retrieved, appraised, organized, and its relevance determined. It therefore seems that Davies thinks that the answer to the question can be read directly off the evidence. Thus, that the practice is in a very direct sense based on evidence. I think this is a misunderstanding: evidence is made up of quantitative data, and you do not base practice on that. You base practice on an idea that if we do X, we are likely to get Y as a result. Nevertheless, adherents to EBP tend to assume that evidence is directly relevant for practice and even sometimes speak about implementing the evidence. In a similar vein, but cast negatively, critics say that “based” implies that the evidence—practice relationship is one of derivation, like an algorithm that allows you to extract a practice or a policy from the
TONE KVERNBEKK

evidence, leaving no room for professional judgment (e.g., Biesta, 2007). Advocates and critics alike thus tend to assume that the evidence—practice relation is direct, such that practice can be derived from or read off the evidence.

This is what we are up against, then: to find a better, more adequate model for the role that RCT evidence can play for practice, given that it is not a direct relationship. You do not derive a practice from RCT data. Practice is based on the causal claim \((X \rightarrow Y)\) that the evidence is evidence for.

Both Toulmin’s argument model and Cartwright’s evidence-in-use model accommodate RCT evidence, and both cast it in an indirect role, as opposed to Davies’ model outlined above. Let me begin with Toulmin’s (2003) model. What I shall present here is an adaptation of Toulmin’s argument model. That is to say, I keep the model intact, but I change its use somewhat by putting it in the service of practical reasoning. This is because EBP is practical in nature.

Toulmin’s model is like a spiral of ever more considerations added as the argumentative need arises. He begins by distinguishing between the claim or conclusion \((C)\) that we wish to establish and the facts we appeal to as a foundation for \(C\), called grounds or data \((D)\) (p. 90). The question “how did you get there” is directed at the step from \(D\) to \(C\) and how this step can be justified. Propositions that provide the justification for inferring \(C\) from \(D\) are called warrants \((W)\) (Toulmin, 2003, p. 91). As David Hitchcock (2003) puts it, the warrant licences the inference. Warrants confer different degrees of strength on the claims that they justify; hence we may need to use a qualifier \((Q)\) to express this: necessarily, probably, possibly, etc. Rebuttals \((R)\), often expressed as “unless”, point to exceptional circumstances which undermine the general authority of the warrant. Both \(Q\) and \(R\) thus have a bearing on \(W\). Next we come to a central concern: whether the warrant is acceptable at all. Suppose you insist that it is reasonable to think \((W)\) that some \(C\) follows from some \(D\) and somebody says “but why on earth do you think that?” This is where backing \((B)\) of the warrant enters the picture, and with it an indirect role for research evidence. “Standing behind our warrants, […], there will normally be other assurances, without which the warrants themselves would possess neither authority nor currency,” Toulmin (2003, p. 96) says.

Toulmin’s model, as he describes it, is about defending claims against challenges—the model lays out a pattern of justification of beliefs. As I indicated above I am putting the model in a practical EBP context and I therefore change the basic question from “how should I defend this claim” to “what should I do”. With this shift a story of the following sort emerges: Suppose I am a teacher; I observe that some of my first-graders are lagging behind in their reading \((D)\), and I tell my colleague that I should give them some extra word and phonological training \((C)\). But why do you want to do that, my colleague asks, inquiring after how I get from \(D\) to \(C\). Different warrants \((W)\) are possible here. For example, it seems to have worked okay with other children, or somebody said it’s a good strategy.

Let us pause here briefly and take a closer look at \(D\), which is also a form of evidence—evidence on the basis of which you form an idea about what you should do. So is this not practice based directly on evidence after all? In one way it is, since \(D\) admittedly comprises evidence. But this is practice-based evidence, picked up from practice on a daily basis; from observations, tests, interactions and intuitions, and leading to a judgment that my students are reading-delayed. This is not research-based RCT evidence and thus not the kind of evidence that EBP is supposed to employ. But granted, it serves as a starting point.

Research-based evidence comes into play in the Toulmin (2003) model, I argue, when somebody problematizes the \(W\) and queries the justification of the \(D\)–\(C\) step: are you sure that
extra phonological training will help? In passing, note the causation involved here: we do extra phonological training, and it will (hopefully) lead to better reading skills in the targeted students. Standing behind and backing up the warrant in backing B: yes, I feel confident that phonological training will work, because there are three solid RCTs showing the effect of phonological training for reading-delayed students. Thus, research-based evidence plays an indirect role; it helps you justify your conclusion to do C. This is also what the epistemologists tend to say: the main function of evidence is that of support of a hypothesis. In this case what is supported is a conclusion to implement a certain strategy. The evidence supports it, but does not tell you how to do it. You cannot derive your practice from it.

We have to cover some ground before we can look at Cartwright’s (2011, 2012; Cartwright & Hardie, 2012) evidence-in-use model. I said above that EBP is causal in nature. Cartwright’s model of evidence is based on a certain understanding of causation; namely that causes (at least in the EBP context) are best understood as INUS-conditions. To make a long causal story short, the INUS approach tells us that results or outcomes are not produced by one single cause, but have contributions from different sources. This sounds commonplace, but is easily forgotten. We tend to look for the cause and if we implement an intervention it is only natural that this intervention is salient for us and that we therefore ignore other factors. But the overall effect on Y depends on how all these factors add up. The gist of the INUS approach is thus that an intervention X is part of a team of causes and enabling factors which work together to produce Y.

The second presupposition for Cartwright’s (2011, 2012; Cartwright & Hardie, 2012) model is that causal connections are local. We cannot simply export a causal connection, insert it into a different context and expect it to work there. Contexts are different, students are different, teachers are different, curricula are different, headmasters are different, parents are different, and school cultures are different. So if an RCT shows that X works well (produces desired results) with reading-delayed children in Exeter, why should I think that X will work for my reading-delayed students in Oslo?

The question at issue in the Cartwright model is thus an effectiveness prediction: will X work here, were I to implement it? That is what practitioners want evidence for; whether a given intervention is worth implementing in their local context. The RCT evidence is indirectly relevant to this question, because it tells you that X has worked somewhere, but in itself it does not tell you that it will also work here, since here is different from there. It can however be made relevant to whether X will work here, if we collect a number of other facts and evidences and put them all together into an argument structure. That is to say: what kind of premises do we need if the conclusion is to be “yes, extra word and phonological training will in all likelihood work here, for my students”? This is precisely what Cartwright’s evidence-in-use model tells us: what I must provide to make the RCT evidence relevant to my conclusion and take us from there to here (Cartwright & Hardie, 2012, p. 133).

Conclusion: Yes, X will in all likelihood work if I implement it here (or no, it will not).

Premise 1: The INUS approach says that X does not work alone but is part of a larger constellation of factors and enablers. The support factors for X are a, b and c. These are necessary for X to be able to do its work.
Premise 2: The causal support factors a, b and c are indeed in place here. If they are not, they can be easily procured.
Premise 3: X can play a causal role here.

These are the three major premises in Cartwright and Hardie’s evidence-in-use model. But where is the RCT evidence? Its role appears only when we begin to examine the premises and look into the next layer down. Below each major premise are the premises in the sub-arguments that support the major premise. Below each of those again are the premises of the sub-sub-argument that supports the sub-argument. One of the sub-arguments for major premise 3 is that X played a causal role somewhere. And how do we know that X helped produce Y there? Because there exists an RCT that shows it. Thus, the RCT serves to back up a sub-premise or is a premise of a sub-argument (Cartwright and Hardie describe it in both ways). Whichever way we choose to describe it, the RCT evidence plays an indirect role for the conclusion.

To return to the conclusion: we see that it requires heterogeneous evidence, from RCTs to various facts on the ground. I put all these pieces of heterogeneous evidence together and I make the judgment that yes, all in all I believe that X has a good chance to work here.

3. Comparing the models

At the outset I judge both Toulmin and Cartwright’s models to be more adequate than the model we can extract from Philip Davies’ five steps, for the reason that Davies (1999) casts research (RCT) evidence in a direct role in relation to practice. A good many advocates and critics of EBP in education do the same; they think that ‘based’ signifies that practice is somehow derived or extracted from the evidence (e.g., Biesta, 2007). I have argued that you cannot derive practice from quantitative data, and that the evidence actually speaks to the truth value of the causal claim that X brings about Y. The practice (teaching method) is based on this claim, not on the evidence supporting the claim.

Toulmin and Cartwright’s models alike cast research (RCT) evidence in an indirect role. The models provide different argument structures, but in this particular respect they obtain the same result. But there the likenesses stop.

Before we proceed, let me reiterate the point that the Toulmin (2003) model as I employ it here is an adaptation—his model was not constructed to answer the question of what one should do, but the question of how to defend claims. It accommodates RCT evidence very nicely, but backing B was not constructed to do that specifically—B comprises various considerations that back up and secure our warrants. Cartwright’s model, on the other hand, is expressly constructed to accommodate RCT evidence in an EBP context. A comparison of the two might therefore seem unfair, but is nevertheless worthwhile.

Let us begin by looking at the shape of the models. Toulmin’s (2003) model is like a spiral with ever wide circles, running on its own internal drive by questions which come naturally as the argument develops (“how do you get from D to C”, “how do you know that phonological training helps?”, etc.). The model is easily employed; it guides you through a spiral that allows you to look at one thing at a time, and you do not get lost in its complexity because there are no sub-sub-premises. The model shows a very sensible role for RCT evidence, namely as backing B. My initial feeling is that our practical reasoning toward a decision about what to do will be clearer and better justified if we employ this model. Cartwright’s model is a classical model with premises and conclusion and is built up like a pyramid. The conclusion on top is
supported by major premises, which are supported by sub-argument premises, which again are supported by sub-sub-argument premises, etc. (Cartwright & Hardie, 2012, p. 17). The role ascribed to RCT evidence is that of a sub-argument premise (also described as a sub-sub-premise). The pyramid, as I understand it, has both a general and a specific point. The general point is the importance of always backing up one’s premises. This is more than just a self-evident statement; it also points to the direction of the reasoning here. We begin at the top, with a conclusion concerning the likelihood that X will work here. Then we work our way downward in the pyramid, by mapping major premises, then securing these premises with new ones, and so on and so forth. The specific point concerns RCT, and is the idea that RCT evidence speaks to the claim that X worked there, which I suggested above is the first premise in an EBP argument to the conclusion that it is likely (or not) that a given intervention will work here. However, given that the argument pyramid is built up of premises, sub-premises, sub-sub-premises, premises backing up sub-arguments, etc., the structure branches out down through the levels and gets enormously complex. The layers may be difficult to keep track of. On the other hand, if we get it right, it does put order into our adduced reasons and it does help us assess the degree of confidence we should have in our conclusion—just how likely is it that X will work here?

Second, let us look at point of departure for the reasoner. Toulmin’s model of practical reasoning begins right where the practitioners are: in practice, facing problems to be solved. Much reasoning about what one should do is set off by perceptions or felt problems in practice; for example the observation that some students are reading-delayed. This is what Michael ERAUT (2004) terms “practice-based evidence” (p. 92) and I propose to understand it in terms of data D. Practice-based evidence and its important role in professional contexts is, I argue, amply captured by the D-C connection in the Toulmin model. This model thus encapsulates what I take to be two of the most important elements of practitioners’ reasonings; what the problem is (or what I take it to be) and what I should do. For that very reason I guess the model would be attractive to employ, especially when coupled with its internal spiral drive which guides the reasoner from issue to issue.

Cartwright’s model starts with a different question, namely an effectiveness prediction: will X work here? But where does that question come from? Practitioners who are using Cartwright’s model begin at a very different place. They are not wondering how to solve a particular problem (whether relating to “clients” or not); rather they have a ready-made X at their hands and have to decide whether to implement it. Where does this X come from if the practitioners do not devise it themselves? It comes from various “warehouses” such as What Works Clearinghouses that conduct meta-analyses and vet interventions and advise you to choose interventions backed by good evidence. This is recognizable from Davies’ model outlined above, with the exception that his model omits the all-important X; that which is to be implemented to produce desired results. But how does this situation look like on the ground, so to speak? I have no doubt that practitioners sometimes find themselves in situations with an already existing X. What, then, are their options? My guess is that most practitioners take it for granted that “it worked there” automatically means “it works in general” which in turn automatically means “it will work here”. If so, they do not see the necessity of running through an argument pyramid, but just implement X. Actually they may not have a choice—the decision to implement X may not even be theirs; it may be made by the local government. But that is another issue.
There is nothing in Cartwright’s model; it seems to me, which plays the same role as D does in Toulmin’s model. Toulmin’s model thus accommodates practice-based evidence, and Cartwright’s model does not.

Third, let us look more closely at the RCT evidence. As suggested above, both Toulmin and Cartwright’s models provide an indirect role for RCT evidence, but there are subtle differences between them. To see this, we have to look at what RCT evidence tells us. The privileging of RCT evidence is a characteristic feature of EBP and much criticized, since it implies that other kinds of evidence that are important in practice (for example professional judgment or personal experience) are under-valued. However, the privileging can be justified in the case of EBP, because what we want here is evidence that speaks to the truth value of a causal hypothesis, not just any hypothesis. In RCTs we compare groups that are the same with respect to all relevant (causal) factors except one: the intervention. Random assignment is supposed to ensure that the groups have the same distribution of properties and factors. The standard result of an RCT is a treatment effect, generally expressed in terms of effect size: average effect in treatment group minus average effect in control group. We assume that the difference between the groups needs a causal explanation, and since all other factors (supposedly) are equally distributed we infer that the treatment, our intervention X, is the cause of the outcome. X works, we might be tempted to conclude.

RCTs are strong on internal validity. If we obtain an average positive result and the conditions of the trial are met, we may safely believe the causal claim in question. Doing X indeed leads to Y, and the evidence shows it. But internal validity is purchased at the expense of external validity, or generality. As Cartwright (2007) argues, what RCT evidence shows is strictly speaking that the X—Y relation holds where the trial was conducted, for that particular study group. It holds there. But the fact that it holds there is not itself evidence that the X—Y relation holds generally across differing contexts, even though we seem to presuppose that it does—somehow we seem to think that if an RCT shows a causal connection, this causal connection holds simpliciter and simply exists everywhere. Hence, we tend to think that if something worked there, it will also work here. This assumption certainly seems to underlie Philip Davies’ (1999) model of evidence as well as the practical guidelines produced by What Works Clearinghouses (often based on many there). This limitation of RCT evidence is really not discussed in the EBP literature, with the exception of Nancy Cartwright. We therefore should not wonder that her evidence-in-use model is developed to take account of precisely this limitation. If causes do not hold simpliciter, then they do not work generally. Then how do we make what worked there relevant to the question of whether it (X) will work here? We look at the context. Causal connections are not general; instead they are dependent on contextual factors for their workings. Causal connections, Cartwright says, only hold ceteris paribus. Hence, to make a judgment whether X will work here, we have to collect information about the necessary contextual factors, the causal support team: what the support factors are, whether they are at hand or whether they can be procured if necessary. This is expressed in her major premises.

This works (sic) differently on the Toulmin model. Suppose I observe that some of my first-graders are reading-delayed. That is my diagnosis of the situation, D, and I infer that the children should receive extra phonological training, C. On my colleague’s query I state that phonological training is a well tried remedy, W. But how do you know that it will work, my colleague asks—that is, he is asking for backing of my warrant. So I refer to an RCT (Hatcher et al., 2006) which shows that compared with the control group, children who received this this
intervention for two consecutive 10-week periods made significant progress on letter knowledge, single word reading and phoneme awareness. This is RCT evidence as backing B, and its relation to my decision is indirect. But, and here is the current crux of the matter, it is natural to think and speak of it as *general*. Why do I think X will work with my students? Because an RCT (probably more than one, too) shows that X works. When I reason my way through Toulmin’s spiral, I then come to this point: In order for the RCT to be able to serve as B and justify our warrant, I am forced into treating it as yielding general results. This is in accordance with widespread beliefs about RCT results, but given their research design they actually yield results limited to the study group. Thus it might seem that on the Toulmin model it comes naturally to think of RCT results as general; thereby providing an easy bridge to “X will work here”, but at the same time misunderstanding what the evidence tells us. But then Toulmin’s model has a trick up its sleeve; qualifiers Q and rebuttals R. That is to say, we should run through the entire spiral and not just stop at B. Q and R both speak to the bearing of W on the inference from D to C; qualifiers by indicating the strength of the warrant, and rebuttals by pointing to possible conditions of exception. Hatcher and his colleagues state explicitly that there are exceptions; a full 25% of the children did not respond to the intervention and this non-responsiveness is tentatively explained in terms of the children’s pre-existing knowledge and skills (2006, p. 825). A piece of practical reasoning running through the Toulminian spiral might sound like this: Some of my students are reading-delayed (D), and *unless* they have extremely low initial scores on word recognition and letter knowledge (R), they will presumably (Q) respond well to word and phonological training (C), since this is a well-tried remedy for reading-delayed children (W); the warrant being backed by sound RCT evidence (B).

What is happening here? With Q and R we have introduced flexibility and probability into the model, but the conditions of exceptions as I have worked them out here, do not apply to B. Thus, the RCT evidence still comes through as being general, at least in how we express this kind of reasoning linguistically. The exceptionality has a different source. Hatcher et al. (2006) locate it in the students; the non-responsiveness of students is explained by the students being below the cut-off point where the intervention becomes useful. It is difficult to gauge how they understand the causation involved in their study, but my guess is they understand it as *simpliciter* and general (since that is what most researchers seem to do…). If you were to implement their intervention in your classroom and fail to attain Y, the failure would be understood to be about the “material” and not about the limited external validity of RCTs and/or the absence of a proper causal support team. Cartwright’s plea for premises to take us from *there* to *here* seems difficult to accommodate in Toulmin’s model.

To sum up, the two models share one central feature: they both cast RCT evidence in an indirect role. This is vital if we are to understand the role of RCT for EBP at all. Casting it in a direct role, as many advocates and critics alike do, leads to understanding the evidence—practice relation as one of derivation, which is clearly not feasible. But since we are discussing the possibilities of merging the models, perhaps the differences are of more interest. I have argued that they answer different questions: “what should I do” and “will X work here”. I conjecture that practitioners, whose job it is to produce desired results, whether by implementing a ready-made X or devising an intervention themselves, might find Toulmin’s model attractive because it speaks directly to issues that are within their power of control, namely what they themselves should do. Toulmin’s model also accommodates what is known in the (educational) EBP literature as “practice-based evidence”; Cartwright’s model has no place for that. Cartwright’s model includes premises to take us from *there* to *here*, Toulmin’s does not. This particular issue
also touches the restrictions of RCT evidence and the idea that causal X—Y relations hold locally but not *simpliciter*. The restricted reach of RCT evidence is not accommodated in Toulmin’s model; as backing B it must be general to serve its purpose for the decision to be made. Toulmin’s model explicitly incorporates exceptions (R) but I am not sure if (or how) that helps to solve the problem. Finally, Toulmin’s model, with its internal drive, is user-friendly and might appeal more to practitioners than Cartwright’s more complicated model.

4. Merging the models?

Each model has strengths and weaknesses and I confess I like them both. Now, if we merge them, can we get the best of both worlds? Can we get a model that speaks to practitioners, has an internal logical drive, accommodates practice-based evidence, understands the role of RCT evidence as indirect, accommodates causes as INUS conditions, allows the restricted stretch of RCT designs and provides room for contextual evidence that takes us from *there* to *here*?

In my professional domain, which is education, models are combined fast and easy. Concerning, for example, didactical models, many textbooks I read as a student began by showing two or three very different models; behaviorist, cognitivist and constructivist models. Then it was argued, quite correctly, that all models taken in isolation are oversimplified and only present a partial picture of the phenomenon. The conclusion would be that all models integrated or combined into one is the solution—the new model supposedly representing a (more) holistic picture of didactics (for example Gundem, 1983). My initial experience with model combination thus taught me that it is no big deal; they can simply be “added” together. The complexity which inevitably resulted was generally interpreted to indicate a holistic understanding, and holistic understanding was seen as an undisputed good. As a student I became skeptical of model combinations because of the sheer ease with which it was done.

But my textbook experience is perhaps more about adding models, piling them on top of each other, rather than merging them, and presumably things have happened in the field of didactics since then. I take the term *model merging* from Brunet et al. (2006), who take it to signify the combination of information from several models into a single model (p.1). That is to say, one does not (necessarily) take whole models and “add” everything in them, one takes certain entities and/or relations from each model and combines them into a new model. This is not a matter of conjunction of models or of linking them, but a matter of performing certain operations on them and putting the entities together again in a different way.

There are obviously many different ways of merging models. Some prefer the term *ensemble* methods (Banerjee & Bandyopadhyay, 2013; Opitz & Maclin, 1999) and some prefer the term *cascade* methods (Heitz et al., n.d.). These have been studied in many domains, but to the best of my knowledge not in argumentation (or in education, for that matter). There are different approaches and the requirements and underlying assumptions of these approaches may differ significantly. In these circumstances we have to be very careful about which if any insights we may extract from them and apply to the problem at hand. It is important to keep original models and the resulting merged model apart, and I shall adopt the term *ensemble* to denote the merged model.
Purpose of merging

We should not take the purpose behind model merging for granted. If there is to be a point to merging two models, we must think that the ensemble has something to offer that the original models taken in isolation do not. That is to say, we have to have a (reasonably) clear idea what sort of job the ensemble is supposed to do, and we have to have some way of comparing the results of original models and ensemble. I assume that the purpose, whatever it is, must be grounded in the context. Our current context is EBP, with its focus on “what works” and its concomitant practical reasoning.

Several of the merging methods seem to come from tasks of classification. Opitz and Maclin (1999) thus write about combining the predictions of multiple classifiers to produce a new classifier, an ensemble. The purpose of merging is that the ensemble should be more accurate than any of the original classifiers when used to classify novel instances. Opitz and Maclin (1999) discuss ensembles concerning neural networks and decision trees and conclude that the ensemble outperforms the original models. Banerjee and Bandyopadhyay (2013) also discuss classification; their case being classification of linguistic formulations in Bengali. Again, the conclusion is that the merged models outperform the existing single models. From this literature there emerges a general lesson: if we are to merge models, it must be because we think that the ensemble performs better than the single models it is constructed from.

Here I would like to register a possible snag in our exploration of the merging of Toulmin and Cartwright’s evidence models. At the outset the original classification models discussed by Opitz and Maclin (1999) and Banerjee and Bandyopadhyay (2013) perform the same kind of job. But the Toulmin and Cartwright models perform different jobs. They answer different questions and a comparison as to their performance would not make sense. Unless they can be tweaked into answering the same question, an aim of a better performing ensemble must be given up. But all is not thereby necessarily lost. Recall my student experience of merged didactic models. It indicates that the main aim of model merging was to create a (more) holistic picture of the phenomenon at hand; holism in this case being understood to mean “several different approaches to the same topic at once”. I shall come back to the creation of more holistic pictures subsequently.

Match and merge

Brunet et al. (2006) list several operations one can perform on models and the following is taken from them. These concern model management in general, but the authors do take a special interest in model merging. The operations include consistency checks, diff, slice, split, match, patch, transform, propagate and, of course, merge. Which of these operations we perform, depends on the type of models in question. And on the purpose, I would like to add—Brunet et al. (2006) seem to take that for granted. Whatever we do, we have to consider the models themselves, their respective properties and the relationship(s) between them.

The basic operator is the merge operator; the other operations are conceived of as supporting. Merge is defined as model x model x relationship → model (Brunet et al., 2006, p. 2); the resultant model being what I have termed an ensemble. The relationship specifies how the models as a whole relate to one another. This is meant to be flexible, in the sense that there may be several ways of relating models and thereby creating ensembles. Most relationships focus on mappings between model elements; hence the importance of match. In our case the relationship...
can be described in different ways. It may, for example, be an overlap. We then employ the supporting operator *match* (model $\times$ model $\rightarrow$ relationship) (Brunet et al., 2006, p. 2). *Match* is used to map out common features of the models and decide if the overlap can be used as a basis for merging the models. We have found that the Toulmin and Cartwright models share at least one feature in common, namely the indirect role they provide for RCT evidence. *Match* would give us $B$ from the Toulmin model, and a sub-sub-premise from the Cartwright model. Typically, Brunet et al. (2006) say, such mappings have to capture different vocabularies and preferences for which vocabulary to use in the ensemble (p. 6). It also seems to me that we face problems of comparison. Brunet et al. (2006) use the example of two organizational charts merged into one by using the *match* operator. The vocabulary issue arises because one chart uses the term “person” and the other uses “employee”, and it is not difficult to make the judgment that these are the same entity. But how about “backing of warrant” and “support of a sub-premise”? In general, one common feature may seem too scant for a *merge*, and in our case the common feature is located at different places in two different argument structures. On a high enough level of abstraction these serve the same function, namely provide an indirect role for RCT evidence in EBP. But is that sufficient for merging purposes? It is an open issue how much commonality is required and I leave the problem open.

Let us look at the inverse strategy; to view the relationship of the original models as one of complementarity. That is, the overlap is still there, but focus is now on concepts in each model that are absent in the other. This is how I intuitively viewed the merging problem and also why I initially thought that the differences between the models perhaps are more interesting than the commonalities: their combined strengths could make up for their perceived weaknesses. I have hinted that I take the strengths to be user friendliness, a structure with an internal logical drive, accommodation of practice-based evidence, an indirect role of RCT evidence, accommodation of causes as INUS conditions and space for contextual evidence that takes us from there to here. That is to say, most aspects of the two models. A complementarity-driven *merge* would therefore be practically like a total merge and the ensemble would encompass all properties and/or content of the two original models. On the assumption that the models are compatible, the ensemble would encompass $D$, $W$, $B$, $Q$ and $R$ as well as premises concerning an INUS condition with its larger constellation of factors and enablers and the judgment that $X$ can play a causal role here. It stands to reason that the conclusions also belong to the models, so we include $C$ and the effectiveness prediction. I admit to having grave doubts about this possible ensemble. Where the constituent models appear tidy with a clear internal structure, the ensemble more looks like a jumble; and we have not even attempted to include Cartwright’s sub- and sub-sub-premises. Brunet et al. (2006) suggest that the choice of relationship as complementary, overlapping or for example contradictory is mostly determined by the model semantics (p. 6). I would like to suggest that model structure might be another determinant.

Yet another possible strategy is to use the *slice* operator (model $\times$ criterion $\rightarrow$ model) (p. 2). *Slice* produces a partial view of a model, based on a given criterion. This operation allows us to extract an aspect from one model and then *match* and *merge* to apply it to the other model. This seems feasible if we have models that originated from different sources, which we have in our case. But are all models amenable to *slice*? At least it would seem that Cartwright’s model can be operated on in this way, given that the three major premises are independent of each other. And interestingly, Cartwright and Hardie (2012) provide an example of what might happen to the model if it were sliced using the role of RCT as criterion (p. 133). If the effectiveness prediction (the conclusion) was to rely on RCT support only, the three major
premises would be reduced to one: X can play a causal role here. The sub-premises backing up this major premise would also be reduced to one: X played a causal role there. It is this sub-premise which is supported by RCT evidence. This slice is reminiscent of Philip Davies’ evidence model in that it indicates that RCT evidence by itself is enough to bolster an effectiveness prediction, but it is a misrepresentation of Cartwright’s original model (and intended by Cartwright and Hardie to show how meager the grounds for an effectiveness prediction would be if RCT evidence was all we had). I am not sure how we should incorporate it into the Toulmin model, though. Perhaps as an extension of backing B?

*Holistic understanding?*

One of the possible purposes of model merging is the achievement of more holistic understandings: an ensemble gives a truer, better, less biased, more comprehensive, etc., picture of the phenomenon in question than the original models taken in isolation. There is obviously something to this idea. Heitz et al. (n.d.) discuss what they call cascaded models, and the purpose is precisely holistic understanding, in their case of a natural scene.

The context of their discussion is computer vision. Their point of departure is that to fully understand a natural scene several sub-problems must be solved simultaneously, but when you wish to consider them jointly you have to combine them into an ensemble. The outputs from each problem-solving task are combined to produce high-quality pictures with scene categorization, object detection, multi-class segmentation and depth reconstruction. These are not simply added together. The “cascading” in question means (as I understand it) that model1 gives input to model2, which gives input to model3, etc., such that the models solve their own problem but also help each other and thereby improve on all. The result is a holistic understanding of the scene which is also more than the mere sum of its parts— the latter being my interpretation.

It is not clear to me whether parts of their discussion may be sliced (sic) and incorporated into mine. They combined models into an ensemble, but with the clear understanding at the outset that these models are different contributions to the same large picture. In our case we have two models which are to be understood as comprehensive in their own right, not as contributions to some defined larger issue. To return to my student experience: if we perform complementarity match and merge operations on the Toulmin and Cartwright models, the ensemble would certainly increase in complexity, if only because the number of elements, factors and properties increases. But complexity is not always a good thing; sometimes it yields mess and confusion rather than holistic understanding. My intuition is that if the purpose of our model merging is a more holistic view of argumentation which accommodates RCT evidence in an indirect role, the merging operations I have discussed fail to attain it.

5. Conclusion

I have in this paper compared two models of evidence; Toulmin’s model and Cartwright’s model. The rationale for comparing them comes from the context I have put them in: the practical reasoning involved in evidence-based practice. Both models are argument structures and both accommodate a crucial feature of EBP, namely the indirect function of research-based evidence in such practical reasoning. Other than that common feature the models are very
different, both in vocabulary and structure, and the question is whether they can possibly be merged into an ensemble.

I have argued that they cannot. Even if the problem of different vocabularies could be overcome, there would remain the fact that the models answer different questions. That would not be a problem if they were different contributions to the same large picture, like in the natural scene case, but these models are whole, comprehensive models in their own right. They also have very different structures; one is a spiral and the other a pyramid.

For the same reasons I would also like to argue that they should not be merged. If we are to take the trouble of making an ensemble, it must be because we have reason to think that the ensemble is somehow better than the original models. But in this case the ensemble is not better, I venture. In the first place it is difficult to get a picture of what such an ensemble might be like, given the big differences between the original models. But if we try to combine them, it is no longer clear what the question is—only that it is not “what should I do” or “will X work here.” We would have different factors criss-crossing without a unified structure to fall into; the original structures would presumably be destroyed in the merge.

I conclude that while there is overlap between the models—I have had a reason to compare them in the first place—they are alternative models that solve different problems.

References


