PRACTITIONER TALES: POSSIBLE ROLES FOR RESEARCH EVIDENCE IN PRACTICE

Abstract: This paper discusses, compares and contrasts four different models for bringing evidence from randomized controlled trials (RCTs) into practice and into practical reasoning. I look at what questions the models can and cannot answer; what role they accord to RCT evidence, and what their possible attraction for practitioners might be. The models are those of Philip Davies, Nancy Cartwright, Stephen Toulmin and Harald Grimen. The two first are constructed for the purpose of using evidence in practice; the third is an adapted argumentation model, and the fourth is based on *phronesis* and represents an alternative to evidence-based education. It is argued that of these, Toulmin's model has the narrative qualities that make it the most usable and practitioner-friendly model.

The present educational climate worldwide is dominated by a language of improvement, raising standards, student outcomes and achievement, accountability, effectiveness, testing and measurement. Educational researchers are called upon to produce better evidence bases for policy and decisions, and to find what works in practice so students can meet expected standards and improve their achievements. This may sound like a good job for research. But in education, the relationship between research and practice has always been troublesome. On the one hand it is common to criticize research for being irrelevant and unhelpful to practice. On the other hand it is increasingly common to criticize research for invading the space of the most important professional tool of practitioners, namely their own judgment. Incidentally, this criticism is not new. A couple of decades ago it was directed against *theory,* but with the growth of evidence-based education (hereafter EBE) it is now largely directed against *research evidence* (especially in the form of data stemming from randomized controlled trials, hereafter RCTs) and its supposed role for practice, with concomitant changes in the role of the teacher and the nature of practice.

While there are many forms of evidence, the focus here is on RCT evidence – RCTs are regarded as the gold standard and RCT evidence is privileged in EBE contexts. For example, the US Department of Education's *User Friendly Guide* (2003) unequivocally takes the "rigorous evidence" which should support practice to be RCT evidence and devotes an entire chapter to RCTs and why they are needed to establish strong evidence of the effectiveness of interventions. In this paper I use the terms 'evidence', 'research evidence' and 'RCT evidence' interchangeably.

RCT evidence can be accorded different roles in/for practice, and in this paper I shall examine four different models. This is to be distinguished from implementation or translation issues, which are also obviously both important and complex and which have received much attention in recent years. My interest is rather how the research evidence might enter into practitioner reasoning. To pinpoint the role of evidence we need to put it into some relationship with other forms of information that practitioners use; in other words, we need a model. I propose to discuss and compare four such models: First, I shall briefly lay out Philip Davies' model of EBE, which is expressly constructed to provide a role for evidence in practice. Second, I shall discuss Nancy Cartwright's model, which is also expressly constructed to provide a role for evidence in practice. Third, I shall adapt Stephen Toulmin's argument model to the problem at hand. Fourth, I shall look into the favorite alternative of EBE critics, namely professional judgment. Like Toulmin's model this is not in and of itself a model for the role evidence might play in practitioners' reasoning, and I shall explore whether a role for RCT evidence can be carved out there at all. I shall especially rely on the "practical syntheses" approach developed by the Norwegian philosopher Harald Grimen. My comparison of the four models will be restricted to the following dimensions: what questions the models answer; the role they give RCT evidence, and their possible attraction for practitioners. But first we need to look briefly at educational practice – the entity that supposedly can be evidence-based.

The Nature of Educational Practice: Brief Remarks

What is educational *practice*, that it can be evidence-based? Many educationists over the years, from classics such as Jean-Jacques Rousseau to contemporary philosophers of education such as Gert Biesta, have taken the view that educational practice overall is *moral* in nature, centering on the relationship between child and educator. Traditionally this relationship is cashed out in three broad purposes for education; qualification (learning of knowledge and skills), initiation (into culture, history, society, democracy, values) and self-formation (forming character/personality and becoming an individual subject, both dependent on and separate from group). Some EBE critics surely refer to practice as moral when they criticize EBE (or downright reject it), on the assumption that the effectiveness vocabulary associated with EBE is incompatible with a vocabulary consisting of judgment, appropriateness, responsibility, educative potential, intrinsic value, etc. The critics thus argue that their own preferred vocabulary captures the nature of practice and the nature of

professional work (e.g. Biesta, 2010, 2015; Hammersley, 1997, 2005; Pring, 2004, 2015). EBE, they suggest, distorts both educational practice and professional work with its vocabulary of effectiveness, measurement, outcomes, accountability, predefined objectives and *what works* (see also Phillips, 2019).

I shall in this paper take the following view: practice is a *systemic* notion that denotes a whole. The whole itself is not amenable to RCTs and evidence-basing, I argue, not because of its alleged moral nature, but because it is a *whole* and must be understood as an open system. It interacts with local, national and international contexts and input to the system can come from anywhere. In Ludwig von Bertalanffy's classical expression a system is "organized complexity" (1984, p.21), where complexity refers to the numbers of parts in the system and organization refers to the level of order in the system. Educational practice is a whole that is made up of students, teachers, learning, socialization, self-formation, teaching, beliefs, norms, intentions, communication, perceptions, emotions, judgments, motivations, curricula, plans, aims, objectives, causal connections, outcomes, behaviors, activities, relationships, materials, results, feedback, evaluation, tests, methods of instruction, institutional values, rules and regulations, successes and failures and the various connections and interactions between them. In the wings we have headmasters, administrators, janitors, parents, school boards, policy-makers, ministries of education and other stake-holders. In other words, educational practice is a highly complex enterprise and the level of order varies from messy to well organized.

If we look at the three basic purposes of education, it would seem that of these three, qualification lends itself to RCTs and evidence-based practices, the other two do not. Initiation is a long-term process with inputs from many sources besides schools. It would be difficult (to say the least) to define the target to make it measurable and it would be impossible to control for confounders. Self-formation is a "second-order" process; it denotes that which you learn about self, others and the ways of human beings alongside being taught knowledge and skills, being tested and evaluated. Not only would an RCT on self-formation face the problems of specifying target and controlling for confounders, there will be a second-order process running parallel to it that you do not capture. Still, I take it that some of the parts of education as a whole are amenable to RCTs and to EBE. EBE can by no means cover all there is to education, and also – as critics argue – we run the risk that those parts of education which are amenable to EBE receive too much attention at the expense of other important educational tasks. But let us focus on qualification, the learning of knowledge and skills – my examples belong to this category.

Evidence in Practice: Four Models

In order to figure out what role evidence plays (or can play) for practice, we need to place it in a model together with other relevant pieces of information. There are several different models available and – I am sure – many yet to be constructed. The selected models have different strengths and weaknesses and thus provide instructive contrasts to one another.

Davies' Model

Philip Davies is an advocate of EBE. Here is his description of what he thinks educational practitioners, i.e. teachers, should be able to do (Davies, 1999, p.109):

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

This is Davies' definition of EBE. Implicitly it also provides a model, with five steps, for the function of evidence. But what kind of practitioner tale can we extract from these five steps?

It is unfortunate that Davies does not provide examples of the kinds of questions he thinks practitioners could or should ask, to which research evidence can provide answers. I suggested above that I think that the EBE approach is restricted to the parts of education that deals with qualification – that is, learning of knowledge and skills. RCT evidence cannot answer questions such as what we should aim for in history lessons, how to strike a reasonable balance between the different purposes of education, or what the educative potential of an excursion to Auschwitz might be. RCT evidence cannot by itself tell you *why* intensive phonics instruction improved the reading comprehension in the study group but not in the control group, but it can tell you *that* it did. Incidentally, David Hargreaves, an early advocate of EBE, argues that this is what teachers want to know – they want to know *what* works, only on rare occasions are they interested in *why* something works (Hargreaves, 1996). Nor can evidence tell you what would or could have happened to reading-delayed first-graders if the teacher had tried a word identification strategy instead of eye movement and use of inner speech (example adapted after Cunningham, 1993). Such questions involve comparing observed results to counterfactual results, and no RCT can provide evidence to

answer those (Pearl & Mackenzie, 2018). That of course does not have to prevent the teacher from trying word identification with her students next time around.

There are thus many kinds of pertinent educational questions that evidence cannot answer. So what kind of questions can it answer? RCTs, if they satisfy all requirements, establish with high probability that intervention x, which was present in the study group but not in the control group, contributed to the differential values of outcome y in the two groups. We tend to assume that such outcome differences must have a cause (Cartwright, 2011). Hence we attribute causal power to the intervention, since all other causally relevant factors are (assumed to be) cancelled out in the randomization process and the treatment of the two groups is otherwise identical. RCTs run on an input-output logic, but with the added assumption that outcome differences require a causal interpretation, we can (and do) extract from RCTs *what works*. If we do x, then in all likelihood, hopefully, we get result y. The standard result of an RCT is a treatment effect, expressed in terms of an effect size (but see Simpson, 2019). Practitioners would do well to remember that an overall positive effect is compatible with negative results for some of the individuals in the intervention group.

Davies' model seems to imply that answers can be read directly off the evidence. Of course we can ask questions about the effect of vocabulary skills on the development of reading comprehension in second language learners and compare this to the effect of decoding skills, and maybe find that the former is slightly bigger. But then what? Something seems to be missing from Davies' model. In fact, I shall suggest that two things are missing from his model, both of which would be useful in a practitioner tale. The first thing missing is a causal hypothesis or a causal model. There must be some causal hypothesis that the evidence is evidence *for*, without such a hypothesis the evidence makes no sense. What the practitioner really is after surely is this hypothesis. Roughly it may tell her that x causes y, for example that teaching vocabulary knowledge leads to improvement in reading comprehension, especially for second language learners. The *evidence* is a servant; it speaks to the truth value of the causal hypothesis. Similarly, when Davies' model tells the practitioner to determine the relevance of the evidence for her own educational needs, surely what is meant is that the practitioner should determine the relevance of the causal hypothesis for her own educational needs. The second thing that is missing is a sixth step; a conclusion. What inferences could the practitioner make after the relevance of evidence (or causal hypothesis) for his or her own educational needs has been determined? What could a practitioner *do* with this evidence?

Cartwright's Model

Nancy Cartwright has developed her evidence-in-use model in a number of publications (e.g. Cartwright, 2011, 2012; Cartwright & Hardie, 2012). Like Davies' model, this model also begins with a question, but a particular question: an effectiveness prediction of the form "will this intervention work *here*, were I to implement it?" This is what practitioners want evidence *for*, Cartwright suggests. Here we should note three things: that this question differs from what researchers want evidence for; that the practitioner assumes that what has worked *somewhere*, for some students, might not work for her students *here*; and finally we should also note that this is a question that Davies' model cannot answer. Research evidence by itself does not tell you whether an intervention will work in your classroom, for your students.

The practitioner tale that emerges from Cartwright's model explicitly invokes a causal claim. Her model is an argument whose conclusion is a particular causal judgment, that an intervention x will likely work in a particular context – for example, that oral vocabulary training will likely improve my second language learners' text comprehension (example adapted from Lervåg & Aukrust, 2010). The following premises provide direct support for the conclusion (Cartwright & Hardie, 2012, p.133):

1. Intervention x is an INUS condition (insufficient but necessary part of an unnecessary but sufficient condition) – x does not work alone but only if a, b and c are present.
2. a, b and c are present *here*, in my particular situation.
3. x can play a causal role *here* after it is implemented.

None of these premises contain *evidence*. The role of evidence appears when we consider that while each of these premises requires further support, premise 3 in particular does so:

3.1. x played a causal role *there* (the context where the data were collected).

This sub-premise to premise 3 in turn is then supported by evidence:

3.1.1. x played a causal role *there* and we know that because an RCT found a difference between treatment group and control group, and this difference was judged to be large enough to be of interest.

Some comments are in order. First, this tale begins with a practitioner who find herself wondering whether to implement intervention x in her classroom, with her students. This story thus seems to start in a different place than the practitioner tale emerging from Davies' 5-step model, which instead invites such questions as "is there a method for improving second language learners' reading comprehension?". Second, the first premise in Cartwright's model explicitly addresses the causal hypothesis we are dealing with; namely that intervention x leads to result y. It also shows how Cartwright understands causality, at least for this kind of problem. The concept of an INUS condition comes from John Mackie (1965) and Cartwright

extends it to multi-valued variables: "X is an INUS contributor to Y: X is an insufficient but nonredundant part of a complex of factors that are unnecessary but together sufficient to produce a contribution to Y" (Cartwright, 2012, p.979). The other factors, here simply called a, b and c, make up the causal support team without which x cannot do its job.

Third, evidence, which plays such a prominent role in Davies' model, shows up as a premise to a sub-argument (premise to 3.1.) or alternatively as a sub-sub-argument – Cartwright and Hardie describe it in both ways. Evidence thus plays an *indirect* role in this practitioner tale; it functions as a sub-sub premise in a sub-sub argument of the reasoning toward the conclusion that supports the practice in question. It tells you that the intervention you consider implementing worked *somewhere*, or perhaps even in many *somewheres.* But it does not tell you that it works generally, and it does not tell you that it will work for you, *here*. Ideally designed RCTs establish with high probability that x, which was present in treatment group but not in control group, produced or contributed to the differential values of y in the two groups – but they establish this only in the circumstances of the trial. The evidence is strong, but its scope is narrow, and meta-analyses cannot in principle solve that problem. That is why the Davies practitioner tale cannot answer Cartwright's question.

The tale does not end with the sub-sub premise that there exists an RCT which shows that x played a causal role *somewhere*; that does not provide an answer to the question of whether x will work *here*. So let us take a detour and come at the problem from a slightly different angle. The formula "x causes y", when y denotes student achievement, is a simplification and does not do justice to the INUS condition approach to causality. Causal production is better expressed by the following equation (Cartwright & Hardie, 2012, p.26):

$$Y(i)c = a_1 + a_2 y_0(i) + a_3 b(i) x(i) + a_4 z(i)$$

The very short explanation of this equation is as follows: The effect y (for individuals *i*) is placed on the left-hand side of the = symbol and the causes on the right. The *c* can be read as "caused by." The *a*s are constants across all individuals; $y_0(i)$ is the base level of y for individuals; the *b*s are the support factors that work along with the intervention; *x* is the intervention variable; the *z*s are contextual factors that do not include x. All these factors *together* produce the outcome, *y(i)* being the value of outcome y for individuals in the study group. This is the power of the INUS approach: the equation tells us that many factors besides x contribute to y and thus points to the importance of contextual factors. In other words, how

(or whether) x will work depends on the individuals receiving it and the temporal/social/material affordances at their disposal.

Now, to implement x in a new context, *here*, we need the assumption that the causal principle expressed in the equation applies in the new context. This practitioner tale is a complicated one. Cartwright and Hardie also present their argument structure in a different way (2012, p.45), which makes for a related but slightly different practitioner tale:

1. x played a positive causal role *there*
2. x can play the same causal role *here* after it is implemented.
3. The support factors (b) necessary for x to play a positive causal role *here* are present.

Conclusion: x will play a positive causal role *here* (for at least some individuals).

We know premise 1, it has an RCT (or several) to support it. But if a practitioner wants to base her own practice on that evidence, she needs to establish sub-arguments in support of premises 2 and 3 in order to bolster her conclusion. It should be noted here that both these premises require local knowledge – both of these thus fall outside Davies' model. The sub-argument for premise 2 demands a "vertical search" (Cartwright & Hardie, 2012, p.91); we describe the causal principle by which x contributed *somewhere* at a level of abstraction high enough that it connects to the context *here.* That is to say, we climb up from *somewhere* to an abstraction that applies in both places and then down to a concretization of the causal principle in our target context. If it cannot be made applicable for your local context, you obviously should not implement it. The sub-argument for premise 3, which I judge to be far more important in education, requires a "horizontal search" (p.91) to determine whether the support factors that figure in the causal principle (the *b*s), concretized for your context, are actually present or whether they can be (easily) procured should they be absent. If the necessary support factors indeed are in place and the causal principle is judged to be locally applicable, you may go ahead, implement x and hope for y.

This is a fairly complex practitioner tale; not one that is easy or straightforward to tell. It has the great advantage of weaving causality, evidence and local context into the same story. It casts RCT evidence in an *indirect* role as a sub-sub argument to the conclusion that x will work here. In fact, you can tell the story without even mentioning the research evidence – the most important evidence is local. Thus, we can say Cartwright's model illustrates the complexity of step 5 in Davies' model – how to determine the relevance of the evidence for your own purposes.

Toulmin's Model

Stephen Toulmin's model is an argumentation model (Toulmin, 2003) and lays out a pattern of justification of beliefs. It is about defending claims against challenges, not about using RCT evidence in practice. But the model is flexible and can accommodate different kinds of stories, including practitioner tales. The basic structure of the model consists of six factors and any practitioner tale will have to encompass them all: First we distinguish between the claim or conclusion (C) that we wish to establish and the facts we appeal to as foundation for C, called grounds or data (D) (p.90). Next we ask how we get from data to conclusion, from D to C. Propositions that provide justification for inferring C from D are called warrants (W). Warrants confer different degrees of strength on claims; hence we use qualifiers (Q) to express this: necessarily, probably, possibly, etc. Rebuttals (R) point to exceptional circumstances that undermine the general authority of the warrant, usually expressed as "unless". But how do we know that the warrant is tenable or trustworthy? Here we appeal to backing (B): "Standing behind our warrants, […] there will normally be other assurances, without which the warrants themselves would possess neither authority nor currency," Toulmin says (2003, p.96). Backing B is highly important for our purposes here. This is where a place can be found for research evidence.

But let us back up a bit. At the outset it would seem that this model can accommodate evidence in two different places; as data D and as backing B. If cast in the role of D, evidence can literally be interpreted as a *base,* as a foundation from which practice (in the role of C) can be inferred or even derived. This has been the concern of a good many critics of EBE. For example, Gert Biesta (2007) attributes to EBE advocates the view that:

> [R]esearch will be able to give us "the truth", that "the truth" can be translated into rules for action, and that the only thing the practitioners need is to follow these rules without any further reflection on or consideration of the concrete situation they are in (p.11).

In a similar vein, David Bridges, Paul Smeyers and Richard Smith (2008), provide the following description:

> The evidence-based policy movement seems almost to presuppose an algorithm which will generate policy decisions: *If* A is what you want to achieve *and if* research shows R1, R2 and R3 to be the case, and *if furthermore* research shows that doing P is positively correlated with A, *then* it follows that P is what you need to do. So provided

you have your educational/political goals sorted out, all you need to do is slot in the appropriate research findings – the right information – to extract your policy (p.9).

This implies understanding evidence as data D. Evidence is *directly* related to practice and you extract your policy (C) from it. Incidentally it also exemplifies Davies' step 5, making research relevant for one's own educational needs. I do not know what Davies would think about this way of cashing out step 5. It is certainly very different from Cartwright's cashing out of the same step.

     I shall argue that research evidence is best cast in the role of B. But we are not quite done with data D yet. I think it has an important role to play. More specifically, I would like to understand D as *practice-based evidence*. The phrase is Michael Eraut's (2004). He makes a useful distinction between research-based evidence and practice-based evidence: research-based evidence comes from published research satisfying the critical reviews in that area, practice-based evidence comes from "professional practices recognized by the relevant profession, and performed in accordance with the criteria expected by the relevant experts within that profession" (p.92). For my purposes here I need a somewhat wider definition and suggest that we understand practice-based evidence as various pieces of information picked up in practice on a daily basis, from observations, tests, interactions and intuitions, without "approval" of experts in the profession. These various pieces of information enter into an interpretation or a "diagnosis" of what is going on, for example that some of the first-graders are delayed in their reading toward the end of the first year. This diagnosis is the starting point of the reasoning that eventually will come to use research evidence as backing B. Most likely Eraut has greater general faith in the quality of practice-based evidence than I do, but that is of no concern here. It seems to me that such practice-based reasoning can be viewed as conductive reasoning; that is, as an argument in which a supervenient status is attributed to some subject, in this case a practice situation (Hitchcock, 2013). A supervenient status is typically evaluative, prescriptive, interpretive or classificatory, according to Hitchcock (p.206), and that is spot-on for my purposes here. A practitioner gathers information in practice and makes an interpretation of what is going on. And after that comes the question: "What should I do?"

     Thus, I suggest, the practitioner tale emerging from Toulmin's model also runs on a question, like Davies and Cartwright's models. I wish to make three observations here. First, "what should I do?" is a quite common question for practitioners to ask. Second, Davies' model cannot answer it, because research evidence in and of itself cannot tell you what to do.

Third, this practitioner tale begins in what is perceived as a problematic practice situation. This is different from Cartwright's model, which does not have anything like data D in it. In this respect it is also different from Davies' model, which has no hints as to where the question comes from.

In Toulmin's model the answer to the question of what one should do is represented by conclusion C. D under-determines C, so the same D is compatible with different Cs, which may lead to different chains of reasoning and therefore invoke different research-based evidence. A fictive example might look like this: A group of first-graders is judged to lag behind in their reading (D) so the practitioner decides to give them exercises to promote their decoding skills (C). She could be asked to substantiate her judgment that D, and reply that she has tested them using the Early Reading Screening Instrument (ERSI) (Morris, 1998) and that the test results are unambiguous. When asked how she got from D to C, she might say that promotion of decoding skills is an effective method (W). Warrants are important. They are the justification for inferring C from D. They are not data or evidence in themselves; rather they are rules which say that "data such as D entitle one to draw conclusions, or make claims, such as C" (Toulmin, 2003, p.91). But to do this job warrants above all need to be valid, reliable, of the right kind. To this end Toulmin introduces the concept of *backing*, as that which justifies the warrant and provides it with the authority it needs to license the step from D to C. Thus, if queried why she thinks promotion of decoding skills is a good strategy (W), our practitioner might say that there are several RCTs showing the effectiveness of this strategy in achieving the desired outcome (B) (Snowling & Hulme, 2011).

And there we have found a role for research evidence in the Toulmin model – as backing B. Again a couple of observations are in order. First, it does not in principle matter for the Toulmin model that B consists of one (or more) RCTs. This is what one would expect, I presume, since the model is an adapted argument model. Methodology is irrelevant; what matters is that B provides validity, reliability and authority to the warrant. This is of course different from both Davies and Cartwright, since their models are constructed to deal with RCT evidence. Second, as in Cartwright's model the role for research evidence is *indirect*. You do not derive conclusions about what to do directly from the evidence, but in its role as backing B it helps you to justify your conclusion about what to do. Third, in Cartwright's model the research evidence functions as a sub-sub premise and is not (necessarily) stated at all, as we have seen. In Toulmin's model it enters the chain of reasoning explicitly, and we can immediately see its contribution. Despite the same indirect role in relation to practice (or to answer the relevant question), research evidence occupies very different roles in their

respective argument structures. Cartwright illustrates her chain of reasoning by a pyramid with boxes for conclusion, premises, sub-premises and sub-sub-premises. Some of the boxes have only question marks in them (Cartwright, 2011, p.223) and premises that are taken to not need further sub-premises to support them have roots at the bottom (Cartwright & Hardie, 2012, p.17). The structure is built up on the principle that you constantly need to check and secure your premises. Thus, while both think in terms of argumentation, their proposed structures are rather different. As briefly described above, Toulmin lays out the following pattern of an argument, illustrated here with reading instruction: Some of my first-graders are delayed in their reading (D). What should I do? Since decoding exercises work well for reading-delayed students (W), something we know because several RCTs show their effectiveness (B), this remedy will probably (Q) help improve their reading comprehension. So unless (R) my students are very different from the study populations, I will design exercises to promote their decoding skills (C). Cartwright's pyramid makes for a practitioner tale that contains much more detail about the specific causal process in each RCT study and how to make it relevant for *here*, whereas Toulmin's model yields a tale that runs easily from a stated problem to a (reasonably qualified) conclusion.

Fourth and related to my first observation that methodology does not really matter for backing B; the Toulmin practitioner tale, presumably due to its oral character and easy flow, portrays RCT results as *general*. I think it is widely assumed that RCTs indeed show what works in general. Cartwright's model is constructed to accommodate the fact that RCTs do not tell you what works in general. RCTs require strong premises and are narrow in range of application since strong premises do not hold widely (Cartwright, 2007). Strictly speaking RCTs hold only for the sample. A Toulmin practitioner tale, as I have laid it out here, makes many assumptions about generality, causal principles that hold everywhere, transferability and similarity. On the other hand, and this is my fifth observation, Toulmin's model explicitly draws attention to conditions of exception that might obtain – a feature that might serve as an antidote to (over)generalization and too quick applications of method of instruction under consideration. Interestingly, the *a*s in Cartwright's equation implicitly point in the same direction: are the students *here* anything like the students *there*, do they have different starting points, etc.

Grimen's Model

This model is about professional practitioners: what kind of knowledge they have, how they organize it, and how they decide what to do. I shall use Harald Grimen's model to represent

the critique of EBE which takes its point of departure in the concepts of *phronesis* and *(practical) judgment*.

Let me begin by providing some context. In the conceptual landscape of much, but not all, EBE critique we find a vocabulary consisting of such terms as aims, intrinsic value, appropriateness, judgment, ethical considerations, phronesis, etc. This vocabulary is set up as a contrast to the EBE vocabulary, which consists of such terms as outcomes, what works, evidence, effectiveness, measurement, standards, accountability, etc. We might say that the two compete for the definition of professional practice and knowledge. A case in point is Richard Smith (2006), who views EBE as "technical reason" geared to finding the best means toward pre-determined ends. This way of thinking is purely instrumental, Smith argues: the ethical is relegated to the status of personal preference, and "the dominance of technical reason threatens to eliminate the idea of *judgement*" (p.162). Thus we have the two, EBE and judgment, pitted against each other, as they quite often are.

The idea of judgment that Smith refers to is intimately connected to the Aristotelian concept of *phronesis. Phronesis* belongs to a trio of concepts, alongside *episteme* and *techne*. Of special interest here is the relationship between *phronesis* and *techne. Techne* is the root of the term "technology", which incidentally may be one of the reasons why many EBE critics refer to EBE as "technical". *Techne* is the province of the craftsman who plans and makes things, Smith says (2006, p.165). It is dominated by instrumental reason. The aim falls outside the activity: the outcome, the product, is what is of interest, not the actions or the process. *Phronesis* is also a form of practical knowledge, but unlike *techne* the action itself is the aim, and that is so because the action is ethical. *Phronesis* denotes the ability to judge how one should act in order to promote what is morally good for people in concrete circumstances. This ability is exercised by the practitioner and focus shifts from outcome in terms of student achievement to actions performed by the practitioners. Whereas EBE is associated with *techne*, the alternative is associated with *phronesis.* I have already suggested that of the three broad purposes for education, EBE is suitable for qualification, the learning of knowledge and skills, only. One might think that the other two, initiation and self-formation, are suited for a *phronesis* approach. But judgment, Gert Biesta (2015) argues, covers all three purposes, in the sense that overall judgments must be made as to the dimensioning of the three, how to handle conflicts between them and to find proper trade-offs between them. If you think that educational practice is moral in nature, a phronetic approach seems well suited.

I would like to briefly interject here that judgments come in different forms and with different intentions, not only ethical ones; we have appraisive, evaluative, diagnostic,

decision-making judgments, etc. Such judgments are central to thought and inferences and are heavily involved in all three models discussed above; for example in the determination of relevance of evidence to one's own educational needs (Davies), in the estimation of the probability of an effectiveness prediction (Cartwright) and in the interpretation of practice-based evidence that sets off a chain of practical reasoning (Toulmin).

One sometimes gets the impression that all judgments are necessarily good because they emanate from practice. But Harald Grimen (2009, p.211) makes it clear that while judgment is necessary because it incorporates things that are highly valued in practice (such as individualized and fair treatment of students, and decision-making in situations where institutional regulations are ambiguous or knowledge is scant), it can also be bad, biased or unreasonable and have negative effects (such as abuse of power, unpredictability or arbitrariness). One also sometimes gets the impression that judgments are made without constraints. Again, Grimen is very clear that professional judgments are made within boundaries; contextual, institutional, legal and normative constraints. Such boundaries must be definitional of professional judgment, Grimen argues, to distinguish it from imagination or instinct.

In my judgment Grimen has fashioned a sober, albeit rather vague, model of professional knowledge bases. He dubs it *practical synthesis* (Grimen, 2008, p.72). In its bare bones it says that the sources of professional knowledge are heterogeneous, that they are fragmented, and that the various constituent elements are integrated in the form of practical syntheses. The synthesis is *practical*, as opposed to theoretical, because the integrating "force" derives from the concrete demands of practice. The heterogeneity in question comes from the fact that all client-oriented professions, including education, must use knowledge from many fields. It stands to reason that heterogeneous knowledge bases also are theoretically fragmented – professional practice is simply not geared to theoretical reflection or integration; its *raison d'être* is a different one. In addition we have the fact that professions have their aims outside themselves, Grimen argues. They do not exist for their own sake, but stand in the service of certain values. It is these external aims that determine the composition of the knowledge base, he thinks. In passing, let us briefly note here that this is a point of contention; whether educational activities are done for their own sake and thus constitute the aim, or whether they are done for the sake of something else. Grimen's view is at odds with most other EBE critics who defend judgment and *phronesis*, who also tend to accentuate intrinsic value and activity for its own sake over instrumental value and activity for external aims (e.g. Biesta, 2007, 2010; Carr, 1992; Elliott, 2004; Pring, 2004; Smith, 2006). If the

activity *constitutes* the aim we do not need research, let alone RCTs, to tell us how to attain the aim.

A practical synthesis is a meaningful whole. Different pieces and types of knowledge, be it theoretical, practical, experiential or tacit, are made to hang together because they act as parts in this whole, brought together by the demands of practice. The attainment of the aims creates the basis for the type of unity professional knowledge enjoys, Grimen argues. His own example is from the medical profession: Medical doctors need to master both biology and communication theory. Not because communication tells you much (if anything) about health and sickness, but because doctors must be able to communicate with their patients to diagnose illness or motivate for treatment. Biology and communication are not tied together through some comprehensive theory; they are tied together through practice – conversations with patients give doctors vital information. And, says Grimen (2008, p.74), it is the methodological requirements of diagnostics, not pathology, which make it necessary for doctors to communicate with their patients. The demands of practice thus make these different elements of knowledge hang together, by placing both in a meaningful whole. This, he suggests, is typical for the way in which heterogeneous pieces of knowledge are brought into relationship with each other in the professions.

This makes for a dynamic, but not chaotic model: the elements will shift with the demands of the job and the whole will shift with the elements; the whole not being the whole of practice but a delimited and focused whole. The making of such meaningful wholes is a form of practical knowledge, expressed in actions, decisions and *judgments*. As already said, I find this to be a sober model; it avoids the dichotomizing tendencies of much EBE criticism, and it is the closest thing I have seen to a *model* of judgment. Baked into the model are also nuanced views about theoretical knowledge, practical knowledge and tacit knowledge, along with a deep understanding of the nature of professional practice. On the other hand it is rather vaguely characterized. Grimen does not give his readers a fully worked out example of the considerations that are implied, nor does he suggest what kind of questions would be suitable as a starting point for construction of such a whole. Given his own example and his emphasis on practical knowledge, it seems reasonable to suggest that questions concern what is going on and what to do. Is there room for research evidence in this model? At the outset one might think that a heterogeneous knowledge base would not by definition preclude RCT evidence. But Grimen is skeptical of EBE. He thinks accumulation of research results and RCT evidence represents an illegitimate simplification of professional knowledge bases (2008, p.83) and that it does not do justice to professions and their missions. This criticism hits

Davies' model, which depicts one source of knowledge only for practitioner questions. His main concern, however, is the wider professional-political implications of EBE and in particular the new epistemic division of labor (2009, p.202ff). Practitioners, he argues, lose control over their own professional knowledge to upcoming new professions or actors in the field (e.g. developers of tests or interventions who offer their products to schools or school boards); to manufacturers of software (e.g. digital learning platforms which shape educational practice); and to organizations that produce systematic reviews and provide guidelines for practice (e.g. the Clearinghouses). Both practitioners and policy-makers become dependent on these organizations, Grimen argues, and the belief that they master advanced statistics and are capable of evaluating the quality of systematic reviews amounts to wishful thinking. Here again he differs from Davies.

But, I claim, one can reasonably use RCT evidence, or the causal hypothesis (method of instruction) that the evidence is evidence *for*, without buying the whole EBE package. If so, evidence would enter into a practical synthesis as one of its parts, alongside other forms of knowledge and considerations. Where do you begin to fashion a practical synthesis? In the demands of practice. What is the end point? To judge by Grimen's own example, a diagnosis is a possible outcome or conclusion, or a decision to involve students in specific learning activities. But in-between? Davies' models runs on steps, Cartwright's model builds premises for an effectiveness prediction, and Toulmin's model has an internal drive from practice-based evidence to a conclusion about what to do. Grimen's model is not very helpful in this regard, as I said above it is not really worked out. Whatever internal structure practical syntheses possess, it comes from the practitioner's understanding of the demands of practice.

Let us again enlist the aid of argumentation theory and view practical syntheses as conclusions of conductive reasoning. This might help clarify the picture somewhat. In my discussion of Toulmin's model I suggested that practice-based evidence (D) could be understood as the conclusion of conductive reasoning, since it amounts to an interpretation of what is going on. I would now like to suggest that a practical synthesis as a whole can be understood as the conclusion of conductive reasoning (Hitchcock, 2013). Such conclusions attribute a supervenient status to the subject of interest, for example a diagnosis of being reading-delayed attributed to a first-grader, on the basis of factors that the practitioner takes to account for or against the first-grader's having that status. This status is taken by the practitioner to be constituted by a complex of types of considerations and counter-considerations. In conductive reasoning we thus cite features of the subject (the first-grader's performance) that count for and in some cases against the conclusion (that she is reading-

delayed). Counting for the conclusion might be results on the ERSI test, counting against it might be observations that the first-grader in question reads happily when believing herself to be unseen by the teacher. One can imagine a good many different considerations being brought to bear on the subject, more or less relevant ones, some relevant ones might be missing and some irrelevant ones might be falsely introduced as relevant. Since Grimen's model is based on *phronesis* it stands to reason that ethical, practical and contextual considerations will take center stage. Conductive conclusions, especially if they take counter-considerations into account, use such phrases as "taking all this into account" or "therefore, on balance" – phrases that indicate a *whole*. Hitchcock (2013, p.226) prefers to call this form of reasoning an appeal to considerations (or to criteria), to avoid restricting conductive arguments to the weighing of pros and cons. However that may be, conductive reasoning is a complex process and conductive conclusions are *judgments.* Most likely many of the considerations that function as premises are judgments as well. The correctness of conductive conclusions, Hitchcock argues, cannot be judged independently of the sorts of considerations adduced to support it. Thus,

> Only after this initial check on the independent relevance of each feature cited, and on the acceptability of each claim that the subject of interest has that feature, can one determine whether the acceptable and independently relevant features on balance provide sufficient support for attributing the supervenient status (2013, p.211).

It may be impossible to assess the correctness of practical syntheses, since they include practical and even tacit elements. Practical knowledge is acquired and exercised only in a first person singular perspective, Grimen argues, and it is largely specific to the activity. Tacit knowledge either is not or cannot be articulated verbally, but is present because there are always things in practice we must take for granted. It influences our thought, experience and actions (2008, p.79). To the extent that practical and tacit knowledge enter into conductive reasoning, the correctness or adequacy of the conclusion is likely beyond our reach. A practitioner tale spun from a practical synthesis might not be entirely articulable.

Tales Not Told

In this section I wish to explore briefly whether there are important tales that cannot be told using RCT evidence or that cannot be accommodated by the models. Let us note first that all four models are prospective, forward-looking, dealing in expectation. Will this work? What

should I do? Grimen's model in addition has an air of the here-and-now about it. This of course fits with education, which in its nature is future-oriented.

Several issues are intertwined in these untold tales; individuals, counterfactuals, retrospection and causal models. Randomized controlled trials standardly yield a treatment effect, a causal effect that is averaged over the (study) population (strictly speaking over the sample). It is important to note that an overall positive average result is compatible with negative results for some of the individuals. RCTs run on an input-output logic; we know the input x and we measure the output y (in educational contexts y usually is cashed out as student achievement, viewed as performance on standardized tests). Keith Morrison (2001) describes RCTs as black-boxes, disabling any identification of the detailed causal mechanism responsible for the observed effect (p.73). In other words, input and output, cause and effect, are not (explicitly) placed in a causal model. Morrison attributes to EBE and RCT advocates the view that both theory and causal models are uninteresting and unnecessary, since the data themselves show us what works. David Hargreaves, as we have seen, attributes the same view to teachers: they want to know *what* works, he claims, only secondarily are they interested in *why* it works. *What* can signal many kinds of questions, including those that yield conductive conclusions. Nancy Cartwright argues that practitioners need not only evidence *that* x worked somewhere, but an understanding of *how* it worked. This requires figuring out how x and its support factors, in situations where x has been shown to work, contribute causally to the desired result. That is, how x in combination with as, bs and zs (and probably some us, unknown factors) contribute positively to the result. RCT evidence can thus provide *part* of an answer to the question "how can I make y happen?" but it can by no means provide the whole answer by itself, as we have seen. It is unclear to me whether Grimen's model can answer the question of how I can make y happen. A practical synthesis is a tale of a practitioner who judges what practice demands, makes decisions accordingly and carries out appropriate responses. Thus, you *respond*, you do not initiate.

Let us turn now to *why. Why* is a causal question. As Morrison observes, RCTs measure input and output but provide no causal mechanism and cannot answer the general question of why x leads to y. Why-questions are tricky. Consider the question "why is this first-grader lagging behind in his reading?" Daniel Kahneman and Dale Miller (1986) suggest that answers to why-questions, at least the sort that interests us here, invariably invoke norms. This is because why-questions are premised on a contrast between an observation, say y_2, and the norm or expected result, say y_1. That is to say, we compare what *is* to what *should have been*. The field of education with its focus on desired changes is of course full of norms, some

explicit, some implicit. Norms may emerge from practitioner experience; or it may emerge from research, from specified standards of reading for first-graders; or perhaps from a meta-analysis showing that y_1 is the normal – average – result for first-graders who have been taught to read using method x_1, in which case a treatment effect has been turned into a norm. But why is this particular student lagging behind, when he really should not have been? Questions of this kind, which concern individual students, are of obvious interest to the practitioner, but cannot be answered by RCT evidence. However there is more to it than that and I shall inquire briefly into the question drawing on Kahneman and Miller's discussion of perspective differences. Suppose our first-grader is mildly dyslexic. He has been tested using the Early Reading Screening Instrument (ERSI) and he scored very low on all four task areas (alphabet knowledge, concept of word, phoneme awareness, word recognition) (Morris, 1998). The practitioner may identify the student's test anxiety as the cause of the result; an external reading expert (or a researcher) may identify dyslexia as the cause and the test anxiety as occasional. This kind of situation illustrates both the importance of norms and the ambiguity of causal why-questions, Kahneman and Miller argue (1986, p.149-150). The practitioner, who knows the student, is concerned with an intrapersonal norm: this student normally does not suffer from test anxiety, so why did he do it on that particular day, for that particular test? An external expert (or a researcher) will invoke an interpersonal norm: why does this student lag behind when the others do not? They use different norms and have different perspectives, and even use different causes to explain the same event. The practitioner selects a feature of the situation that distinguishes this situation from other situations; the external expert selects a feature that distinguishes this student from other students. This, I submit, reflects a difference in purpose, interest and perspective of practitioners and researchers.

Why-questions of the kind discussed by Kahneman and Miller are retrospective. They are generally asked when an event is perceived to be unexpected and in violation of some norm. The question "why is our first-grader lagging behind in his reading?" thus indicates a belief that he should not have been lagging behind. Reality is compared to an alternative, as Kahneman and Miller put it, and out of that comparison arises the why-question. There is a short step from there to the question of what would have happened if the first-grader had been taught to read using method x_3 instead of x_1. This is a retrospective, counterfactual question and I have already indicated that RCT evidence cannot answer it. As described by Judea Pearl and Dana Mackenzie,

> … to answer [counterfactual questions] we must go back in time, change the history, and ask, "What would have happened if I had not taken the aspirin?" No experiment in the world can deny treatment to an already treated person and compare the two outcomes, so we must import a whole new kind of knowledge (2018, p.33).

What the language of counterfactuality gives us is the possibility to talk about causation at the level of individuals and particular events. Counterfactuals can for obvious reasons only be imagined and not observed, and yet, Pearl and Mackenzie point out, they explain the causal asymmetry we intuitively attribute to many observed correlations. Roosters crow at sunrise, but we do not believe that the rooster's crow causes the sun to rise: "We know that if the rooster was sick one day, or capriciously refused to crow, the sun would rise anyway" (p.266).

We all use counterfactual talk every day, also in professional practice, for example to explain events and assess results by comparing reality to its alternatives. Grimen's model is interesting in this respect – does it support counterfactuals? My inclination is to say no. Counterfactual reasoning presupposes a causal model. You need to have some idea of how things hang together before you can change some factor mentally and imagine how that would affect other factors and change the result. But I find no traces of causation in practical syntheses. Critics of EBE who promote *phronesis* and judgment are generally skeptical of causation – causation and counterfactual reasoning belong to *techne*, not to *phronesis.*

Conclusion

I think practitioners like Grimen's model. Emphasizing practical and tacit knowledge, it is a tale that allows the practitioners themselves to take center stage. Individuals are in the foreground rather than populations, and context is vital – that is where the demands of practice come from, and where the practitioner makes her judgments. To me it is a weakness that the model, as I understand it, cannot support counterfactual reasoning because it does not incorporate a causal model. RCT evidence may enter into professional knowledge bases; these are heterogeneous and fragmented in nature and are kept together in delimited meaningful wholes geared to meeting the practical demands. If RCT evidence is allowed, it would make up a *part* of the knowledge base and probably play an indirect role parallel to what we found in Cartwright's and Toulmin's models.

Cartwright's model accords an indirect role to research evidence, as a sub-sub-premise to an effectiveness prediction. The model rests on an explicit understanding of causation, an

explicit understanding of the scope of RCT evidence, and an explicit belief in the diversity of the world. It represents a sophisticated way of reasoning and is quite demanding to use – indeed it requires much sound judgment in practitioners. Toulmin's model, as I have used it here, also accords an indirect role to RCT evidence, as backing (B) of a warrant (W). Toulmin's model is an argumentation model, and Cartwright's model too is fashioned on an argument structure. Both illustrate the power of argumentation theory. I think Davies' model is more difficult to use than he probably imagines. In the absence of spelled out examples it is hard to assess the model. It seems to accord a direct role to RCT evidence, which I find problematic because you cannot derive policy or practice from evidence. Moreover, the model does not mention causation or the causal hypotheses that the evidence is evidence *for*. The model can clearly answer "what works" questions and also "how can I make y happen" questions, if the causal hypothesis is implemented. But there is a whole range of important questions RCT evidence cannot answer, including why-questions, counterfactual questions concerning individuals and questions such as "what should I do?".

Of the four models discussed, Toulmin's model has become my favorite. It can answer practically relevant questions, such as "what should I do?", it has a narrative quality to it that gives it a nice flow from beginning to end, it has a good place for RCT evidence, it accommodates context and individuals (through its "unless" clause), and it is reasonably adequate. Of the four possible practitioner tales, this is the easiest one to tell.

References

Bertalanffy, L. von (1984). *General System Theory. Revised Edition.* New York: George Baziller.

Biesta, G. (2007). Why «what works» won't work: Evidence-based practice and the democratic deficit in educational research. *Educational Theory, 57,* 1, 1-22.

Biesta, G. (2010). Why "what works" still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education, 29,* 5, 491-503.

Biesta, G. (2015). What is education for? On good education, teacher judgement, and educational professionalism. *European Journal of Education, 50,* 1, 76-87.

Bridges, D., Smeyers, P. & Smith, R. (2008). Educational research and the practical

judgement of policy makers. *Journal of Philosophy of Education, 42*, Supplementary Issue 1, 5-11.

Carr, D. (1992). Practical enquiry, values, and the problem of educational theory. *Oxford Review of Education, 18,* 3, 241-251.

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

Cartwright, N. (2011). Predicting what will happen when we act. What counts for warrant? *Preventive Medicine, 53,* 221-224.

Cartwright, N. (2012). Will this policy work for you? Predicting effectiveness better: How philosophy helps. *Philosophy of Science, 79,* 5, 973-989.

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

Cunningham, J. W. (1993). Whole-to-part reading diagnosis. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 9,* 31-49.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47,* 2, 108-121.

Elliott, J. (2004). Making evidence-based practice practical. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (164-186). Maidenhead: Open University Press.

Eraut, M. (2004). Practice-based evidence. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (91-101). Maidenhead: Open University Press.

Grimen, H. (2008). Profesjon og kunnskap [Profession and knowledge]. In A. Molander & L.I. Terum (Eds.), *Profesjonsstudier [The Study of Professions]* (71-86). Oslo: Universitetsforlaget.

Grimen, H. (2009). Debatten om evidensbasering – noen utfordringer [The debate about evidence-based practice – some challenges]. In H. Grimen & L.I. Terum (Eds.), *Evidensbasert profesjonsutøvelse [Evidence-Based Professional Practice]* (191-222). Oslo: Abstrakt forlag.

Hammersley, M. (1997). Educational research and teaching: A response to David Hargreaves' TTA lecture. *British Educational Research Journal, 23,* 2, 141-161.

Hammersley, M. (2005). Countering the "new orthodoxy" in educational research: A response to Phil Hodkinson. *British Educational Research Journal, 31,* 2, 139-155.

Hargreaves, D. (1996). Teaching as a research-based profession: Possibilities and prospects. Teacher Training Agency Annual Lecture, London.

Hitchcock, D. (2013). Appeals to considerations. *Informal Logic, 33,* 2, 195-237.

Kahneman, D. & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93,* 2, 136-153.

Lervåg, A. & Aukrust, V. G. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *Journal of Child Psychology and Psychiatry, 51,* 5, 612-620.

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

Morris, D. (1998). Assessing printed word knowledge in beginning readers: The Early reading Screening Instrument (ERSI). *Illinois Reading Council Journal, 26,* 2, 31-40.

Morrison, K. (2001). Randomised controlled trials for evidence-based education: Some problems in judging "what works." *Evaluation & Research in Education, 15,* 2, 69-83.

Pearl, J. & Mackenzie, D. (2018). *The Book of Why. The New Science of Cause and Effect.* New York: Basic Books.

Phillips, D. C. (2019). Evidence of confusion about evidence of causes. Comments on the debate about EBP in education. *Educational Research and Evaluation, this issue*

Pring, R. (2004). Conclusion: Evidence-based policy and practice. In G. Thomas & R. Pring (Eds.), *Evidence-Based Practice in Education* (201-212). Maidenhead: Open University Press.

Pring, R. (2015). *Philosophy of Educational Research.* London: Bloomsbury.

Simpson, A. (2019). Separating arguments from conclusions: The mistaken role of effect size in educational policy research. *Educational Research and Evaluation, this issue*

Smith, R. (2006). Technical difficulties: The workings of practical judgment. In P. Smeyers & M. Depaepe (Eds.), *Why 'What Works' Doesn't Work* (159-170). Dordrecht: Springer.

Snowling, M. J. & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology, 81,* 1-23.

Toulmin, S. E. (2003). *The Uses of Argument. Updated Edition.* Cambridge: Cambridge University Press.

US Department of Education (2003). *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide.* Washington, DC: Coalition for Evidence-Based Policy. https://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf