EBP: Where Rigor Matters.

Nancy Cartwright and Alexandre Marcellesi[1]

1. A plea for rigor in evidence-based policy

Pat Suppes teaches two great lessons about rigor:

- Rigor matters.
- A little rigor can be a dangerous thing.

Our focus in this paper is on an area where rigor is badly needed and where it is highly touted but where Nancy is in trouble for insisting on it: the movements that go under the labels 'evidence-based medicine' and 'evidence-based social policy', which typically assert---in the interest of rigor---that randomized controlled trials (RCTs) are the gold standard in evidence. (We shall refer to both together as 'EBP' for short.) For instance, at a recent conference on evidence and causality, Sir Iain Chalmers, who was founding director of the UK Cochrane Collaboration (which oversees evidence-based medicine) and whose knighthood is for his contributions to healthcare, put some formulas of Nancy's, similar to ones we shall use here, up on the screen and made fun of them as both useless and unintelligible. But EBP is an area where rigor matters. The

[1] Both authors would like to thank the Templeton Foundation's project 'God's Order, Man's Order and the Order of Nature' as well as the AHRC project 'Choices of evidence: tacit philosophical assumptions in debates on evidence-based practice in children's welfare services' for support for the research and writing of this paper. Nancy Cartwright would in addition like to thank the British Academy, which funded her project 'Evidence for Use', as well as the Institute for International, Comparative and Area Studies at UC San Diego for the 'Political civility and scientific objectivity' grant.

problem is that Suppes's second lesson is often ignored. There's too much rigor at one stage of the argument it takes to support a policy recommendation and very little thereafter---and that without mention. Yet it is well known that an argument is only as strong as its weakest premise. Downplaying the other premises that do not have solid support and stressing the rigor of the one premise gives a false sense of security. And this can be dangerous when policy decisions are at stake.

The EBP movement has generated a number of evidence hierarchies, grading systems for evidence, organizations and methods to review evidence pertaining to proposed treatments/interventions/policies, and warehouses where policies that pass review can be found. The evidence hierarchies rank not individual pieces of evidence but rather methods for the production of evidence, with well-conducted RCTs or systematic reviews or meta-analyses of well-conducted RCTs at the top. The hierarchies are then deployed by the various review organizations to evaluate how strongly supported treatments or policies are.

What you should have noticed right away is how vague our description is. Evidence is always evidence for some specific claim. Treatments aren't the kinds of things that have evidence. So what are the claims about treatments or policies that evidence regarded as good by the EBP movement is supposed to be good for? There's the rub. It is really difficult to get a clear statement about this. We'll give a short survey of some of the major sites and what they say in section 2. In section 3 we will look at the favored method in the EBP movement---the RCT. We will first propose one general form that an evidence claim that an RCT produces can take: "C causes E in (study population) A" or, to use a loose slogan: "It works somewhere". Second,

using a version of Suppes's probabilistic theory of causality, we will sketch one rigorous argument that RCTs can certify this kind of claim in the ideal. Section 4 first proposes a form for the final hypothesis EBP is trying to provide evidence for: "C causes E in (target population) A'" or, "It will work here". Second we will sketch a rigorous argument for establishing claims of this form taking the "It works somewhere" claims RCTs can establish as a starting premise. Thus we restore rigor by laying out an entire argument that starts with the assumption of a successful outcome in an ideal RCT and ends with the desired conclusion.

So…What's to make fun of about this?

2. What the 'rigor' in EBP frameworks looks like

2.1 Grading of Recommendations Assessment, Development and Evaluation (GRADE)

The GRADE working group is one of the most prominent advocates of the use of standardized grading schemes to assess the quality of evidence. It is particularly influential in evidence-based medicine: its grading schemes have been adopted by healthcare organizations such as the World Health Organization and the American College of Physicians, but also by vetting agencies such as Iain Chalmers' own Cochrane Collaboration.

GRADE offers two different grading schemes, one for the strength of recommendations to treat (strong, weak) and one for the quality of evidence (high, moderate, low, very low). A recommendation for a treatment is the output of a decision process that takes three elements as inputs: (i) an analysis of the (health-related) costs and benefits of this treatment, (ii) an assessment of the quality of the evidence supporting this cost-benefit analysis, and (iii) the

values and preferences of patients. What the GRADE evidence scheme ranks is the quality of the evidence supporting the estimates of the benefits and harms of a treatment that are to feed into the cost-benefit analysis for this treatment.

These estimates of benefits and harms are estimates of treatment effects. Witness, for instance, what the GRADE authors says about the impact of deficiencies in RCTs on the quality of evidence: "Our confidence in the evidence decreases if the available randomized controlled trials suffer from major deficiencies that are likely to result in a biased assessment of the *treatment effect*." (http://www.gradeworkinggroup.org/FAQ/index.htm, emphasis added).

GRADE's approach to grading evidence is detailed in two series of articles published in the *British Medical Journal* (*BMJ*) and *the Journal of Clinical Epidemiology* (*JCE*). Throughout these articles, the authors of GRADE talk about 'the effects of treatments' in general, not in particular populations at particular times, implicitly assuming a 'narrow' conception according to which the effect of a treatment depends exclusively on what the treatment is and not on who receives it at what time and in what setting. The only place in which populations are discussed is the *JCE* article warning users of GRADE about the threats of 'indirect' evidence.

One way in which evidence for the effectiveness of an intervention can be indirect is by there being a difference between the target population and the study populations (Guyatt et al. 2011b, §2.1). The GRADE authors make the following recommendation about how to react when a worry of indirectness arises:

> In general, one should not rate down [evidence] for population differences unless one has compelling reason to think that the biology in the population of interest is so different from that of the population tested that the magnitude of effect will differ substantially. Most often, this will not be the case. (op. cit., 1304-1305)

The GRADE authors do not provide a justification for the claim made in the last sentence of this quotation and, most importantly, they ignore potential behavioral and environmental differences between populations that may make an important difference to the effect of the treatment (just think of a case in which the condition targeted is high blood pressure, for instance). The problem created by the possible existence of relevant differences between study and target populations is thus dismissed without much argument. And users of GRADE are advised to make the default assumption that estimates of treatment effects obtained from a particular study population at a particular time can easily 'travel' to other populations and other times.

It is true that, without making this assumption, interpreting the results of meta-analyses becomes very difficult. Consider the example of a meta-analysis of the effect of antibiotics on acute otitis media in children given in (Guyatt et al. 2011a, 387-388, tables 1 & 2). The GRADE authors report estimates for various effects of this treatment, e.g. pain at 24h, that are aggregates of the estimates produced by several studies, e.g. five RCTs for pain at 24h. Unless one assumes that these five estimates produced by RCTs conducted on five distinct study populations are all estimates of the same thing, something like the effect of antibiotics on pain at 24h for children suffering from acute otitis *in general*, i.e. in any population, then it makes little sense to aggregate them into a single estimate. The problem, of course, is that neither the problematic assumption that there is such a thing as the effect of antibiotics on pain at 24h *in general* nor the

assumption that the five estimates that are aggregated are estimates of *this* particular effect are supported by either argument or evidence.

This does not prevent GRADE from rating the quality of the evidence supporting this aggregated estimate as 'High'. The reason for this rating is that the five studies from which this estimate is obtained all *individually* score highly on the GRADE criteria for quality of evidence, since they are all RCTs with no serious limitations, no serious inconsistencies, etc. According to the GRADE framework, to say that the quality of the evidence supporting this aggregate estimate is 'High' is to say that, "We are very confident that the true effect lies close to that of the estimate of the effect." (Balshem et al. 2011, 404, table 2).

But the "true effect" of what? Of the treatment in the particular population you are interested in treating? Of the treatment *in general*, i.e. in any population you might want to treat? Of the treatment in a superpopulation composed of the five study populations involved in the five RCTs from which the aggregate estimate was obtained? Of the treatment in a superpopulation composed of the populations which the five study populations were sampled from? Without a clear and principled answer to this question, it is difficult to interpret the aggregate estimate produced by the meta-analysis presented by the GRADE authors and, as a consequence, it is difficult to assess the quality of the evidence supporting this estimate. The evidence supporting the aggregate estimate *might* be good if what the "true effect" is happens to be the effect of the treatment in a superpopulation composed of the five study populations, but need not be if it is the effect of the treatment in the population you are interested in treating.

2.2 Oxford Center for Evidence-Based Medicine (CEBM)

The CEBM promotes EBP and produces evidence grading schemes. The most recent CEBM levels of evidence (http://www.cebm.net/index.aspx?o=5513) offer different rankings of study designs, and of the evidence they produce, depending on the question one is interested in answering (e.g. Does this intervention help? What are the common harms? Etc.) The evidence that gets ranked by the CEBM levels of evidence thus is assumed to be evidence supporting particular answers to these questions.

Consider the question that most resembles ours: 'Does this intervention help?'. The CEBM rankings tell you that the best evidence for answers to this question is produced by systematic reviews of RCTs, systematic reviews that often take the form of meta-analyses. Just as in the case of GRADE, then, meta-analyses of RCTs are considered to produce the best evidence. But just as in the case of GRADE, the CEBM levels of evidence do not tell you exactly what the evidence produced by these meta-analyses of RCTs, or by RCTs individually, is supposed to be evidence for.

One will say: ‘But they are evidence that the intervention helps (or doesn’t help)!’ What does it mean, however, to say that an intervention ‘helps’? The same questions arise as before: Is it to say that it helped in some study population in which it was implemented? That it helps in every population in which it is implemented? That it will help in the population in which you intend to implement it? Again, as in the case of GRADE, it is not clear how one can rate the quality of evidence without a clear answer to this question. An individual RCT might provide very good evidence if the question asked is whether the intervention helped in the study population on

which this very RCT was conducted, but not if the question asked is whether the intervention helps in every population in which one might implement it.

2.3 California Evidence-Based Clearinghouse for Child Welfare (CEBC)

The CEBC is a vetting agency that, unlike the Cochrane Collaboration, has its own evidence grading scheme. This grading scheme, which its authors call a 'Scientific rating scale', has five levels (from 'Well-Supported by Research Evidence' to 'Concerning Practice'), each level being defined by a list of criteria (http://www.cebc4cw.org/ratings/scientific-rating-scale/). According to its authors, this 'Scientific rating scale' is "a 1 to 5 rating of the strength of the research evidence supporting a practice or program." Of course, as for the EBP frameworks considered above, it is not made clear whether what is assessed is the evidence supporting the effectiveness of the program in the study population, in the target population, or in any population one might want to treat.

A quick look at the grading scheme, however, is enough to see that even their own vague description is not right. The CEBC grading scheme mixes together (i) whether a program's positive effects outweigh its negative effects and (ii) the strength of the evidence supporting this cost-benefit analysis. The fifth and lowest level, for instance, clearly is not a level of quality of evidence. This level, called 'Concerning Practice', is the level at which should be classified interventions such that "the overall weight of evidence suggests the intervention has a negative effect upon clients served". To present this 'Scientific rating scale' as ranking evidence thus is misleading and has the potential to confuse its users.

The tutorial video accompanying the CEBC evidence grading scheme does little to clarify the way this ranking scheme is supposed to work (http://www.cebc4cw.org/ratings/scientific-rating-scale/scientific-rating-scale-tutorial/). Consider, for instance, its use of the metaphor of the 'solidness of evidence': This video tells you to see evidence as a foundation, with the five levels of the evidence ranking scheme corresponding respectively to rock, gravel, sand, water, and gas foundations (from level 1 to level 5). It does not tell you, however, what this foundation is supposed to be a foundation for: What are you to build on top of your evidence? And the metaphor of the 'solidness of evidence' also fails to be faithful to the content of the CEBC grading scheme since, as we argued above, this ranking is not properly seen as a ranking of evidence. Consider again the fifth level of the ranking: If the overall weight of evidence suggests that a practice has negative effects, then the verdict that the practice has negative effects is presumably not based upon a gas foundation (otherwise why trust that any practice ranked at this level *really* has negative effects?).

The CEBC ranking scheme provides a striking example of pretend rigor: The use of expressions such as "Scientific Rating Process" or "Scientific Rating Scale" that is "Based on a Continuum" to describe the CEBC grading scheme stands in stark contrast with the lack of rigor in either the grading scheme itself or in the explanations and tutorial video accompanying it.

2.4 Substance Abuse and Mental Health Services Administration's National Registry of Evidence-based Programs and Practices (NREPP).

The NREPP is an agency that vets policies in the domain of mental health. Like the CEBC, it has its own system for grading the quality of evidence. This system grades six aspects (Reliability of

measures, Validity of measures, Intervention fidelity, Absence of confounders, etc.) of studies on a 0.0-4.0 scale. The authors of NREPP's evidence grading scheme claim that, "NREPP's Quality of Research ratings are indicators of the strength of the evidence supporting the outcomes of the intervention. Higher scores indicate stronger, more compelling evidence." (http://nrepp.samhsa.gov/ReviewQOR.aspx) As in the cases examined above, however, it is not made clear what is meant by "the intervention". Is it the intervention as it was implemented in the study population? The intervention as it might be implemented in any population? The intervention as it might be implemented in the population you are interested in?

Some of the criteria that serve to grade studies, moreover, are stated in rather vague terms. To get a 4 on 'Appropriateness of [statistical] analysis', for instance, a study must satisfy the following conditions: "Analyses were appropriate for inferring relationships between intervention and outcome. Sample size and power were adequate." (http://nrepp.samhsa.gov/ReviewQOR.aspx) What does it mean for sample size and power to be "adequate"? And adequate for what purpose? In their presentation of it, the authors of NREPP's grading system do not state clearly whether the six different scores a study receives are to be aggregated (nor, if so, how) to give an overall 'Quality of Research' rating. Looking at the NREPP's database of evaluations, however, one notices that the six scores received by a study are in fact aggregated. How are they aggregated? Again, no explicit information is given regarding the method followed. In fact, the overall 'Quality of Research' rating attributed to a study simply is the average of the six scores it receives.

This is an odd choice, since not all the criteria determining the quality of the evidence produced by a study seem equally important. 'Intervention Fidelity', which requires that the intervention be implemented exactly as it was designed to be, does not seem nearly as important as (controlling for) 'Potential Confounding Variables' for instance. One would think a weighted average to be more appropriate. One might even think that a study that receives a score of 0 on 'Potential Confounding Variables' should receive an overall score of 0.

So, not only is it not clear what the evidence ranked by the NREPP's scheme is supposed to be evidence for, it is also not clear why anybody should believe that higher overall scores "indicate stronger, more compelling evidence."

2.5 Scottish Intercollegiate Guidelines Network (SIGN)

SIGN is a vetting agency that produces guidelines or recommendations regarding healthcare policy for Scotland's National Health Services. It is a user, rather than a producer, of evidence grading schemes. SIGN, like the Cochrane Collaboration, has adopted (in 2009) GRADE's scheme for grading evidence. It is interesting to see how a vetting agency such as SIGN understands the evidence grading scheme it relies on. The authors of SIGN's handbook explicitly take the evidence grading scheme used by their framework to rank evidence for *effectiveness predictions*:

> It is important to emphasise that the grading does not relate to the *importance* of the recommendation**,** but to the strength of the supporting evidence and, in particular, to the predictive power of the study designs from which these data were obtained. Thus, the grading assigned to a recommendation indicates to users the likelihood that, if that

> recommendation is implemented, the predicted outcome will be achieved. (SIGN 2011, 34, emphasis original)

This passage illustrates the assumption that seems to underlie EBP frameworks in general, including the ones examined above, namely that the 'best' study designs (i.e. systematic reviews of RCTs, or individual RCTs) automatically and straightforwardly produce evidence that is relevant, and sufficient, to warrant predictions regarding the effectiveness of policies that have yet to be implemented.

Unfortunately, this assumption is mistaken and one needs an argument of the kind to be presented in section 4 in order to go from the result of an RCT to a well-supported effectiveness prediction. There is thus little sense in talking about the "predictive power" of systematic reviews of RCTs, for instance, and in interpreting evidence grading schemes as ranking studies according to their predictive power.

2.6 So what?

The five EBP frameworks considered above all value methodological rigor highly. This is why they systematically rank RCTs (and systematic reviews of RCTs) at the top of the evidence grading schemes they use and expert opinion at the very bottom. What's more rigorous than a well-conducted RCT? And what's less rigorous than unchecked opinion, even if that of an expert? What an examination of a sample of these frameworks reveals, however, is that they lack rigor in key places. If you put forth a scheme for grading evidence, then, unlike the CEBC scheme, it should rank evidence, and evidence only. If you give an overall numerical score to studies, as the NREPP rating system does, then you should clearly explain how this score is

computed and justify the choices involved in this computation. Most importantly, your evidence grading scheme should state clearly what the evidence it ranks is evidence for.

All the evidence grading schemes considered above equivocate on this last point. They never clearly answer the questions we keep underlining: What is the evidence ranked evidence for? Is it evidence that the intervention was effective in the study population? Is it evidence that the intervention will be effective in most population? Is it evidence that the intervention will be effective in every population? Is it evidence that the intervention will be effective in the population in which you want to implement it? We repeat this point once more not at the risk of boring the reader because it is of crucial importance. Ranking evidence without a clear answer to this question is vain. You cannot evaluate how good the evidence for a particular claim is unless you are clear on what this claim says.

What we do below is to explain what claims the evidence ranked by evidence grading schemes is generally in fact evidence for and explain why RCTs are thought to be very good at supporting claims of this kind. We also explain how to bridge the gap from the kinds of claims supported by RCTs, i.e. claims of the form "It works somewhere", to predictions of the effectiveness of interventions, i.e. claims of the form "It will work here".

3. The starting point: RCTs

3.1 The probabilistic theory of causality

What' so good about RCTs? They are supposed to be a very good way---some insist, the only way---for controlling for unknown confounders. See for instance what the webpage of MIT's

Abdul Latif Jameel Poverty Action Lab (J-PAL) has to say about RCTs: The reader is told that "randomized evaluations do the best job" at controlling for unknown confounders because they "generate a *statistically identical* comparison group, and therefore produce the most accurate (unbiased) results" while "other methods often produce misleading results" (http://www.povertyactionlab.org/methodology/why/why-randomize). And how did 'confounders', known or unknown, enter the discussion? Let's start back, way behind where the usual defence of RCTs begins, to get a more rigorous account. Confounders enter when we are trying to establish causal claims. So we shall begin with the probabilistic theory of causality. We shall not, though, use the theory in exactly the form Suppes first put it, but in a modified version Nancy has developed building from Suppes' account (Cartwright 1979). For simplicity we will consider only yes-no variables.

Suppose then that the notion of causality at stake satisfies the following constraint:

> *Probabilistic causality*: For any population A, C causes E in A iff for some A(i) ⊆ A every member of which satisfies K_i, $P_{A(i)}(E/C\&K_i) > P_{A(i)}(E/\neg C\&K_i)$ where K_i is a state description[2] over a full set of causes of E, barring C itself.

The expression 'a full set of causes' takes some further paraphernalia to characterize. It can be done either relative to a formulation of the causal principles that govern A or to a set of causal pathways into E that obtain for A, where a full set of causal factors for E will contain one node from every pathway into E. We shall here leave it undefined. The factors that go into the Ks are just the 'confounding factors' that RCT advocates are concerned about. We shall use the term

[2] A state description over factors $A_1,\ldots,A_n$ is a conjunction on n conjuncts, one for each A_i, with each conjunct either A_i or $\neg A_i$.

'causal structure' from now on. A causal structure for outcome E in A is a set $\{C_A, P_A\}$, where C_A is a full set of causal factors for E in A and P_A is a probability measure that holds in A over the space generated by $C_A \cup \{E\}$.

Note first that this theory uses the notion of causality on the right-hand-side and hence cannot provide a reductive definition for causation. It does however provide an important constraint between probability and causality, which is a good thing for our enterprise since the immediate results of RCTs are statistics. Second, a direct application of the formula requires a huge amount of antecedent causal knowledge before information about probabilistic dependencies between C and E can be used to determine if there is a causal link between them. The RCT is designed specifically to finesse our lack of information about what other causes can affect E. Third, the theory allows that C may both cause E and prevent E (i.e., cause ¬E) in one and the same population, as one might wish to say about certain anti-depressants that can, it seems, both heighten and diminish depression in teenagers. This is especially important to note when it comes to RCTs since the effect size measured in an RCT averages over different arrangements of confounding factors so that the cause may increase the probability of the effect in some of these arrangements and decrease it in others and still produce an increase in the average.

3.2 Ideal RCTs

We shall describe the simplest basic structure, to make the argument outline clear. RCTs have two wings---a treatment group where every member receives the cause under test and a control group, in which any occurrences of the cause arise 'naturally' and which may receive a placebo. In the design of real RCTs three features loom large:

a. *Maskings* of all sorts. The subjects should not know if they are receiving the cause or not; the attendant monitors should not know; those identifying whether the effect occurs or not in an individual should not know; nor should anyone involved in recording or analyzing the data. This helps ensure that no differences slip in between treatment and control wings due to differences in attitudes, expectations, or hopes of anyone involved in the process.

b. Random assignment of subjects to the treatment or control wings. This is in aid of ensuring that other possible reasons for dependencies and independencies between the cause and effect under test will be distributed identically in the treatment and control wings.

c. Careful choice of a *placebo* to be given to the control, where a placebo is an item indiscernible both for subjects of the experiment and for those administering the experiment from the cause except for being causally 'inert' with respect to the targeted effect. This is supposed to ensure that any 'psychological' effects produced by the recognition that a subject is receiving the treatment will be the same in both wings.

These are in aid of bringing the real RCT as close as possible to an *ideal RCT*. An RCT is ideal for testing "C causes E in A" iff the probability of all combinations of causal factors in A of E are the same in both wings except for C and except for factors that C produces in the course of producing E, whose distribution differs between the two groups only due to the action of C in the treatment wing. Suppose for simplicity $P_A(C)$ in the treatment wing = 1 and $P_A(C)$ in the control wing = 0. An outcome in an RCT is *positive* if $P_A(E)$ in the treatment wing > $P_A(E)$ in the control wing.

As before, designate state descriptions over a full set of causal factors other than C for E in A by K_i. In an ideal RCT each K_i will appear in both wings with the same probability, w_i. Then $P_A(E)$ in treatment wing = $\sum w_i P_A(E/C\&K_i)$ and $P_A(E)$ in control wing = $\sum w_i P_A(E/\neg C\&K_i)$. So a positive outcome occurs only if for some i, $P_A(E/C\&K_i) > P_A(E/\neg C\&K_i)$. Thus *a positive outcome in an ideal RCT for C cause E in A occurs only if C causes E in some A(i) ⊆ A,and hence only if C causes E in A by the probabilistic theory of causality*.

The RCT is neat, at least in the ideal, because it allows us to learn causal conclusions without knowing what the confounding factors are. By definition of an ideal RCT, these are distributed equally in both the treatment and control wing, so that when a difference in probability of the effect between treatment and control wings appears, we can infer that there is an arrangement of confounding factors in which C and E are probabilistically dependent and hence in that arrangement C causes E. It is of course not clear how closely any real RCT approximates the ideal.

Notice that a positive outcome does not preclude that C causes E in some subpopulation of the experimental population and also prevents E in some other. Again, certain anti-depressants are a good example. They have positive RCT results and yet are believed to be helpful for some teenagers and harmful for others.[3]

[3] See for instance the U.S. Food and Drug Administration medication guide at www.fda.gov/cder/drug/antidepressants/SSRIMedicationGuide.htm

4. The destination: a prediction of effectiveness

Out of the morass of vague expressions reviewed in section 2 of what the evidence in EBP is supposed to be evidence for, let's take SIGN's formulation to express the basic idea: "Thus the grading assigned to a recommendation indicates to users the likelihood that, if that recommendation is implemented, the predicted outcome will be achieved." We take it that this is meant to be some kind of causal claim; and also that the users constitute some new population A' different from any experimental population A. Let's suppose then that the target claim that we aim to produce evidence for is *C causes E in A'*. This is a fairly weak claim recall, since it is consistent with it being true that C also causes ¬E in A'. Why should a positive RCT result for *C causes E in A* speak in any way for the truth of *C causes E in A'*?

It will do so if these conditions for RCT relevance are both satisfied:

R1. Populations A and A' have the same causal structure for E.

R2. One of the K_i that picks out a subset of A such that "C causes E in A(i)" holds also picks out a subset of A' that has members.

Of course it will also do so under weaker conditions. The weakest seem to be if both R3. and R4. are satisfied:

R3. C is in C_A → C is in $C_{A'}$.

R4. $P_{A'}(E/C\&K_i) > P_{A'}(E/\neg C\&K_i)$ for some i, where K's are state descriptions over $C_{A'}$---{C}.

That is, C is a cause in A only if it is a cause in A' and there's at least one causally homogeneous subpopulation of A'---picked out by the causal structure *that holds in A'*---where C acts positively and that subpopulation has a non-zero probability.

The lesson to be learned is that although (ideal) RCTs are excellent at securing causal claims about the study population, there is a very great deal more that must be assumed---and defended---if those causal claims are to be exported from the experimental population to some target population. Advice on this front tends to be very poor indeed however. Recall GRADE's recommendation to take as a default the assumption that experimental and target populations are sufficiently similar unless there's good evidence to the contrary. Or consider the US Department of Education website, which teaches that two successful well-conducted RCTs in 'typical' schools or classrooms 'like yours' are 'strong' evidence that a programme will work in your school/classroom (USDE 2003, 10).

This problem often goes under the label 'external validity'. A study has external validity when the claim established in the study population (here A) holds in a target population (A') as well. The Department of Education's advice about external validity is typical: A study will have external validity with respect to a given target if the two populations involved are sufficiently similar. The great advantage of a little rigor is that it can give content to this uselessly vague advice. From the point of view of the probabilistic theory of causality, 'like yours' can mean that R1. and R2. hold. At the very least, R3. and R4. must hold or the RCT results in A will be totally irrelevant to A'. Admittedly, these conditions are abstract so do not give much practical purchase on how to decide whether they obtain or not. But they are not, like the usual advice, without content. At least we know now just what kinds of similarity in what respects we need to look for.

5. Conclusion

We have briefly explained how to make up for the lack of rigor in EBP frameworks when it comes to justifying the relevance of RCT results to effectiveness predictions, that is, when it comes to bridging the gap between “It works somewhere” and “It will work here”. The account sketched here using the probabilistic theory of causality that originates with Pat Suppes has been developed in detail by Nancy together with Jeremy Hardie (2012). The argument taking one from RCT results to effectiveness predictions must be rigorous every step of the way in order for its conclusion to be properly supported by evidence. We urge practitioners and advocates of EBP not to focus solely on rigor in establishing the “It works somewhere” premise at the expense of rigor in establishing other premises that are equally necessary to yield the conclusion that “It will work here”.

References

Balshem, H. et al. (2011). 'GRADE guidelines: 3. Rating the quality of evidence'. *Journal of Clinical Epidemiology*, 64: 401-406.

Cartwright, N. (1979). 'Causal Laws and Effective Strategies'. *Noûs*, 13: 419-437.

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

Guyatt, G. et al. (2011a). 'GRADE guidelines: 1. Introduction---GRADE evidence profiles and summary of findings tables'. *Journal of Clinical Epidemiology*, 64: 383-394.

Guyatt, G. et al. (2011b). 'GRADE guidelines: 8. Rating the quality of evidence---indirectness'. *Journal of Clinical Epidemiology*, 64: 1303-1310.

SIGN. (2011). *SIGN 50: A guideline developer's handbook*. Edinburgh: NHS Scotland.

USDE. (2003). *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington, DC: Coalition for Evidence-Based Policy. http://www2.ed.gov/rschstat/research/pubs/rigorousevid/index.html.