A UNESCO Child and Family Research Centre Working Paper

Evaluation Study Design – A Pluralist Approach to Evidence

Dr Allyn Fives
Dr John Canavan
Professor Pat Dolan

UNESCO Child and Family Research Centre, School of Political Science and Sociology, National University of Ireland, Galway.

Paper presented to the Board of Eurochild, May 20\textsuperscript{th}, 2014, Brussels.

How to cite this paper:
ABSTRACT

There is significant controversy over what counts as evidence in the evaluation of social interventions. It is increasingly common to use methodological criteria to rank evidence types in a hierarchy, with Randomized Controlled Trials (RCTs) at or near the highest level. Because of numerous challenges to a hierarchical approach, the current paper offers a Matrix (or typology) of evaluation evidence, which is justified by the following two lines of argument. First, starting from the principle of methodological aptness, it is argued that different types of research question are best answered by different types of study. The current paper adopts a pluralist interpretation of methodological aptness that stands opposed to the polarized debate between those for whom RCTs are the “gold standard” and those in contrast who contend that RCTs are often inappropriate in social settings. The two opposing paradigms are based on a series of ethical and methodological arguments concerning RCTs that this paper attempts to refute. The second line of argument is that evaluations often require both experimental and non-experimental research in tandem. In a pluralist approach, non-causal evidence is seen as a vital component in order to evaluate interventions in mixed methods studies (as part of evidence-based practice) and also is important for good practice itself (as part of practice-based evidence). The paper concludes by providing a detailed description of what an Evaluation Evidence Matrix can and cannot do.

1. Introduction

Unlike the situation twenty or even ten years ago, policy makers and practitioners now have access to a range of interventions that are underpinned by strong research evidence; evidence adequate to support a decision to implement them. Most commonly identified as “evidence-based practice” (EBP) the existence of these interventions has led to the hope that the policy making process, and specifically decision-making on policy options, will be transformed into one that is justified by empirical evidence. Supporting the development of EBPs, organizations such as the Social Care Institute for Excellence (SCIE) and the Society for Prevention Research (SPR) seek to disseminate information about “what works” in areas such as health, education, welfare, social care, etc. (SPR, 2004; Nutley, Powell, & Davies, 2013). However, the development of evidence-based interventions has elicited a critical response that is epistemologically and methodologically based – concerning our understandings of the nature of knowledge and how it is established; and
what type of evidence is required to underpin policy and practice and how this can be obtained.

One approach common in social and health studies is to categorize evidence types in a hierarchy, with the highest level of confidence in the value of an intervention usually provided by randomized controlled trials (RCTs) (GRADE Working Group, 2004; NICE, 2004; Veerman & Van Yperen, 2007; IES, 2008; Puttick & Ludlow, 2012). However, others believe that experimental studies are in many if not all cases inappropriate when evaluating social interventions and that, for ethical and methodological reasons, non-experimental studies should be considered the most appropriate source of evidence (Morrison, 2001; Hammersley, 2008; Stewart-Brown, Anthony, Wilson, Wintsanley, Stallard, Snooks, & Simkiss, 2011). In the context of this polarized debate on evidence, the current paper arose in response to a request from Eurochild to assist them in developing a position on evidence for evaluation. An Evaluation Evidence Matrix was offered as an orientating guide for those like Eurochild responsible for decision making on evaluation approaches and/or program adoption. The Matrix represents a pluralistic approach to evidence and builds on the efforts of others to dispense with a hierarchy and instead to categorize evidence types in a typology (Muir Gray, 1996; Petticrew & Robert, 2003).

The position put forward in this paper is based on two lines of argument. First, starting from the principle of methodological aptness, it is argued that different types of research question are best answered by different types of study. The current paper adopts a pluralist interpretation of methodological aptness that stands opposed to both positions in the polarized debate on RCTs. This paper will argue that both positions are based on a series of arguments that we will attempt to dispel. The second line of argument is that often both experimental and non-experimental research is required in tandem. This paper illustrates the value to evaluation work of a pluralist approach to evidence, by drawing attention in particular to the questions of “why” and “how” an intervention works as well as highlighting the benefits of practice-based evidence (PBE) to practice itself.
2. Evidence: From Hierarchies to a Matrix

2.1. Evidence Hierarchies

Finding a standardized and manageable approach to the role of evidence in policy making has been seen as a key challenge for the policy and research community. One method is to sort evidence types into a hierarchical classification. In such a hierarchy, typically, the best evidence is provided by an RCT. In an RCT, because treatments are randomly allocated to participants, confounding explanatory variables can be ruled out, and for that reason an RCT evaluation can provide causal evidence for the effectiveness of an intervention.

<table>
<thead>
<tr>
<th>Level of evidence</th>
<th>Type of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1++</td>
<td>High quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias</td>
</tr>
<tr>
<td>1+</td>
<td>Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias</td>
</tr>
<tr>
<td>1-</td>
<td>Meat-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias*</td>
</tr>
<tr>
<td>2++</td>
<td>High-quality systematic reviews of case-control or cohort studies</td>
</tr>
<tr>
<td></td>
<td>High-quality case-control or cohort studies with a very low risk of confounding, bias, or chance and a high probability that the relationship is causal</td>
</tr>
<tr>
<td>2+</td>
<td>Well-conducted case-control of cohort studies with a low risk of confounding, bias, or chance and a moderate probability that the relationship is causal</td>
</tr>
<tr>
<td>2-</td>
<td>Case-control or cohort studies with a high risk of confounding, bias, or chance and a significant risk that the relationship is no causal*</td>
</tr>
<tr>
<td>3</td>
<td>Non-analytic studies (e.g. case reports, case series)</td>
</tr>
<tr>
<td>4</td>
<td>Expert opinion, formal consensus</td>
</tr>
</tbody>
</table>

Figure 1 – NICE Levels of Evidence
Source: Adapted from NICE (2005)

The overall assessment of each study is graded using a code ‘++’, ‘+’ or ‘−’, based on the extent to which the potential biases have been minimized. *Studies with a level of evidence ‘−’ should not be used as a basis for making a recommendation.

According to the National Institute for Clinical Excellence (NICE) (2004) (Figure 1), and also the Grades of Recommendation Assessment, Development and Evaluation (GRADE) system (GRADE Working Group, 2004), the highest form of evidence is provided by an RCT with a low risk of bias, or high quality meta-analytic studies or systematic reviews of RCTs. According to the What Works Clearinghouse guide to evidence (IES, 2008), only RCTs with low levels of attrition provide “strong evidence” (or “meets evidence standards”). Both RCTs with high levels of attrition and quasi-experimental studies (QESs) where there is “equivalence” at baseline can provide only “weaker evidence.” At the bottom of the hierarchy, “insufficient evidence” is provided by a QES where there is no “equivalence” at
baseline. Unlike an RCT, in a QES treatments are not randomly allocated to participants, which has implications for any attempt to rule out confounding explanatory variables.\(^1\)

The quality of study design implementation also matters. The NICE hierarchy makes clear that an RCT study may have a high risk of bias (coded ‘-1’ in Figure 1) and when this is the case the findings should not be used as a basis for making a recommendation. Bias may result in mistakenly attributing causal efficacy to an intervention, and may arise (*inter alia*) when participants move from one arm of the study to another (i.e., they choose the treatment they were not assigned to) or due to study attrition affecting one arm of the study more than the other (i.e., participants leaving the study).

Other evidence hierarchies make more explicit their application to policy and practice. According to the National Endowment for Science Technology and the Arts’ (NESTA) Standards of Evidence for Impact Investing, while an RCT evaluation can demonstrate causality, and independent replications can confirm causality, a higher level of confidence in the impact of an intervention comes with the development of “manuals, systems, and procedures to ensure consistent replication and [a] positive impact” (Puttick & Ludlow, 2012, p. 2). In addition, the evidence hierarchy provided by Veerman and Van Yperen (2007) associates different levels of evidence with different stages in the development of an intervention (see also Urban, Hargreaves, & Trochim, 2014). While non-experimental studies may be appropriate when an intervention is at an early stage in its development, the highest type of evidence is provided by an experiment and it can be sought only for interventions at their most developed. Veerman and Van Yperen (2007) also note that causal evidence can be provided by either an RCT or a repeated case study. The latter is a single subject study design where the participant is exposed to different treatments or different doses of a treatment over time and data are collected at multiple time points.

---

\(^1\) A within-groups quasi-experimental study collects data from one group, usually before and after exposure to the program being evaluated. A between-groups quasi-experimental study collects data from two groups, only one of which received the program, but the treatments are assigned to participants by a non-random process.
2.2. Evidence Matrix

An alternative approach is to categorize evidence in a matrix. The Evaluation Evidence Matrix (see Figure 2) builds on the work of Muir Gray (1996) and Petticrew and Roberts (2003). A matrix draws attention to the wide variety of research questions that can be addressed in evaluation but also to the varied study designs appropriate to answering them. Figure 2 includes research questions posed in exploratory, implementation, and outcomes studies, and also a variety of experimental and non-experimental study designs.

As we saw, experimental studies can be used to analyze the causal effects of an intervention. However, as the Matrix shows, many of the same questions asked about the effectiveness of an intervention can be addressed by a range of study designs. For instance, a cohort study, which is a non-experimental longitudinal study involving the observation of participants over a number of time points, can be used to answer the question “Does X do more good than harm?” Also, a QES may be more appropriate than an RCT in the first stage of an evaluation to answer the question “Does X work?”; and if the QES evaluation shows significant gains by participants then there may be a strong rationale to proceed to a RCT evaluation of efficacy (i.e., under optimal conditions) and then of effectiveness (i.e., under real-world conditions) (SPR, 2004).

In addition, RCTs are not the only study design capable of generating understanding of causality (AEA, 2003). An RCT with a very low risk of bias is thought to provide strong evidence for causality because randomization rules out confounding explanations: “alternative explanations ... are randomly distributed over the experimental conditions” (Shadish, Cook, & Campbell, 2002, p. 105). Because of the lack of randomization in QESs, they are thought to provide indicative rather than causal evidence (Veerman & Van Yperen, 2007). Nonetheless, there are various ways to show that alternative explanations are implausible in a QES, and therefore for a QES to provide evidence of causality, including observations at multiple prettest time points and multiple follow-up time points (Shadish et al., 2002).
### Research Questions

#### Exploratory Studies

| Are needs being met by existing services? | X | X | X | X | X | X | X | X | X | X |
| Are practitioners satisfied with services? | X | X | X | X | X | X | X | X | X | X |
| Are users satisfied with services? | X | X | X | X | X | X | X | X | X | X |
| Is the current service cost-effective? | X | X | X | X | X | X | X | X | X | X |

#### Implementation and Outcomes Studies

| Does X work more than Y? | X | X | X | X | X | X | X | X | X | X |
| Does X do more good than harm? | X | X | X | X | X | X | X | X | X | X |
| Does a proven program work here? | X | X | X | X | X | X | X | X | X | X |
| How does X work? | X | X | X | X | X | X | X | X | X | X |
| Is it meeting its target group? | X | X | X | X | X | X | X | X | X | X |
| Are users satisfied & will they take up? | X | X | X | X | X | X | X | X | X | X |
| Are practitioners satisfied? | X | X | X | X | X | X | X | X | X | X |
| Was it implemented with fidelity? | X | X | X | X | X | X | X | X | X | X |
| Is it cost-effective? | X | X | X | X | X | X | X | X | X | X |
| Is it socially valuable? | X | X | X | X | X | X | X | X | X | X |

---

**Figure 2. An Evaluation Evidence Matrix: A pluralist guide to appropriate study designs.**

*Note: RCT = Randomized Controlled Trial; QES = Quasi Experimental Study*
Other non-causal quantitative study designs are available as well, and they can be matched to various research questions. A theory of change study, a norm referenced study, or a benchmark study may be appropriate to answer many of the questions also addressed by experimental and quasi-experimental studies, as the Matrix shows. In addition, because these study designs do not deny a treatment to participants, they may be more appropriate than an RCT to answer the question “Does a proven program work here?” We return to this issue below in more detail.

Finally, the range of possible research questions is broader than those concerned with effectiveness and outcomes. Some implementation questions (e.g. “Was it implemented with fidelity?”) and exploratory questions (e.g. “Is the current service cost effective?”) are better answered through documentary analysis, qualitative studies, and survey-based approaches. The Matrix also indicates that an RCT can be employed to answer the “How does X work” question. Various approaches to statistical analysis (including structural equation modeling) can be used to test the causal model that underlies the intervention, for example, that the impact of a reading program on children’s reading will be mediated through its impact on children’s perceived self-efficacy as readers and/or frequency of reading (Fives et al., 2015). Therefore, even when an RCT returns findings showing that the intervention caused the observed outcomes, the underlying causal model can and should be tested in addition.

2.3. Arguments for an Evaluation Evidence Matrix

Although we have drawn attention to the wide variety of research questions that can be addressed in evaluation and the varied study designs appropriate to answering those questions, others may still be unwavering in their support of an evidence hierarchy and continue to insist that only a small minority provide “strong evidence.” The current paper adopts two lines of argument in justifying an evidence matrix as an alternative to an evidence hierarchy.

The first is that often both experimental and non-experimental research is required in tandem, and this will be addressed below in section 4.

The second line of argument is that a hierarchy of evidence pays insufficient attention to methodological aptness. If “different types of research questions are best answered by different types of study” then in certain circumstances the “hierarchy may
even be inverted,” for example, if a qualitative study is more appropriate to answer the research question being posed than an RCT (Petticrew & Roberts, 2003, p. 528). Although this paper will support the principle of methodological aptness, it is still a matter of great importance how that principle is to be interpreted. This is the case as the principle can be interpreted in such a way as to limit the plurality of evidence types, by supporting either side in the highly polarized debate on RCTs.

On one side of the debate are those who believe RCTs are the gold standard in evaluation (Axford, Morpeth, Little, & Berry, 2008; Boruch et al., 2009; Oakley et al., 2003; Ritter & Holley, 2008; Shadish et al., 2002). Although they can accept that different study designs are appropriate for answering different research questions, they can go on to argue that, for ethical and methodological reasons, in fact the RCT is the gold standard in evaluation. On the other side is a grouping of those who for different reasons conclude that, in most if not all cases, RCTs are not appropriate in evaluating social interventions (Bonell, Hargreaves, Ciousens, Ross, Hayes, Petticrew, & Kirkwood, 2011; Cheetham, Fuller, McIvor, & Perth, 1992; Ghate, 2001; Greene & Kreuter, 1991; Morrison, 2001; Pawson & Tilley, 1997; Speller, Learmonth, & Harrison, 1997; Stewart-Brown et al., 2011). Although they acknowledge that RCTs can generate evidence of causality in appropriate contexts, they also believe that, unlike the controlled environment of bio-medical studies, very often the fluid and dynamic environment of social interventions is not appropriate for RCTs.

As the principle of methodological aptness can be understood in conflicting ways, it is necessary to address directly this polarized debate on RCTs. In the next section we argue that both positions are based on arguments concerning the ethical and methodological qualities of RCTs that should be rejected. This will not be an exhaustive review of all issues pertinent to decisions on appropriate study designs and evidence types, but rather will focus on a number of contentious and influential arguments. The debate will be illustrated with reflections on the RCT evaluation of the Wizards of Words (WoW) reading program (Fives et al. 2013; Fives et al., 2014).

3. Against the Two Extremes in a Polarized Debate on RCTs

In an RCT, treatments are assigned randomly and treatments are withheld from participants (Altman, 1991; Boruch, Weisburd, Turner, Karpyn, & Littell, 2009; Jaded, 1998).
Participants do not all receive the same treatments and treatments are not assigned on the basis of criteria used in everyday professional practice, in particular the professional’s judgment of the recipient’s needs. The seminal text on clinical trials states that an RCT is justified only if (among other things) (1) it is a “crucial experiment” to determine which intervention is superior and shows “scientific promise” of achieving this result; and (2) there must be a state of “equipoise,” as the members of the relevant expert community must be “equally uncertain about, and equally comfortable with, the known advantages and disadvantages of the treatments” (Beauchamp & Childress, 2009, p. 323, p. 320). However, the opposing arguments in the current debate on RCTs show that there is deep disagreement on whether these requirements can, or should, be met.

3.1. The position that RCTs are the ‘gold standard’ in social evaluation

3.1.1. Argument 1: When making ethical judgments of RCTs therapeutic duties of care are irrelevant

The first argument concerns the fact that in RCTs treatments are withheld. In the field of clinical trials it is argued by some that, if physicians are to act from a duty of care, it is never permissible to withhold what is believed to be a beneficial treatment from their patients. Therefore, it is only ever permissible to recruit subjects in a clinical trial, where treatments will be withheld, when the physician-scientists are confident that no patients will be denied a beneficial treatment (Freedman, 1987; Miller & Weijer, 2006). There must be what is called a state of “equipoise.” This explains why it is ethically objectionable to run an RCT to “confirm” the effectiveness of a program. When a program has already been shown to be effective through an RCT, all things being equal, it is inappropriate to run an RCT again for this program (Beauchamp & Childress, 2009), although it may be necessary to reconsider this question if and insofar as the program is being implemented among significantly different population groups.

However, an opposing line of argument has developed which states that therapeutic duties of care are irrelevant to the ethical justification of RCTs and for that reason equipoise is unnecessary for an ethically justified RCT. Some of their proponents have claimed that RCTs in clinical studies should be evaluated “entirely ... on a scientific orientation ... without
making any appeals to the therapeutic norms governing the patient-physician relationship” (Joffe & Miller, 2008, p. 32). Because the overriding objective in an RCT is scientific rather than therapeutic, the ethical justification of RCTs requires only non-exploitation of participants and that the evaluation is a crucial experiment and shows scientific promise (Miller & Brody, 2007; Veatch, 2007).

If we accept this position, the source of possible moral objections to any one RCT study is reduced considerably. However, even though the RCT design departs from therapeutic criteria in the way that it allocates treatments, and even though study subjects should be made fully aware that participation does not guarantee receipt of any one treatment (to avoid the therapeutic misconception), nonetheless scientists and non-scientist collaborators in an RCT continue to have a duty of care to participants, both qua study subjects and also in many instances qua clients of their services. Because they are in a position to exercise “discretionary” powers, whether as scientists or as professionals, they have a duty to demonstrate “reasonable care, diligence and skill” (Miller & Weijer, 2006, p. 437).

Therefore, there are strong grounds to argue that the duty of care is relevant to the ethical evaluation of RCTs. It also follows that our duty of care may be a source of ethical concern about the permissibility of a RCT evaluation.

3.1.2. Argument 2: Study conditions must be equal at baseline

When RCT findings are reported, often authors will begin by stating that there were no statistically significant differences observed between the two conditions at baseline (i.e., baseline equality) and for that reason the random allocation process was successful and the study will allow causal inferences to be drawn about the effectiveness of the treatment evaluated (Tierney et al., 1995; Oakley et al., 2003; Savage et al., 2003; Ritter & Holley, 2008), and many textbooks and guides to RCTs repeat this contention (Rossi et al., 2004; SPR, 2004; GSR, 2007; IES, 2008). However, random allocation does not, and need not, guarantee baseline equality. Although across all possible RCTs the two groups will be equal at baseline, in any one study, precisely because allocation is random, it may lead to two unequal groups at baseline (Altman, 1985; Senn, 1994; Fives et al. et al., 2013). If there are inequalities between groups at baseline, this can be controlled for in the analysis of outcomes but only if the variables in question predict outcomes (Assmann, Pocock, Enos,
Kasten, 2000). To make such a judgement requires “prior knowledge of the prognostic importance of the variables,” “clinical knowledge,” and “common sense” along with knowledge of the relevant research (Altman, 1985, p. 130 – 132).

The actual role of random allocation in an RCT is to protect against one threat to internal validity, “selection bias.” An RCT is “internally valid” when the differences between the two study conditions observed at the completion of the treatment (i.e. at posttest) can be ascribed to the different treatments (along with random error) and not to other confounding variables (Juni et al., 2001). Random allocation protects against confounding variables because it removes selection bias, i.e. there is “no systematic relationship” between membership of the intervention or the control groups “and the observed and / or unobserved characteristics of the units of study” (GSR, 2007, p. 4). The statement on the Consolidated Standards of Reporting Trials (CONSORT) recommends that the success of the random allocation process depends on both the generation of the random allocation sequence and the steps taken to ensure its concealment (Altman et al., 2001). For example, in the WoW evaluation, as the process of allocation was random and as it was concealed from both the researchers and the program providers, we can be confident random allocation was successful, irrespective of a “small” and non-significant (Cohen’s $d = 0.20$, $p = .10$) difference between study conditions at baseline on the measure of single word reading.

Therefore, it is the case that random allocation of participants can protect against selection bias and ensure that causal inferences can be drawn about the effectiveness of the intervention evaluated. Nonetheless, it is not the case that the random allocation process in an RCT will guarantee baseline equality or that this explains the robustness of findings from a social experiment.

3.2. The position that RCTs often are not appropriate in evaluating social interventions

3.2.1. Argument 2: Study conditions must be equal at baseline (continued)

As we saw, many who believe RCTs are the gold standard hold to the argument that random allocation must ensure baseline equality between study conditions. Many of those opposed to RCTs in social settings share this belief. They assume that, if study conditions are not equal at baseline, this is evidence that the randomization process has failed and the
internal validity of the study has been deeply compromised. They also claim that, because social settings are more fluid than the laboratory or the hospital for instance, it is not possible to guarantee that the two groups of participants will be equal at baseline, as there may be differences between groups on both measured and unmeasured variables, and for this reason RCTs in social settings cannot be scientifically promising studies (Morrison, 2001).

As we have already said, although the role of random allocation is to safeguard against selection bias, this does not require the creation of baseline equality. Therefore, even if the two study conditions are not “the same” in all respects at baseline, this by itself does not undermine the “scientific promise” of the RCT. An RCT study can allow causal inferences to be drawn because random allocation safeguards against selection bias and this is the case even in supposedly more fluid social settings such as children’s education. Therefore, skepticism concerning RCT evaluations of social interventions is not justified by this particular belief.

3.2.2. Argument 3: RCTs in social settings are not ethically permissible as equipoise cannot be attained

The third argument again concerns the fact that in RCTs treatments are withheld. As we have seen, according to some, it is only ever permissible to recruit subjects in a clinical trial, where treatments will be withheld, when the physician-science are confident that no patients will be denied a beneficial treatment. We have already argued that the duty of care is relevant to the ethical judgment of a trial and we have criticized the argument that claims to show such duties are irrelevant. Now we turn to the question of whether it follows that equipoise also is necessary for the ethical permissibility of a trial. It has been claimed (this is the third argument) both that equipoise is necessary for an RCT to be justified, and also that equipoise is difficult or impossible to attain in social settings. Given the variability in professional judgments concerning which treatments are most effective for specific clients, for example, each teacher’s judgments of what each pupil needs to improve their literacy acquisition, equipoise is unlikely to be achieved and impossible to retain for very long (Hammersley, 2008). Therefore, in social settings, in most instances an RCT will be unjustified because it will deny what is thought to be a beneficial treatment to some or all participants in a study.
However, a community of experts can be out of equipoise even when for scientific reasons there are strong grounds for an RCT evaluation. For example, in the WoW study, even though the community of experts had a decided preference for programs similar to WoW, a volunteer reading support for children at risk of reading failure (Brooks, 2002; NESF, 2005), still there was a lack of evidence for its effectiveness. Also, this study helps illustrate that it may be possible to ethically justify an RCT even in the absence of equipoise. This is the case because even when there is a decided community-based preference for it, an intervention may turn out to be ineffective or actually harmful (Oakley et al., 2003). It suggests there may be a moral dilemma to be addressed: a conflict between the duty of care to those over whom we exercise discretionary power (e.g. child participants in an RCT) (Miller & Weijer, 2006) and the duty to go ahead with a crucial experiment that shows scientific promise. And in some cases, as in the WoW study, considerations about the need for causal evidence about the effectiveness of a program that could be harmful or ineffective may be judged important enough to out-weigh opposing considerations about the need to ensure no one is denied a preferred treatment (Fives et al. et al., 2014).

3.2.3. Argument 4: RCTs cannot be scientifically promising in social settings

As we have seen, to be ethically permissible, one requirement is that the proposed RCT shows scientific promise and is a crucial experiment. The final argument to address is that, in most if not all cases, an RCT cannot be scientifically promising in social settings. We look at just two of the numerous methodological criticisms of RCTs here. First, it is argued that in social settings participants often will be aware of the treatment type (control or intervention) they have received, and therefore RCTs cannot be “blind,” and this lack of blindness will undermine the validity of the results (Morrison, 2001; Stewart-Brown et al., 2011). For instance, a positive self-concept may be more likely among some subjects because they know they have received the intervention. Second, in a social setting it is not possible for a trial to be truly controlled, due to the large number of possible confounding variables (Morrison, 2001; Hammersley, 2008). For that reason, it is not possible to infer that the intervention evaluated was the cause of the observed outcomes.

However, these are threats to the internal validity of an RCT that are (along with many others) routinely addressed in clinical trials. And if these threats to validity can be
addressed in clinical trials then there is no *prima facie* reason why they cannot be addressed in social trials. For instance, even though students in the WoW evaluation were aware of their study condition, this was not likely to undermine study validity because external (rather than self-report) measures were used to collect data on outcomes: i.e. standardised tests of reading achievement (Shadish et al., 2002; Webb, Campbell, Schwartz, & Sechrest, 1981). Also, continual monitoring reduced the likelihood of control participants receiving the WoW program (treatment diffusion) or control participants receiving supplementary reading supports (compensatory equalization). In that way, the “controlled” nature of the study was preserved and confounding variables ruled out.

Although there are grounds to reject this argument against RCTs in social settings, it does not follow that RCTs are always preferable on methodological grounds. Even when the aim is to analyse whether an intervention works, it does not follow the RCT is the “gold standard.” First, RCTs require a large sample size to guarantee statistical power, and recruitment of a large sample size is not always feasible for programmatic reasons (Shadish et al., 2002, p. 46). Therefore, in such a scenario, on methodological grounds a quasi-experiment or one of the various non-experimental study designs may be preferable. Second, in highly individualized programs, such as speech and language therapy (and others in areas such as psychotherapy, family support, parenting), where interventions are matched to the needs of the individual, other study designs may be more suitable. In particular, single case study designs, which can provide evidence of causality, may be more appropriate (Vance & Clegg, 2012).

As we have argued, the claim that RCTs are the “gold standard” in evaluation is based on two commonly made but highly questionable arguments: that random allocation should create two groups equal at baseline and that therapeutic duties of care are irrelevant to the ethical justification of RCTs. The opposing argument that RCTs are inappropriate in social settings is based on the following three commonly made but highly questionable arguments: that random allocation should create two groups equal at baseline (but cannot do so in social settings), that RCTs cannot be ethically permissible as equipoise cannot be attained, and that RCTs cannot be scientifically promising. We have concluded that RCTs may be both scientifically promising and ethically justified, but also that ethical objections to RCTs may arise from our duty of care and also alternative study designs may be chosen for methodologically sound reasons.
4. A Plurality of Evidence Types: Combining Practice-Based Evidence and Evidence-Based Practice

Starting from the principle of methodological aptness, we have questioned the two extremes in a polarized debate on RCTs, and argued that ethical and methodological considerations will determine if and in what circumstances an RCT is appropriate. In so doing, we have questioned the tendency to place value in only a small minority of study designs, and provided grounds for greater pluralism in evaluation evidence. We now turn to the second line of argument for an evidence matrix, namely that often both experimental and non-experimental research is required in tandem. We will address what role non-causal evidence types can play in mixed methods studies, and also in practice-based evidence (PBE) and practice wisdom.

4.1. Adaptation of Evidence-Based Practices

There is a close connection between experimental evidence on the one hand and manualized programs on the other. As we have seen, according to the NESTA hierarchy of evidence (Puttick & Ludlow, 2012), while multiple RCTs can confirm that an intervention is effective, manualization provides confidence that the intervention will be effective consistently. While an “efficacy” experimental study tests a hypothesis about the causal impact of a single intervention, implemented in a standard way, in targeting a single disorder or problem (Barkham & Mellor-Clark, 2003, p. 320), an “effectiveness” RCT is to provide evidence regarding the program when implemented by practitioners in real world settings (Green, 2001; Wandersman, 2003). However, in real world settings children and parents may have multiple and complex problems and disorders (Mitchell, 2011, p. 208; Jack, 2011), any one family may be receiving services from multiple sectors, and not all social interventions are discrete, manualized programs (Mitchell, 2011, p. 208). Therefore, while RCTs are ideally suited to analyze the causal impact of standardized programs under ideal conditions, if social interventions and social contexts do not all fit this model, this indicates why other forms of evidence, whether as an adjunct to or in place of experimentally-derived causal evidence, are of value when decisions are made about implementation.
The PBE paradigm examines interventions as they are in routine practice. This can be pursued through “practice research,” research often carried out by (a network of) practitioners themselves so as to inform and improve practice (Barkham & Mellor-Clark, 2003, p. 321) which is also referred to as “empowerment evaluation” (Green, 2001; Wandersman, 2003). Efforts have been made to “integrate” the two paradigms, and in particular to show that evidence-based programs are and must be adapted when they are implemented because “client characteristics, needs and contexts” will not be the same as they were among participants in the original evaluation (Mitchell, 2001, p. 209; Cf Vandenbroeck, Roets, & Roose, 2012).

As an example of an integrative approach, the American Psychological Association (APA) Presidential Taskforce defined EBP as “the integration of the best available research with clinical expertise in the context of patient characteristics, culture, and preferences” (APA, 2006, p. 273). An integrative approach is also proposed by the Institute of Medicine (2001; Sackett, Rosenberg, Gray, Haynes, Richardson, 1996). An integrative approach places considerable importance on the idea of clinical expertise, or practice wisdom (Rubin & Babbie, 2014). Practice wisdom is defined as “practice-based knowledge that has emerged and evolved primarily on the basis of practical experience rather than from empirical research” (Mitchell, 2011, p. 208). However, if an integrative approach is to be pursued, the following question must be addressed: how is practice-based knowledge to be integrated with for instance causal evidence in a way that is rigorous and rationally justifiable?

4.2. Mixed methods studies

The current paper, as part of a pluralist approach, has rejected a hierarchy of evidence that places experimental studies at the apex. While it is the case that some evidence types make a greater contribution to knowledge of efficacy or effectiveness because they better rule out threats to internal validity, it does not follow that causal evidence is by definition superior to other evidence types (APA, 2006, p. 275). Even when an experiment provides evidence of efficacy or effectiveness, “other types of research tools and designs are often needed to yield important information about when, why, and how interventions work, and for whom” (SPR, 2004, p. 1; emphasis added). Some have gone on to argue that because RCTs pay insufficient attention to the circumstances in which an intervention works and for whom it works, RCTs are “not always the best for determining
causality” as they “examine a limited number of isolated factors that are neither limited nor isolated in natural settings” (AEA, 2003; see Bagshaw & Bellomo, 2008). It should be noted that experimental studies can address questions of when, why, how, and for whom, in particular through sub-group analyses. However, the point being made here is that these questions can be addressed as well by various non-causal study designs such as qualitative interviews, observations, and documentary analyses, but also by the various ways evidence may emerge from practice (i.e. practice-based evidence).

Experimental studies may provide causal evidence to support a particular policy, for example, mentoring with at-risk youths to reduce risky behavior, and home visiting in the first 24 months to reduce child maltreatment in vulnerable families. However, even when such evidence has been accrued, other types of evidence become necessary for questions of how and why interventions work with specific groups, questions that need to be answered so as to make informed decisions on the adoption of programs and/or the reform of existing interventions (Pawson & Tilley, 1997; Chen, 2004). Qualitative methods, which may identify how and why interventions work, have been in use in social science research particularly in social work contexts since the 1960s. An important instance is the use of reflective practice in the UK and Australia via “process recording” whereby professionals use discrete methods (Dolan, 2006) to establish with colleagues and those they work with how and in what way they are deemed useful, ineffective, or a hindrance in their service provision. This has a value in breaking down into identifiable components why an intervention works for the end user. For instance, Devaney, in her research with veterans of Family Support, both in practice and academia, established that the person may be as important as the program, in that the quality of the caseworker-client relationship was deemed crucial (Devaney, 2011).
5. An Evaluation Evidence Matrix: A Pluralist Guide to Appropriate Study Designs

If there is a plurality of evidence types how can informed choices be made about what type of evidence is appropriate in evaluation and program adoption? The Evaluation Evidence Matrix (Figure 2) is truly informative in making such choices only if it is understood in light of a number of qualifying statements arising from the foregoing discussion of a pluralist approach to evidence. Below we list both what the matrix can do but also what it cannot do.

5.1. What the Evaluation Evidence Matrix can do

Provide guidance on exemplary questions. We do not claim that the list of research questions in the Matrix is exhaustive or comprehensive. Rather, our aim was to include exemplary questions. The value of the matrix is in providing an orientation to any decision about the appropriate study design / evidence type for specific research questions.

Rule out an evidence type as inappropriate for a specific research question. For instance, the type of evidence that can be produced from a documentary study or a cohort study are appropriate for answering some questions, but are not appropriate for answering others, for example, the question “Does X work more than Y?”

Indicate what evidence type may be appropriate for a specific research question. If we wish to answer the question “Does X work more than Y?” the type of evidence that an RCT study provides may be the most appropriate. However, as we saw, many considerations must be addressed before an RCT can be judged the most appropriate study design.

Show if a study design can be used to answer more than one type of question. While RCTs are associated with the “what works” questions, they can also be used to answer the question “How does X work?” if specific data analysis methods are used. For instance, structural equation modeling can be used to analyze the variables that mediated impact in an RCT study, and in this way help determine how an impact came about.

Show if more than one study design can be used to help answer the same question. No one study design can be said to offer the only appropriate approach to answer any one of the research questions listed, for instance the question “How does X work?” In addition, any one piece of research may combine different study designs or elements from different study designs to answer one question (or a number of questions). Therefore, what are referred to in the Matrix as different study designs in the real world of research may be
combined together in a single evaluation. The Matrix therefore acknowledges the reality of mixed-methods studies.

Help make methodological decisions that are ethically informed. The construction of the matrix has been informed by ethical considerations. For instance, it is assumed unethical to run an RCT to “confirm” that a proven program works, as in such a study some participants would be denied a known benefit. Therefore, although on methodological grounds there is no reason why an RCT could not answer the question “Does a proven program work here?” on ethical grounds it would be inappropriate.

5.2. What the Evaluation Evidence Matrix cannot do

Distinguish between strong and weak evidence types. In the literature on evaluation evidence considerable attention rightly has been given to distinguishing between strong and weak evidence types based on considerations of validity, and also between study designs that have been well implemented and those that have not (NICE, 2004; see sections 3.1.2., 3.2.1., and 3.2.3.). However, these distinctions are not included in the Matrix. For example, when the Matrix identifies the RCT design as appropriate for answering the question “Does X work more than Y?” it has to be assumed that the RCT in question is well implemented and has adequately addressed threats to validity. Therefore, while the Matrix can identify appropriate evidence types for research questions, this must be supplemented with further information on study design implementation.

Show how to combine different study design elements. In a mixed methods study different sources of evidence are combined (see section 4.2.). However, a decision is required on how to combine different types of evidence and in particular whether one type of evidence will be used to support (e.g. expand on or clarify) evidence of another sort (Creswell and Clark, 2007; Wight and Obasi, 2003). The Matrix cannot decide for us which of the two sources of evidence should play a supportive role and once again the Matrix must be supplemented, in this instance with information on mixed methods.

Resolve moral dilemmas. Although ethical considerations have informed the construction of the Matrix (see sections 3.1.1. and 3.2.2.), in some instances ethical reasons can be raised to question the guidance given in the Matrix. For example, an RCT is appropriate to help answer the question “Does X work more than Y?” However, even when there is no evidence from other studies that X is effective, the community of practitioners
may have a decided preference for it (i.e. they are out of equipoise) and therefore may believe it is unethical to run an RCT where some will be denied this program. Just such a clash of principles arose in the WoW study (as we saw in section 2) and those involved had to work through this dilemma and make an ethical judgment. What the Matrix cannot do is make that ethical judgment for us.

**Make a normative judgment of programs and policies.** Although practice and policy should be increasingly evidence-based and evidence-informed, normative (i.e. moral or ideological) judgments are also made about what problems are most pressing or significant, and such judgments are not and cannot be based solely on empirical evidence. This is most clearly the case when a decision needs to be made between two programs but the empirical evidence does not show conclusively that one is better. Again the Matrix cannot decide for us what normative judgments should be made concerning priorities in practice and policy.

6. Conclusion: Towards a New and Long Overdue Perspective

The current paper has proposed a pluralist approach to evidence, represented in the Evaluation Evidence Matrix. Although critical of hierarchies of evidence for their too narrow focus on experimental study design and the “what works” question, the underlying argument is not one of skepticism, and indeed this paper has critically addressed those who contend that RCTs are in most cases inappropriate in social contexts. The pluralist approach to evidence arises from an awareness of a need for new thinking on the relationship between experimental and non-experimental evidence.

First, in many contexts, it is appropriate for both methodological and ethical reasons to conduct an RCT evaluation to generate causal evidence for an intervention’s impact. RCTs are ideally suited to answer the “what works” question, but the same question can be addressed by other study designs (e.g. repeated case studies, QESs). In addition, in an RCT study the causal model underlying an intervention should be tested (for example, through structural equation modeling) to help answer the question “how” or “why” an intervention worked.

Second, in many if not most evaluations, mixed methods research is required. In particular, RCTs should be combined with analyses of the processes by which outcomes were achieved and the contexts in which the intervention was implemented. Such a mixed
methods evaluation can help answer questions about “how” an intervention worked, “for whom,” and “in what contexts?”

Third, it may not always be appropriate to repeat RCTs with the same program when it is implemented in different countries, in particular for the ethical reason that participants should not be denied a beneficial treatment. What may be required is a deeper investigation using other quantitative methods (including time series analysis, longitudinal studies) to identify the key components of successful programs over time.

Fourth, a loosening of the relationship between overly strict following of manuals towards greater respect for practice wisdom as a form of evidence may be worthwhile. The danger of doing so is that it raises problems concerning how to ensure program fidelity. For that reason, limits must be placed on practitioner innovation, and in particular such innovation must not entail a departure from what are known to be the key ingredients of a successful intervention. In this way professional wisdom can be mined as one of a number of sources to establish practice-based evidence, but its function should be seen as complimentary to experimental evidence on “what” works.

References


