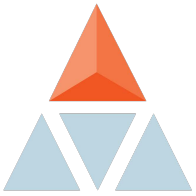

Education
Endowment
Foundation

Implementation and process evaluation (IPE) for interventions in education settings: A synthesis of the literature

Neil Humphrey, Ann Lendrum, Emma Ashworth, Kirsty Frearson, Robert Buck and Kirstin Kerr

MANCHESTER 1824
The University of Manchester

Authors

Neil Humphrey, Ann Lendrum, Emma Ashworth, Kirsty Frearson, Robert Buck and Kirstin Kerr

Manchester Institute of Education
University of Manchester
Oxford Road
Manchester
M13 9PL

Tel: 0161 275 3404
Email: neil.humphrey@manchester.ac.uk

Contents

Executive summary

This research was commissioned by the Education Endowment Foundation (EEF) in order to draw together existing knowledge regarding the aims, functions and methods of implementation and process evaluation (IPE) in relation to educational interventions, with a view to subsequently informing guidance for researchers akin to that already produced on the design and conduct of randomised controlled trials (RCTs) (e.g. Torgerson & Torgerson, 2013). This guidance has subsequently been written by the authors and is available here.

From 'what works' to 'how and why' interventions work in education

RCTs are considered to be the 'gold standard' research method to determine whether a prescribed intervention produces desired effects on a specified set of outcomes. They are thus widely considered to be the optimal means of testing 'what works' (Haynes, Service, Goldacre, & Torgerson, 2012; Maughan, 2013; Torgerson & Torgerson, 2001; Tymms, Merrell, & Coe, 2008).

However, knowing 'what works' in education is insufficient. We also need to know what it means to say that a given intervention 'works', understand how it works, and determine, "why various programs do or do not work, for whom and under what conditions they work, what is needed to scale up certain proven programs, and what policy supports are needed to scale them up without losing their effectiveness" (Slavin, 2012, p.xv).

IPE can help us to answer these important questions by providing researchers with theoretical, methodological and analytical tools that enable insights into the processes and mechanisms underpinning the impact (or lack thereof) of educational interventions. It refers to the generation and analysis of data to examine how an intervention is put into practice, how it operates to achieve its intended outcomes, and the factors that influence these processes.

Implementation is the process by which an intervention is put into practice (Lendrum & Humphrey, 2012). It is a multi-dimensional construct, and may be described in terms of fidelity, dosage, quality, participant responsiveness, reach, programme differentiation, monitoring of control/comparison conditions, and adaptation. A range of factors affecting implementation have been identified in the IPE literatures, including preplanning and foundations, the implementation support system, the implementation environment, implementer factors, and intervention characteristics.

Interventions themselves may be defined as, "purposively implemented change strategies" (Fraser & Gallinsky, 2010, p.459). The 'lifecycle' of intervention development and research generally follows a sequence involving attempts to define and understand the problem to be addressed, designing and describing a proposed solution (including the articulation of an intervention theory), piloting and refinement, establishment of intervention efficacy, demonstration of intervention effectiveness, and finally scale-up and dissemination.

Our primary thesis is that IPE is integral to *all* of these stages of development and evaluation (Moore et al., 2015). As such, our approach to structuring this document reflects this. In terms of our methodology, we undertook a process of *thematic review*, through which we identified the key ideas, models and debates relevant to IPE that could inform the development of subsequent guidance for evaluators. In practical terms, this involved the creation and execution of 12 different work packages comprising of focused literature searches on a given topic (e.g. logic models and theories of change; adaptation; factors affecting implementation).

Summary of findings

IPE in the early stages of intervention development

In the early stages of intervention development, IPE can help developers and/or evaluators to articulate the nature of a given intervention (e.g. its form and function) and build a theory about how it operates to achieve its intended outcomes. In terms of the former, the Template for Intervention Description and Replication (TIDieR - Hoffman et al, 2014) framework represents an attempt to improve and standardise how interventions are described by developers and evaluators. With reference to the latter, a logic model and/or theory of change approach offers great utility.

At the pilot stage of evaluation, IPE may be used to assess the social validity (acceptability, feasibility, and utility) of an intervention. Pilot studies may be used as an opportunity to refine intervention design, materials and theory, but can also be used to optimise evaluation design. The primary focus on process at this stage means that IPE is essential.

IPE in randomised trials

A mixed-methods approach to trial IPE involving both quantitative and qualitative data is preferable to just one or the other. A variety of designs may be utilised; these vary in terms of the level of priority afforded to the different strands (e.g. equal, unbalanced), the amount of interaction between them (e.g. independent, interactive), their timing (e.g. concurrent, sequential, multi-phase), and the protocol for mixing them (e.g. during development, data collection, analysis, or interpretation)

Quantitative methods of data generation in IPE tend to use bespoke (e.g. intervention specific) instruments, which should be subject to the standard considerations in measure development (e.g. establishment of psychometric properties, including inter-rater reliability)**.** There are three primary sources of data for quantitative IPE: researchers (e.g. independent observer), administrative and external support staff (e.g. coaches), and implementers (e.g. teachers). Each has particular strengths and limitations, although direct observation by independent researchers is generally considered to be the preferred approach because of the increased objectivity and rigour afforded by this method**.**

In terms of frequency, aggregated implementation ratings from multiple time points may be more strongly associated with outcomes than a single measurement; however, there are associated increases in evaluation costs and data burden. Quantitative IPE data may be generated for descriptive or relational purposes. Relational analyses typically utilise regression models and can be used to establish whether implementation variability is predictive of outcome variability, empirically validate programme theory, and/or identify critical intervention components.

Qualitative IPE data can help evaluators to determine stakeholder views of intervention implementation, document the implementation of the intervention, highlight contextual influences on implementation and/or validate or modify the intervention theory by illuminating the mechanisms underpinning the change process. Rigorous qualitative IPE is evidenced by an appropriate research design, clear sampling rationale, analysis close to the data, evidence of explicit reflexivity, comprehensiveness of documentation, and negative case analysis. Such evidence enhances the quality, validity and credibility of qualitative IPE.

IPE in randomised trials can serve a number of purposes. First, it can be used to document implementation as comprehensively as possible. Data generated can be used to enhance the internal validity of trial findings (e.g. to demonstrate that the intervention took place, and/or that implementation variability is related to outcome variability)**.** Second, assessment of usual practice through IPE is critical in terms of establishing the counterfactual, exploring the extent to which this changes through the course of a trial,

and establishing programme differentiation. Third, IPE to examine the adaptation process helps us to understand, "when adaptations are being made, why they are being made, what they consist of, and how they align with the program's goals and theory" (Moore et al., 2013, p. 150)**.** Adaptations may be understood in terms of fit (e.g. are changes made for logistical philosophical reasons?), timing (e.g. are changes made pro-actively, or reactively?) and valence (e.g. are changes in keeping with programme theory?). Finally, pre-specified sub-group analyses can help evaluators to understand whether an intervention produces differential benefits for certain participants.

Additional IPE functions in efficacy trials

The first stage of trial-based evaluation examines whether an intervention achieves its intended outcomes under optimal conditions. Put another way, it is the stage at which evaluators ask, "Can it work?" (Haynes, 1999). This is most commonly known as an efficacy trial (Flay et al., 2005; Gottfredson et al., 2015).

In addition to the four core functions noted above (documenting implementation, assessment of usual practice, researching adaptations, and sub-group analysis), the IPE in an efficacy trial may also serve to examine the relationship between 'on treatment' status and intervention outcomes and empirically validate the intervention theory.

On treatment analysis involves comparison of overall intervention effects with those for whom the intervention protocol is known to have been completed to a pre-specified standard. Empirical validation of programme theory may include modelling of mechanisms and processes that are thought to trigger change to demonstrate their association with outcomes, and/or exploring the temporal relationship between outcomes.

Additional IPE functions in effectiveness trials

If an efficacy trial asks, "Can it work?" in relation to an intervention, an effectiveness trial (also referred to as a 'pragmatic' trial; Roland & Torgerson, 1998) poses the question, "Will it work?" (Haynes, 1999). The emphasis here is on delivery of the intervention in 'real world' conditions.

In addition to the four core functions noted above (documenting implementation, assessment of usual practice, researching adaptations, and sub-group analysis), the IPE in an effectiveness trial may serve to enhance our understanding of contextual influences, and distinguish between different intervention components and identify those that are critical (also known as 'active ingredients') through analysis of natural variation or experimental manipulation (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012).

The role of IPE in taking interventions 'from science to service'

At the point in the development and evaluation cycle in which both efficacy and effectiveness have been established, the focus shifts to developing knowledge and understanding of how the now 'proven' programme may successfully be brought to scale. Two related issues are central to IPE at this stage – *scaling up* (bringing the intervention to a wider audience) and *sustainability* (maintaining effective use and impact of the intervention) (Forman, 2015).

At this stage, implementation becomes the outcome of interest. IPE can be used to understand the factors that influence intervention engagement and reach (e.g. who takes on the intervention and why?), implementation quality (e.g. when the intervention is delivered well, what supports this?) and sustainability over time (e.g. what is sustained? How?) (Greenberg, 2010). Given the likelihood for further drift and dilution at the scale-up stage, IPE can also serve an important function in documenting how and why the intervention evolves as it goes to scale (e.g. what changes and why?).

1 Foundations of implementation and process evaluation (IPE)

Randomised controlled trials (RCTs) are considered to be the 'gold standard' research method to determine whether a prescribed intervention produces desired effects on a specified set of outcomes. They are thus widely considered to be the optimal means of testing 'what works' (Haynes, Service, Goldacre, & Torgerson, 2012; Maughan, 2013; Torgerson & Torgerson, 2001; Tymms, Merrell, & Coe, 2008). Recent years have seen significant proliferation of RCTs in education. For example, to date there are more than 140 school-based trials registered on the International Standard Randomised Controlled Trial Number (ISRCTN) registry, of which around two-thirds were registered from 2010 onwards. The Education Endowment Foundation (EEF) alone have funded more than 90 RCTs since their inception in 2011. However, knowing 'what works' in education is insufficient. The term 'what works' alone is viewed by many as problematic, and especially so in relation to complex interventions (Biesta, 2010). Such programmes cannot be easily controlled and often aim to achieve multiple interrelated outcomes, relating both to process and endpoint outcomes for intended beneficiaries that may not lend themselves to easy measurement. Hence, we also need to know what it means to say that a given intervention 'works', understand how it works, and determine, "why various programs do or do not work, for whom and under what conditions they work, what is needed to scale up certain proven programs, and what policy supports are needed to scale them up without losing their effectiveness" (Slavin, 2012, p.xv).

Implementation and process evaluation (IPE) can help us to answer these important questions by providing researchers with theoretical, methodological and analytical tools that enable insights into the processes and mechanisms underpinning the impact (or lack thereof) of educational interventions. However, there is also a much more fundamental reason for an increased focus on IPE in intervention research. Before we can answer 'what works?', we surely first need to know *what was done*, and how this relates to *what was intended.* Research across multiple disciplines, including education, has consistently demonstrated that interventions are rarely, if ever, implemented as designed and that, crucially, variability in implementation is related to variability in the achievement of expected outcomes (Durlak & DuPre, 2008; Lendrum & Humphrey, 2012).

This review was commissioned in order to draw together existing knowledge regarding the aims, functions and methods of IPE in relation to educational interventions, with a view to subsequently informing guidance for researchers akin to that already produced on the design and conduct of RCTs (e.g. Torgerson & Torgerson, 2013). In this sense, the review/synthesis is unique and, we hope, offers a substantive and meaningful contribution to knowledge in relation to the application of implementation science in education. Our primary thesis is that IPE is integral to *all* stages in the design and testing of interventions (e.g. development and piloting, efficacy, effectiveness, scale-up and dissemination) (Moore et al., 2015), and as such our approach to structuring the review reflects this. Thus, following this introductory chapter, each subsequent section deals with a different stage in the development and evaluation cycle[1]. We highlight particular issues and concepts at the point in the cycle at which they are perhaps most salient, but this is not meant to imply that they are not relevant before or after this. So, for example, the issue of social validity is addressed in the chapter on development and piloting (section 2.5), but this concept is also pertinent in later stages. Similarly, our discussion of adaptations is located in the chapter dealing with randomised trials, but with the clear acknowledgement that understanding adaptation is important in IPE from the very inception of an intervention.

In terms of our methodology, we did not undertake a systematic and exhaustive review of IPE and related areas of study. The scale and scope of the literatures that underpin IPE are simply too vast - this would have been an impossible task. Instead, we

[1] We note that this might imply a completely linear process in terms of development and evaluation, and we acknowledge that this is not the case in reality (indeed, this is discussed in section 2.1), but the basic stages of intervention research do at least provide a coherent structure within which to organise the IPE literature.

undertook a process of *thematic review*, through which we identified the key ideas, models and debates relevant to IPE that could inform the development of subsequent guidance for evaluators. In practical terms, this involved the creation and execution of 12 different work packages comprising focused literature searches on a given topic (e.g. logic models and theories of change; adaptation; factors affecting implementation). Three members of the research team (Ashworth, Buck and Frearson) searched electronic databases (e.g. PsycINFO, ERIC, Google Scholar) for relevant outputs that were then summarised and annotated. The resultant repository contained more than 460 sources, which were ultimately used to inform the development of this review. This was supplemented with references drawn from the collective knowledge of the research team, and a small number of additional sources highlighted by peer reviewers of the initial draft of this document (three internationally renowned experts in IPE and related areas of study). The literature presented throughout this document explores findings relating to a wide variety of interventions and spans a number of disciplines, countries and cultures, and samples within them. We note from the outset the consistency evident in this diverse and rapidly expanding evidence base: implementation matters.

In line with the increasing acknowledgement that the privileged status of RCTs (and the experimental method more generally) as the 'gold standard' evaluation methodology can and should be critiqued and problematised, especially as it relates to education (see above, and Stewart-Brown et al, 2011), we also caution against the uncritical acceptance of the theories, methods and analytical techniques that predominate in the IPE literature and are thus documented in this review. In doing so we acknowledge the position adopted by some that answering the 'how' and 'why' questions posed in IPE should only be answered with a particular set of methods and analyses that allow the identification of discrete causal effects through modelling of instrumental variables (e.g. Angrist, Imbens & Rubin, 1996) and/or more elaborate experimental designs that extend the basic RCT model (e.g. multiphase optimisation strategy design (MOST) - Collins et al, 2010). While such work does indeed yield great promise for advancing implementation science, its application to understanding educational interventions has been somewhat limited to date, perhaps as a result of an overly narrow interpretation of what IPE means (see 1.2 below). Thus, we present a view of IPE that represents the state of the field, with the accompanying caveat that much current work presented in these pages should be described in terms of relationships and associations rather than causal effects until the kind of work noted above is more widely and successfully applied to the study of educational interventions.

1.1 Implementation science

Implementation science provides the academic foundation for IPE. It is a multi-disciplinary and inter-disciplinary field with roots in health, education, psychology, business, medicine, sociology and other arenas (Forman, 2015). Implementation science involves the use of scientific methods, "to understand barriers to and facilitators of implementation, to develop new approaches to improving implementation, and to examine relationships between an intervention and its impact. Implementation science has investigated a number of issues, including: influences on the professional behaviour of practitioners; influences on the functioning of… practice organizations; the process of change; strategies for improving implementation, including how organizations can support the implementation efforts of staff members; appropriate adaptation of interventions according to population and setting; identification of approaches to scaling-up effective interventions; implementation measurement methods; and implementation research design" (Forman et al., 2013, p.83)

There are essentially three primary, inter-related goals of implementation science that are supported through IPE:

1. Improving knowledge and understanding of implementation processes and their influence on intervention effectiveness. This includes, for example, understanding how change occurs, and the factors that affect implementation (such as implementer characteristics and organisational conditions).

2. Supporting and improving implementation, for example through the development of research-informed frameworks and guidance with the aim of optimising intervention delivery in diverse contexts.
3. Evaluation of the implementation of interventions. Evaluation may be formative or summative, and will vary according to the stage of intervention development. For example, in pilot studies a formative examination of implementation processes may support refinement and improve feasibility of an intervention. In a summative evaluation, examination of implementation may be integrated with an impact evaluation to explore the relationship between implementation and outcomes.

The main focus of this review is the third goal. However, at appropriate junctures our discussion turns to the first and second goals.

1.2 Defining implementation and process evaluation

The study of implementation evolved as a result of increasing awareness that dissemination of a 'good' innovation was insufficient to guarantee its successful adoption and delivery. From the 1970s, researchers began to observe that interventions were not always delivered exactly as designed (Blakely et al., 1987; Dusenbury, 2003; Rogers, 2003) as implementers were "active modifiers" rather than "passive acceptors" (Rogers, 2003, p.180). Alongside this, evidence began to emerge supporting a general association between the level of implementation and outcomes achieved (Durlak & DuPre, 2008). Thus began the study of implementation, which was initially defined in rather reductive terms that emphasised the need to deliver an intervention with maximum fidelity to the original design (e.g. "the degree to which treatment is delivered as intended" Yeaton & Sechrest, 1981, p.160). While this view is still evident in some aspects of contemporary implementation science (e.g. Elliott & Mihalic, 2004), understanding of implementation generally now recognises it as a multi-dimensional construct (Berkel, Mauricio, Schoenfelder, & Sandler, 2011; Durlak & DuPre, 2008), and accompanying definitions are therefore more wide-ranging.

Implementation is…

"...what a programme consists of when it is delivered in a particular setting" (Durlak & DuPre, 2008, p.329)

"…a specified set of activities designed to put into practice an activity or programme of known dimensions" (Fixsen, Naoom, Blase, Friedman, & Wallace, 2005, p.5)

"…the set of activities that compose the process of putting a practice or programme into place in an organisation" (Forman, 2015, p. 10)

"…put[ting] an innovation into use" (Rogers, 2003, p.20)

"…the process by which an intervention is put into practice" (Lendrum & Humphrey, 2012, p.635)

Box 1. Definitions of implementation.

Other definitions of implementation are more specific and refer to the *transfer of evidence-based* interventions (e.g. Damschroder & Hagedorn, 2011; Rubenstein & Pugh, 2006). However, it is debatable whether this level of specificity is useful, as it risks excluding the study of implementation processes at all stages of the intervention development process, and/or all types of intervention. For the purpose of this review (and accompanying guidance), a broader definition (such as Fixsen et al's (2005) above) that moves beyond the simple 'putting into use' explanation but stops short of specifying that interventions must be evidence-based, is a useful compromise. The *specified set of activities* or *known dimensions* may be interpreted as referring to either a well-defined programme, or the input stage of a clear and explicit theory of change or logic model, thus supporting and encouraging IPE at all stages of intervention development and evaluation.

The term *implementation* is broadly used within literature on education and the evaluation of school-based interventions. However, it is sometimes used interchangeably with *process evaluation*, which has its origins in the field of public health (Steckler & Linnan, 2002). Although there are clear overlaps between the two concepts there are also some subtle variations and the potential for these to inform the evaluation of school-based interventions is worth considering. Theories of process evaluation began to emerge in public health from the mid-1980s, where the primary concern was the avoidance of Type III errors: the erroneous attribution of cause (Basch, Sliepcevich, Gold, Duncan, & Kolbe, 1985; Steckler & Linnan, 2002)[2]. For example, if an evaluation only measures whether an intervention 'works' - that is, it achieves its intended outcomes - any apparent failure to do so may be automatically attributed to programme failure. This assumption may be a Type III error, if, in fact, the cause was the failure to implement the intervention to a sufficient level for change to be triggered and outcomes to be achieved. Basch et al (1985) argued for the evaluation of the implementation of health education programmes alongside impact assessment to avoid Type III errors, and this, "laid the groundwork for modern process evaluation" (Linnan & Steckler, 2002, p.5).

A primary aim of process evaluations is thus to, "distinguish between interventions that are inherently faulty (failure of intervention concept or theory) and those that are badly delivered (implementation failure)" (Oakley, Strange, Bonell, Allen, & Stephenson, 2006, p.413). An equally important aim is to look, "inside the so-called black box to see what happened in the program and how that could affect program impacts or outcomes" (Saunders, Evans, & Joshi, 2005, p.134). This may be achieved through monitoring and assessing the extent to which the different components of an intervention are implemented, and exploring the relationship of these with programme outcomes (Saunders et al., 2005; Steckler & Linnan, 2002) (see sections 4.1 and 5.3). However, the utility of process evaluation for understanding programme outcomes does not imply that it is used only in summative evaluations. An examination of process is equally important for formative reasons, where it may inform understanding of the links between theory and outcomes (e.g. how a programme works) and support refinements to improve feasibility, including requirements for staffing, facilities and resources (Steckler & Linnan, 2002). There is also considerable convergence with the notion of *realist evaluation*, where the focus is on identifying, "what works for whom, in what circumstances and in what respects, and how?" (Pawson & Tilley, 2004, p.2).

Definitions of process evaluation are consistently similar in breadth and scope to those of implementation noted above, but extend them to include an empirical/evaluative element (e.g. Domitrovich, 2009; Rossi et al, 2004; Scheirer, 1994). Collectively then, our working definition of implementation and process evaluation (IPE) for this review is *the generation and analysis of data to examine how an intervention is put into practice, how it operates to achieve its intended outcomes, and the factors that influence these processes*.

1.3 The rationale for IPE

The IPE literatures propose multiple reasons for its examination. These include programme development and refinement, which may comprise the testing of programme theory, the identification of key components and causal mechanisms, and examining moderating factors; informing and improving future dissemination/scale-up of interventions; interpreting outcomes and the avoidance of Type III errors; promoting fidelity; and developing understanding of implementation processes more generally and how these influence outcomes. These may be broadly separated into *formative*, *summative*, and *knowledge generation* arguments.

[2] In the interests of clarity and contextualisation, Type I and II errors refer to false positive and false negative conclusions respectively.

Formative arguments for IPE

An intervention passes through several stages between its inception and broad dissemination (see chapters 2-6), and IPE can provide a critical feedback loop at each phase that enables the programme to be refined and optimised. For example, IPE can be used to examine and test the intervention theory of change, provide feedback on feasibility, clarify causal mechanisms and help to identify critical components, detect moderating factors, and inform future dissemination and on-going quality improvement of interventions.

Summative arguments for IPE

IPE is also extremely useful for summative purposes, where it can help to explain programme effects (or lack thereof). At the most basic level, IPE helps us to understand what actually happened in an intervention. This is necessary to establish the internal validity of the programme and strengthen claims about its role in changing outcomes (Domitrovich & Greenberg, 2000). Furthermore, it is commonly accepted that interventions are not delivered in a uniform fashion that mirrors their design perfectly: "When faced with the realities of human services, implementation outcomes should not be assumed any more than intervention outcomes are assumed" (Fixsen et al, 2005, p.6). IPE helps to clarify the relationship between this variability in delivery and the achievement of intended outcomes, and in doing so aids in the avoidance of Type III errors. Finally, it also enables a deeper understanding of the settings in which, and the individuals for whom, a given intervention is more or less effective.

Knowledge generation arguments for IPE

The final rationale for IPE relates to the need to develop a greater general understanding of how interventions work. IPE findings across multiple individual programmes have collectively generated knowledge that informs the study of implementation and feeds into the intervention design process. Examples include the codification of different dimensions of and factors affecting implementation (Durlak & DuPre, 2008), the distinctions between and influences on different types of adaptations made by implementers (Hansen et al., 2013; Moore, Bumbarger, & Cooper, 2013), and the development of conceptual frameworks for implementation fidelity (Carroll et al., 2007).

1.4 The role of theories, models and frameworks in IPE

Recent years have seen the proliferation of numerous theories, models and frameworks whose broad aim has been to further knowledge and understanding regarding how and why implementation succeeds or fails in a given context. These are summarised in Nilsen's (2015) review. Nilsen (ibid) proposes five categories of theory which are used to achieve three overarching aims: (i) to describe and/or guide the process of translating research into practice (process models); (ii) understanding what influences implementation variability (determinant frameworks, classic theories, and implementation theories); and (iii) informing the evaluation of implementation (evaluation frameworks). The theoretical approaches outlined in this taxonomy (Figure 1 below) may be a very useful tool to inform the aims, design, and data interpretation in IPE at different stages of the intervention development and evaluation cycle.

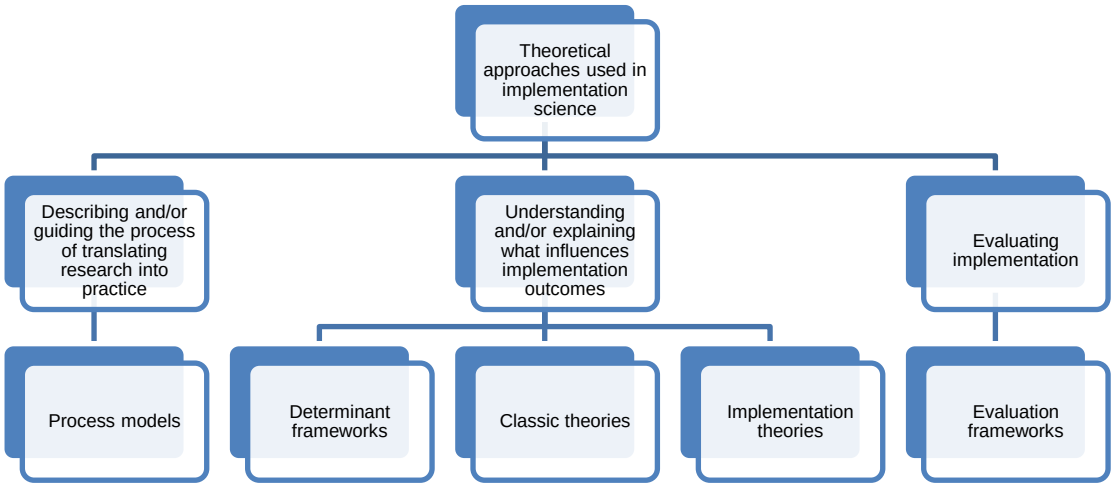

Figure 1. Nilsen's (2015) taxonomy of theoretical approaches in implementation science.

By way of example, consider Domitrovich et al's (2008) Implementation Quality (IQ) model, a 'determinant framework' in Nilsen's (2015) terms. This model positions the IQ of a given intervention as part of an ecological system with influences at different levels that interact with one another. The macro-level includes such factors as community influences, government policies, leadership and human capital. The school-level includes features such as the culture and climate of the school, its characteristics, the resources available and the expertise of staff. Finally, the individual-level includes the attitudes and competencies of individual staff, in addition to their psychological characteristics. In the IQ model, interventions are proposed to vary in terms of their core elements (akin to the 'critical components' noted in section 5.3), standardisation (e.g. the extent to which intervention practices and procedures vary across different sites and contexts) and delivery model (e.g. the mode, frequency, duration and timing of the activities associated with the intervention). The support system underpinning the intervention is framed in a similar manner. Determinant frameworks such as the IQ model may be used to guide the focus of an IPE. Thus, it may be used to inform design (e.g. a nested case study approach), sampling decisions (e.g. ensuring that perspectives are gained from stakeholders at different levels of the ecological system), research questions (e.g. what is the relative influence of factors operating at different levels of the system on implementation variability?) and/or analysis and interpretation of data (e.g. one might use the model in a deductive analysis of data generated in an IPE).

An alternative framing is offered by Forman (2015), who focuses primarily on 'classic' theories (e.g. those originating outside of implementation science), tracing the theoretical roots of implementation science to three psychological paradigms. First, the systems theory perspective attempts to understand implementation through consideration of the functioning of and relationships within social systems, which, "consist of individuals in patterned relationships working to produce a desired outcome" (Forman, 2015, p.23). Schools are viewed as important social systems that form part of a nested social ecology. From the systems theory viewpoint, understanding the boundaries, roles, norms, values, culture and climate of the social system(s) that provide the context for implementation is critical to avoiding implementation failure. Second, social learning theory emphasizes learning and behaviour change as being influenced by factors in the individual's social environment. Hence, the attitudes and behaviours of teachers towards an intervention are theorised to be influenced by those around them (e.g. other teachers). Finally, behaviourism offers a lens through which reinforcement provides a trigger for behaviour change. For example, a behaviourist view can be used to understand the significance of the real or perceived benefits of an intervention among implementers as a key driver of its maintenance and sustainability. Classic theories such

as these can inform IPE by offering different perspectives through which implementation variability and the factors affecting it can be understood.

1.5 Dimensions of implementation

As noted above, there is general consensus that implementation is a multi-dimensional construct. Key reviews and syntheses of the literature (e.g. Domitrovich & Greenberg, 2000; Durlak & DuPre, 2008; Forman, Olin, Hoagwood, & Crowe, 2009) indicate 8 distinct but related dimensions. Before each is addressed in turn, we highlight the variability evident in the IPE literatures regarding how these dimensions are conceptualised. Researchers' understanding and use of the term *fidelity* is particularly noteworthy. Though there is broad agreement that fidelity refers to the extent to which an intervention has been delivered as intended by its developer(s), there are important differences in how it is actually operationalised and where it sits in conceptual frameworks of implementation. The *macro* view of fidelity casts it as a superordinate construct made up of adherence, exposure or dosage, quality of delivery, participant responsiveness, and programme differentiation (Carroll et al., 2007; Nelson, Cordray, Hulleman, Darrow, & Sommer, 2012). By contrast, the *micro* view casts implementation as the superordinate construct, with fidelity, quality, adaptation, and participant responsiveness theorised as subordinate dimensions (Berkel et al., 2011; Durlak & DuPre, 2008). Further confusion eventuates from the manner in which certain terms are used almost interchangeably in some sections of the literature – particularly fidelity and quality (e.g. Steckler & Linnan, 2002). Finally, it is impossible to ignore the influence of intervention characteristics on how different implementation dimensions are defined and assessed. For example, the conceptualisation of fidelity is likely to vary greatly between a highly prescriptive programme and one that is inherently flexible. All of these issues have implications for both the design and interpretation of IPE data.

Fidelity

In the interests of clarity, we adopt the micro view of fidelity in this review. Viewed through the micro lens, fidelity is typically understood and operationalised in *procedural* terms, and can be defined as the extent to which the implementer adheres to the structure and sequence of activities outlined by the intervention developer (e.g. in guidance manuals and/or lesson plans) (O'Donnell, 2008; Odom et al., 2010). For those researchers who take the broader macro view, the 'delivery as intended' concept is taken to incorporate procedural fidelity alongside other aspects such as dosage to give an overall indication of intervention integrity (Dusenbury, Brannigan, Falco, & Hansen, 2003).

Dosage

Dosage is typically conceptualised in terms of the *amount* of an intervention that has been implemented (Moore et al., 2015). For example, where possible, researchers may assess the number of specific 'units' (e.g. lessons/sessions) of an intervention that were actually delivered, often as a proportion of the total number that would be expected/intended within a given period of time (Domitrovich et al., 2008). A distinction may be drawn here between the dosage *delivered* (completeness) and the dosage *received* (exposure) (Naylor et al., 2015). For example, a teacher may deliver most or all lessons in a curriculum-based intervention (high dosage delivered), but children may not necessarily be exposed to them consistently (e.g. because of absence, or if they are withdrawn from the classroom to complete other work).

Quality

While fidelity and dosage provide indications of what and how much was delivered, quality may be conceived of in terms of *how well* intervention components were implemented (Durlak & DuPre, 2008). Here then, the emphasis is on process as opposed to structure (O'Donnell, 2008), with assessment focusing upon implementers' preparedness (including understanding of the basic concepts upon which the

intervention is based), affective engagement, sensitivity, skills and responsiveness in their delivery of intervention content, and where appropriate their integration of concepts into activities beyond the immediate intervention setting (e.g. generalisation techniques) (Domitrovich, 2008; Odom et al., 2010).

Participant responsiveness

Participant responsiveness refers to the extent to which the intended recipients of an intervention are engaged by the activities and content that is delivered during implementation (Dusenbury et al., 2003). Here then, researchers would focus on how well the programme appears to stimulate the interest and hold the attention of participants (Durlak & DuPre, 2008). Consideration may also be given to evidence relating to their levels of participation in core intervention activities and any indications that key objectives are being met.

Reach

Reach is most commonly defined in terms of the extent to which the intended recipients of an intervention actually come into contact with it (Moore et al., 2015). This may be thought of as the proportion of the eligible population that took part in the intervention (for example, in a universal school-based intervention, were all children present during delivery, or were some absent or withdrawn in order to complete other tasks?). Researchers may also consider the extent to which participants are representative of the intended target population (Berkel et al, 2011; Durlak & DuPre, 2008).

Programme differentiation

Interventions do not happen in a vacuum and are rarely completely 'revolutionary' in terms of the ideas and concepts that they introduce. As such, it is important in IPE to consider how distinctive the theory, activities and practices of the programme are when compared to those that may already be in use in the primary intervention setting (e.g. school) (Naylor et al., 2015). The extent to which intervention activities can be distinguished from existing practice provides vital information that can help us to understand what has led to any change in outcomes (or lack thereof) (Dusenbury et al., 2003).

Monitoring of control/comparison conditions

The other dimension of implementation that is concerned with existing/usual practice is a key issue for evaluators. This is particularly pertinent in a trial context, where understanding what is happening in control/comparison conditions helps to establish the counter-factual (Nelson et al., 2012). More broadly, it is recognised that the idea of an 'untreated' control group in a school-based intervention trial is a fantasy (Durlak, 1995). Given the high likelihood that children in these settings will be exposed to a variety of different forms of intervention, monitoring of control/comparison conditions can enable the evaluator to assess the extent to which core intervention components are present in such contexts (Nelson et al, 2012).

Adaptation

Adaptation refers to changes that are made to the intervention during the process of implementation (Durlak & DuPre, 2008; Naylor et al., 2015). For obvious reasons, reference to adaptation is often bound up in discussions of fidelity (see above). However, the inevitability of implementer adaptations and the emerging evidence on their nature and association with outcomes (e.g. Hansen et al., 2013) means that they warrant consideration as a distinct implementation dimension (Durlak & DuPre, 2008). Broadly speaking, adaptations may be classified as *surface* (relatively minor changes to provide a better match to target audience, such as substituting cultural references) or *deep* (more substantial modifications, such as removing core components), and/or *fit* (e.g.

logistical, philosophical), *timing* (e.g. pro-active, reactive) and *valence* (e.g. positive, neutral, negative) (Moore et al., 2013).

Although these different dimensions may be treated separately in a study, they are undoubtedly related (Durlak & DuPre, 2008), and it is highly likely that they will interact with one another. However, little is known about the nature of these interactions and how they may influence outcomes. Berkel and colleagues' (2011) model of programme implementation distinguishes between implementer behaviours (e.g. fidelity, quality, adaptation) and participant behaviours (e.g. participant responsiveness) and makes a number of predictions regarding their inter-relationships and direct and indirect influence on outcomes. For example, the model proposes that the effects of implementer behaviour (e.g., fidelity, quality, and adaptation) on intervention outcomes are mediated or moderated by participant responsiveness. For example, an intervention characterised by high levels of fidelity but with limited participant responsiveness (e.g. low levels of engagement) is unlikely to achieve its intended outcomes. Clearly, empirical verification of such predictions will yield significant advances in our understanding of implementation.

1.6 Factors affecting implementation

With increasing acceptance that implementation of interventions is variable, and robust evidence that this variability influences the achievement of intended outcomes, attention has naturally turned to the factors that influence implementation. The study of these factors in IPE is important in and of itself for knowledge generation purposes, but also has huge potential in terms of helping to optimise future implementation efforts. That is, if we know what factors influence implementation variability, then we can build a model for the optimal conditions for effective implementation that can be communicated to schools as part of the programme dissemination process (Humphrey, 2013). Implementation theories (e.g. Domitrovich, 2008), research reviews (e.g. Durlak & DuPre, 2008) and empirical studies (e.g. Forman et al., 2009) are in broad agreement regarding the factors operating at different levels that are believed to be influential, but their relative importance and the nature of their influence on different dimensions of implementation remains underexplored.

Preplanning and foundations

The preplanning and foundations in place prior to the adoption of an intervention provide the base for effective implementation. Greenberg et al (2005) highlight the importance of exploring the need, readiness and capacity for change in the organisation where the intervention will take place. The authors also note the need to consider the level of commitment and engagement in the change process, the perceived incentive for change, and any available evidence of history of successful change. Here we might also examine the foundations that are in place in terms of existing practice (e.g. is there a history of implementation of similar or related innovations?).

Implementation support system

This aspect includes the variety of methods and practices that are used to support effective implementation for a given intervention. The provision and form of initial (and subsequent) training for implementers and the nature of any on-going support for them (such as coaching) are both likely to influence their knowledge, skills and attitudes in relation to the intervention, and by extension their implementation behaviour (Durlak & DuPre, 2008; Dusenbury et al., 2003). These are therefore important considerations for IPE, particularly given their likely variability at different stages of evaluation (e.g. efficacy vs effectiveness).

Implementation environment

This factor refers to the primary context within which implementation takes place. Data may be recorded on basic compositional and contextual characteristics of schools –

such as size, geographic location, and student population diversity (e.g. proportion of students eligible for free school meals). However, the body of work in this area also points to the primacy of aspects of implementation environment that actively support or impede implementation, such as school leadership support (Gottfredson & Gottfredson, 2002), and the availability of resources – whether these be human, informational, technological, financial and/or physical (Greenberg et al., 2005).

Implementer factors

The primary delivery agent of the intervention will obviously yield a powerful influence on what is implemented and how. In research on this factor, a broad distinction is drawn between implementers' professional characteristics (e.g. education/training, skills and experience), their perceptions and attitudes regarding the intervention (e.g. comfort with and commitment to the intervention), and their psychological characteristics (e.g. stress, burnout, self-efficacy) (Domitrovich, 2008). Put more simply, IPE may therefore seek to examine implementers' *will* and *skill* as predictors of implementation behaviour (Humphrey, 2013).

Intervention characteristics

As we will explore more fully in the next chapter, interventions vary greatly according to a number of characteristics including (but not limited to) form, function, complexity, prescriptiveness, resources and procedures. Each of these is likely to have an influence on implementers' perceptions of the ease with which the intervention can be delivered. For example, complex multi-component interventions requiring action at multiple levels and the commitment of multiple participants take longer to implement successfully (Durlak & DuPre, 2008), are more likely to be discontinued (Yeaton & Sechrest, 1981), and typically become diluted due to their broader scope (Wilson & Lipsey, 2007).

1.7 Chapter summary

- IPE is the generation and analysis of data to examine how an intervention is put into practice, how it operates to achieve its intended outcomes, and the factors that influence these processes.
- If RCTs tell us 'what works', IPE helps us to understand *how*, *why*, *for whom*, and under *what conditions or circumstances* interventions work.
- The academic foundation for IPE is the field of implementation science, whose primary goals are to improve knowledge and understanding of implementation processes, support and improve implementation processes, and evaluate the implementation of interventions.
- The theoretical underpinnings of IPE include process models, determinant frameworks, classic theories, implementation theories, and evaluation frameworks
- Implementation is a multi-dimensional construct, and may be described in terms of fidelity, dosage, quality, participant responsiveness, reach, programme differentiation, monitoring of control/comparison conditions, and adaptation.
- The available evidence suggests that variability in implementation can predict outcome variability.
- A range of factors affecting implementation have been identified in the IPE literatures, including preplanning and foundations, the implementation support system, the implementation environment, implementer factors, and intervention characteristics.

2 IPE in the early stages of intervention development

Interventions may be defined as, “purposively implemented change strategies” (Fraser & Galinsky, 2010, p.459). In the interest of clarity, it is worth briefly breaking this definition down. First, an intervention is *purposive*. It is intentional, not accidental. Second, an intervention is *implemented*. It therefore represents a set of activities, processes and actions - things that are done that can be observed or inferred. Third, interventions are about *change*. One level of change refers to the differences in the activities, processes and actions that characterise the intervention as compared to what was done before it was introduced. The second level of change refers to the intended outcomes that are the ultimate product of the intervention. Fourth, interventions are *strategic*. That is, their constituent components form a coherent, organised plan to bring about the change noted above.

2.1 The development of interventions

Much has been written about the processes that characterise the ‘lifecycle’ of interventions. Here, development and evaluation activities are interwoven over time. Models that attempt to capture and provide sequence to these processes include:

- Rothman and Thomas' (1994) model for the design and development of interventions in human services
- Fraser and Galinsky's (2010) 5-step model of intervention research in the development of prevention programmes for children
- Craig et al.'s (2008) development-evaluation-implementation process guidance for complex interventions
- Campbell et al's (2000) framework for the design and evaluation of complex interventions
- Rohrbach's (2014) overview of the design process for preventive interventions

One immediate limitation to note in this body of literature is an implicit assumption of linear sequencing in the development of interventions (that is, *x* is done first, then *y*, then *z*), when in fact the real world dictates a more complex picture, in which the temporal relationship between different development and evaluation activities is cyclical, interactive and non-linear (Bopp, Saunders, & Lattimore, 2013). For example, even the basic starting point of ‘defining and understanding the problem’ may need to be revisited at different points in the development process as the nature of the problem or the understanding of it changes over time. Similarly, some stages may occur concurrently rather than sequentially (e.g. the development of intervention materials and intervention theory often interweave), and others may overlap (e.g. piloting and the establishment of efficacy may intersect in a pilot efficacy trial). Finally, a linear model presents an oversimplification as it implies that a given stage of development and evaluation only becomes pertinent once previous stages have been completed. For example, we might infer from a linear model that scaling of an intervention only becomes a consideration once efficacy and effectiveness are established, yet there is a strong case for saying that it needs to be given thought much earlier in the process (e.g. the idea of ‘designing for scale’). Indeed, this issue alone may shed light on why so-called ‘evidence-based’ interventions are rarely implemented at the scale to which their developers aspire (Walker, 2004).

2.2 Defining and understanding the problem

Synthesising these key models and drawing upon other relevant sources (e.g., Connell & Klem, 2000; Connell & Kibusch, 1998), we propose that the starting point for the development of any intervention is some kind of assessment of the current situation from which change is sought. This includes some theorisation and analysis of how the current situation came to be. Thus, there is an attempt to *define and understand the problem*. For a useful example, we need look no further than the reason that the Education Endowment Foundation was created. Research evidence shows that children from socioeconomically disadvantaged backgrounds begin primary school behind their

more affluent peers and that this attainment gap widens as they make their way through the phases of compulsory education (Raffo et al., 2010). The point here is that intervention developers need to clearly delineate the problem and provide robust evidence of its scale and significance. For example, 28% of children in the UK were living in poverty in 2013-14, and we know that this has a powerful negative influence on later life chances that is at least partly mediated by its impact on academic attainment (Child Poverty Action Group, 2015). Accompanying theories provide a range of explanations for the poverty-attainment relationship, focusing on individual 'micro-level' factors; immediate family, community, school and peer group 'meso-level' social contexts; and/or social structures and notions of power and inequality at the 'macro-level' (Raffo et al., 2009).

Theory also provides the bridge to ideas about the possible ways through which change may be sought. In doing so, there is a need for clarity regarding the focus of the intervention, and what may be realistically achieved. For example, as noted above, poverty is a principal influence on children's attainment. Few – if any - interventions attempt to resolve poverty, but may instead attempt to intervene to modify or improve factors that mediate the link between poverty and attainment (such as parental engagement). Theorisation may take different forms that vary as a function of the nature of a given problem and the complexity of its underpinnings, in addition to the orientation and disposition of the developer(s). Intervention development in prevention science, for example, often uses theory relating to risk and resilience processes in human development (e.g. Wright, Masten, & Narayan, 2013), ecological systems theory (e.g. Bronfenbrenner, 2005) or a synthesis of these ideas (e.g. Powers, 2010) as its starting point.

2.3 Designing and describing a proposed solution

The next logical step in the development process is design. Here we see the generation of ideas regarding the nature of the intervention, which typically occurs in tandem with the articulation of its underpinning theory (see section 2.4 below). This should build upon existing relevant literatures pertaining to both the nature of the substantive problem to be addressed and the science of effective intervention design. When available, recent systematic reviews and/or meta-analyses are very useful at this stage as they can provide a helpful summary of the current knowledge base(s) that can serve as a foundation for the intervention (Fraser, Richman, Galinsky, & Day, 2009; Craig et al, 2008). For example, in the field of social and emotional learning, a recent meta-analysis of over 200 studies by Durlak, Weissberg, Dymnicki, Taylor, and Schellinger (2011) identified several core intervention design features that were associated with improved outcomes. The authors found that 'SAFE' interventions – those that use a **s**equenced step-by-step training approach, **a**ctive forms of learning, **f**ocus sufficient time on skill development, and have **e**xplicit learning goals - produced larger effect sizes for a range of outcomes than those that did not make use of these practices. An alternative example can be seen in Hattie's (2009) comprehensive study of the factors that influence children's academic attainment. One can see how the findings of this meta-analysis might inform the design of future interventions (indeed, this is the case - Hattie, 2012). This process, in which basic science is applied to inform intervention development, is known as 'Type 1 translational research' (Woolf, 2008)[3].

However, it is important to be clear that not all interventions are designed in this way. Many are initially developed simply through a practitioner trying something different. Consider the case of the Good Behaviour Game (GBG - described in more depth below), which began life as the innovative attempts of a new fourth grade teacher to address recalcitrant discipline problems in her classroom (Barrish, Saunders, & Wolf, 1969), and went on to become one of the most well-known and widely validated behaviour management systems in education. Nonetheless, rigorous information gathering and

[3]Type 2 translational research – which explores how effective interventions can be scaled-up and sustained – is addressed later in this review (see section 4.4).

synthesis in the early stages of design are strongly recommended[4], not least to avoid merely replicating the efforts of others (Fraser et al., 2009). Success cases and failure case analysis are of equal importance here. That is, the developer(s) will benefit from knowing not just what appears to work in a given area/context (and why), but also what has been shown to not work (and why).

New interventions also need to be clearly articulated. Taxonomies and frameworks in the published literature (e.g. Forman, 2015; Foxcroft, 2014; Humphrey, 2013; Moore et al., 2015) indicate that they can be characterised by a number of basic features, including:

- Form (e.g. universal, selective, indicated)
- Function (e.g. environmental, developmental, informational)
- Level and location (e.g. individual, group, family, school, community, societal)
- Complexity and structure (e.g. single component, multi-component)
- Prescriptiveness and specificity (e.g. manualised, flexible)
- Components (e.g. curriculum, environment/ethos, parents/wider community)
- Intervention agents (e.g. teachers, external staff)
- Recipients (e.g. teachers, pupils)
- Procedures and materials (e.g. what is done, how often)

By way of illustration, consider the following description of the *Achievement for All* (AfA) pilot programme (Department for Children Schools and Families, 2009a):

> AfA was developed by the now Department for Education in response to concerns about the school experiences and outcomes of children and young people with special educational needs and disabilities (e.g. Lamb, 2009). The aim of the programme was to improve the academic attainment, parental engagement and confidence, and wider outcomes (e.g. behaviour, attendance, positive relationships, bullying, wider participation) of these students. In terms of form, it combined universal (for example, whole school social and emotional learning programmes) with selective and/or indicated activities (e.g. social skills interventions for children with autism). The three core components of the intervention were: (i) assessment, tracking and intervention; (ii) structured conversations with parents; and (iii) developing provision for wider outcomes. AfA operated at a number of different levels, including individual, family and school. It was a complex, multi-component intervention that contained a blend of prescriptive (e.g. structured conversations with parents were manualised) and flexible (e.g. schools were able to prioritise two of the five wider outcomes noted above and use project funding to enhance their provision in these areas) forms of activity. The intervention agents were designated 'Key Teachers' (defined as the member of teaching staff who knew an individual student best), support staff, the school special educational needs co-ordinator, and the school head-teacher, with additional support from external professionals (e.g. Local Authority advisory teachers). The recipients of AfA were those pupils on the special educational needs register in a given school. Finally, each strand of the programme included a bespoke set of procedures. For example, the assessment tracking and intervention strand included the use of Assessing Pupil Progress and Progression Guidance (Department for Children Schools and Families, 2009b) to track students' progress, setting of curriculum targets for pupils with SEND and implementation of appropriate interventions in support of this (e.g. 1:1 tuition, use of specialist resources), sharing of academic targets and data with parents (e.g. emails, text messages, meetings), and adaptation of interventions based upon frequent target/progress reviews. Information management systems (e.g. 'Assessment Manager') were used to support these procedures, and data generated was aggregated to allow assessment and tracking at class and school level (Humphrey & Squires, 2011).

Box 2. Characteristics of the AfA programme.

Inadequate description of interventions is a perennial problem in a range of disciplines. For example, in a recent review by Hoffmann, Erueti, and Glasziou (2013), less than 40% of non-pharmacological interventions were found to be described adequately in papers, appendices and/or associated websites. Without a complete description of an intervention: (i) the person/s responsible for delivery cannot reliably implement it[5]; (ii) recipients do not know what they are 'signing up for'; (iii) researchers cannot properly replicate or build upon existing findings; and (iv) it is difficult, if not impossible, to

[4] Indeed, future refinement of the GBG was informed by the evidence base on contingency management.

[5] We note, following Kerr, Dyson and Raffo (2014), that the possibility of whole or partial replication is not necessarily an explicit concern for certain kinds of interventions, particularly if these are locally developed. However, this does not detract from the need for comprehensive description.

understand how and why it works. In the reporting of randomised trials, insufficient intervention description represents a serious threat to validity. As a means to address this issue, Hoffmann et al. (2014) developed the Template for Intervention Description and Replication (TIDieR). TIDieR offers a simple, standardised framework that can improve the quality of how interventions are described and understood. Here we present a version adapted to suit school-based interventions:

1. Brief name
2. Why: Rationale, theory and/or goal of essential elements of the intervention
3. Who: Recipients of the intervention
4. What: Physical or informational materials used in the intervention
5. What: Procedures, activities and/or processes used in the intervention
6. Who: Intervention providers/implementers
7. How: Mode of delivery
8. Where: Location of the intervention
9. When and how much: Duration and dosage of the intervention
10. Tailoring: Adaptation of the intervention
11. How well (planned): Strategies to maximise effective implementation
12. How well (actual): Evidence of implementation variability

It is important to note that item 12 would typically only be included once an intervention had actually been piloted or trialled. With this exception, the other 11 items represent, in our view, the fundamental information required to be able to articulate a school-based intervention at any stage in the development process. By way of illustration, consider the following application of TIDieR to the Good Behaviour Game, currently being trialled by the EEF:

(1) Brief name: The Good Behaviour Game (GBG)
(2) Why (rationale/theory): The GBG draws upon the principles of contingency management, in that children receive reinforcement when they engage in appropriate behaviours. However, the group-based orientation of the intervention means it also uses social learning theory, because pupils at-risk of developing conduct problems are able to learn from the appropriate behaviour being modelled by other team members. Finally, the GBG is informed by social field/life course theory, which posits that successful adaptation at different life stages is contingent upon an individual's ability to meet particular social task demands. In school, these task demands include being able to pay attention, work well with others, and obey rules. Success in social adaptation is rated both formally and informally by other members of the social field (e.g. teachers, peers). Social field theory predicts that improving the way in which teachers socialise children will improve their social adaptation. It is also predicted that early improvements in social adaptation will lead to better adaptation to other social fields later in life (Sheppard G Kellam et al., 2011).
(3) Who (recipients): The GBG is a universal intervention and is therefore delivered to all children in a given class.
(4) What (materials): Participating schools receive GBG manuals that detail the programme theory, goals and procedures. Other materials include some tangible rewards (e.g. stickers), displays (e.g. scoreboard), and data forms for recording and monitoring purposes.
(5) What (procedures): The GBG is described by Tingstrom et al (2006) as an, "interdependent group-oriented contingency management procedure" (p. 225). Pupils in a given class are divided into mixed teams with up to 7 members. Strata can include a range of factors such as behaviour, academic ability, and gender. The teams then attempt to win the game as a means to access particular privileges/rewards. During the game period, the class teacher records the number of infractions to the following four rules among the teams: (1) We will work quietly, (2) We will be polite to others, (3) We will get out of seats with permission, and (4) We will follow directions. The game is 'won' by the team with the lowest number of infractions when it ends, although any team with fewer than four infractions also accesses the agreed reward (Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012; Kellam et al., 2011). Over the course of implementation of the GBG, there is a natural evolution in terms of the types of rewards used (from tangible rewards such as stickers to more abstract rewards such as free time), how long the game is played for (from 10 minutes to a whole lesson), at what frequency (from three times a week to every day), and when rewards are given (at the end of the game, end of the day, and at end of the week) (Elswick & Casey, 2011; Tingstrom et al., 2006). Good behaviour achieved during the relatively brief 'game' periods is increasingly generalised to other activities and parts of the school day. Thus, the intervention leads to behaviour modification and intrinsic reinforcement so that modified behaviour is retained even after external reinforcement is removed (maintenance) and will be exhibited in all settings (generalization). These processes are documented by 'game' and 'probe' data collected by teachers during implementation (Chan et al, 2012).
(6) Who (provider): The GBG is implemented by class teachers.
(7) How: The GBG is implemented face-to-face during the normal school day. As it is a behavior management strategy rather than a taught curriculum, it requires no space in the class timetable.
(8) Where: The GBG is implemented on-site in participating schools.
(9) When and how much: The GBG is played throughout the school year. As noted above, dosage evolves throughout the period of implementation in terms of both the duration of the game (from 10 minutes to a whole lesson), and the frequency at which it is played (from three times a week to every day).
(10) Tailoring: The GBG is a manualised intervention and participating teachers receive technical support and assistance from coaches in order to optimize fidelity of implementation. However, it is now accepted that some form of adaptation is inevitable and indeed may be desirable in order to improve local ownership and fit to context (Durlak & DuPre, 2008; US Department of Health and Human Services, 2002). An important aspect of the GBG coach role is to support teachers to make adaptations that are in keeping with the goals and theory of the intervention (Moore et al., 2013).
(11) How well (planned): Teachers receive three days of training (2 days initial; 1 day follow-up). On-going technical support and assistance is provided by trained coaches. This comprises modelling, observation and feedback, ad-hoc email and telephone support, and provision of additional/booster training or information sessions as required.

Box 3. Application of the TIDieR framework to the Good Behaviour Game intervention.

In spite of its obvious advantages in terms of providing a standardised format for description and replication, TIDieR may be seen as somewhat reductive and thus problematic in cases where interventions are more complex and/or flexible in nature. Nonetheless, it is at the very least a useful 'primer' for developers and evaluators alike to consider when they attempt to describe the characteristics of a given intervention.

2.4 Articulating an intervention theory

Decisions made regarding intervention design features reflect a set of ideas about how and why change is expected to occur. The importance of articulating this intervention theory cannot be overstated. As Stinchcomb (2001) notes, "seasoned travellers would not set out on a cross country motor trip without having a destination in mind, at least some idea of how to get there, and, preferably, a detailed map to provide direction and

guide progress along the way" (p.48). In intervention development and research, this typically takes the form of a logic model and/or a theory of change (Forman, 2015).

A *logic model*, often also referred to as a 'chain of reasoning' (Mclaughlin & Jordan, 1999), "describes the sequence of events for bringing about change by synthesising the main program elements into a picture of how the program is supposed to work" (Centers for Disease Control and Prevention, 1999, p.9). Logic models enable developers to clarify the strategy of an intervention, articulate assumptions about conditions for programme effectiveness, provide a frame of reference for evaluation, and can strengthen claims about causality (Centers for Disease Control and Prevention, 1999). Although there is some variability in how they are presented, it is common for logic models to describe what goes in (inputs), what happens to trigger change (processes and mechanisms), and what changes as a result (outcomes) (Forman, 2015). Below we present an exemplar logic model for the aforementioned AfA pilot programme (Figure 2, taken from Barlow, Humphrey, Lendrum, Wigelsworth, & Squires, 2015). Of note here is the fact that the model also incorporates *factors affecting* inputs and processes in acknowledgement of the evidence base for these in IPE (see section 1.6). What is not included in this example is a sense of when particular outcomes are expected to change (e.g. short or long term) (Forman, 2015) and their magnitude and/or temporal relations (e.g. direct, proximal vs. indirect, distal) (Fraser & Galinsky, 2010). Thus, like interventions, logic modelling can be simple or complex, and can be performed using a variety of methods. For example, Yampolskaya, Nesman, Hernandez, and Koch (2004) describe the development of a logic model for a mental health agency in the United States that drew upon a concept mapping approach.

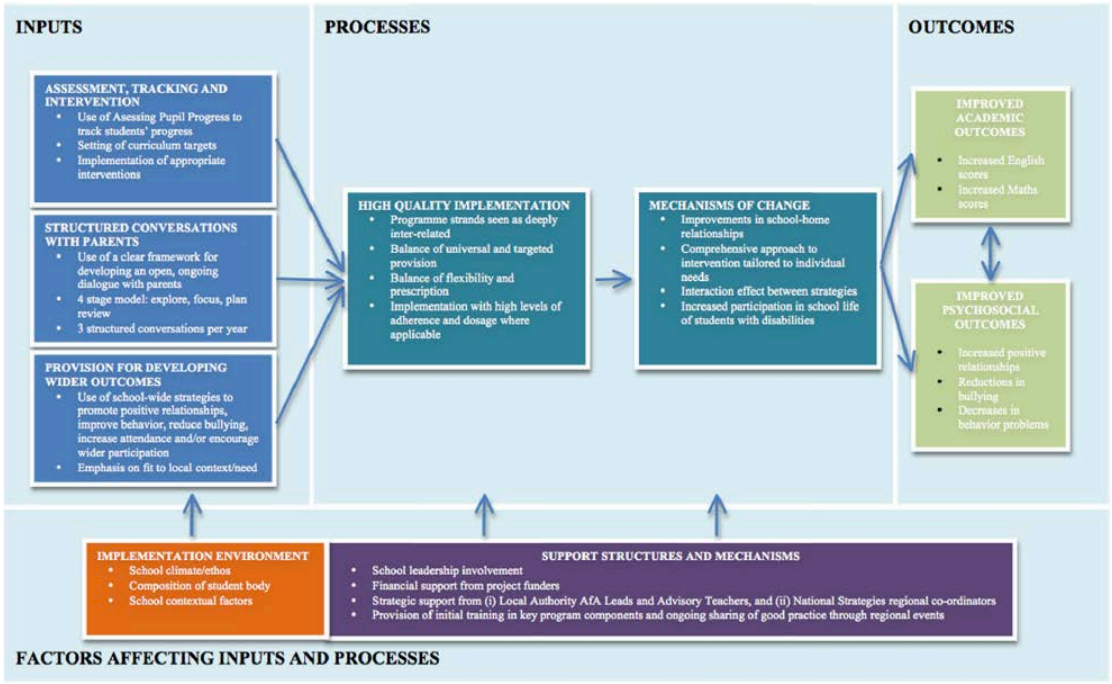

Figure 2. Example of an intervention logic model (taken from Barlow, Humphrey, Lendrum, Wigelsworth, & Squires, 2015).

There is inconsistency in the IPE literature regarding the extent to which a logic model, described above, is distinct from a *theory of change* (ToC). As with other terms discussed in this review, they are sometimes used interchangeably. However, in some contexts (most notably, the evaluation of comprehensive community initiatives, e.g. Connell & Kibusch, 1998) reference to a ToC approach signifies a particular methodology that aims to address a chief limitation of logic modelling – simplicity. As Rogers (2008) states, "Life is not simple, but many of the logic models used in programme theory evaluation are" (p.29). For complicated and/or complex interventions, a logic model approach may be too reductive. Blamey and Mackenzie (2007) and

Rogers (2008) have argued that 'real world' theories of change should reflect the fact that interventions are not unified entities, people are not simply passive recipients upon which change is enacted, and that context (individual, social, organisational) is critical to understanding how and why intended outcomes are achieved. Thus, if a distinction between logic models and theories of change is to be made, it is that the latter actively embraces complexity rather than aiming for simplicity. Blamey and Mackenzie (2007) outline a 6-step model for eliciting a ToC in which evaluators and intervention stakeholders work collaboratively. Throughout the process, all parties are encouraged to highlight conflicting and discrepant ideas and to reflect on the contextual factors that influence their decisions. At the first stage, the focus is on the long-term vision for the intervention, linked to a problem at the local or national level (for example, eradicating the aforementioned achievement gap). Importantly, the long-term vision is likely to extend beyond the immediate lifespan of the intervention itself. With the ultimate aim agreed, in the second stage stakeholders explore the necessary outcomes that would need to be achieved by the end of the intervention if the longer-term vision is to be realised. In the third and fourth stages, interim outcomes that would facilitate the achievement of these targets are considered. In the fifth stage, the activities and strategies that are the most appropriate to bring about the required change are generated by stakeholders. Consideration of context is embedded here. Finally, in the sixth stage, stakeholders consider the resources that are required to enact the planned intervention.

In a ToC paradigm, quality is judged by four criteria – the extent to which the theory developed through the above process is *plausible* (is there a consensus among stakeholders that the logic and assumptions of the theory are correct?), *doable* (are there necessary resources (e.g. human, material, financial) to implement the strategies outlined in the theory?), *testable* (are there credible methods to evidence the veracity of the theory?) and *meaningful* (are the outcomes important to stakeholders and worth the effort involved to achieve them?) (Connell & Klem, 2000). ToC generation often results in a detailed intervention map against which the reality of implementation can be contrasted during an evaluation. This enables each aspect of the theory to be interrogated with a high degree of specificity and amendments to be made as necessary. Put another way, ideas and assumptions that made sense 'on paper' but were not borne out in the practice of the intervention can be uncovered and changed.

2.5 Piloting and refinement

The piloting stage involves the first 'real world' testing of a given intervention and is typically conducted on a small scale. On the basis of pilot findings, refinements may be made in an attempt to 'optimise' it ahead of the next stage of evaluation (e.g. an efficacy trial). Refinements are not solely confined to intervention design and materials here. Pilot studies may also provide vital methodological information that can inform subsequent evaluation design (e.g. estimation of likely recruitment and retention rates, intervention effect sizes, sensitivity of proposed outcome measures; Craig et al., 2008). The data generated can also be used to refine intervention theory, particularly if the processes are operating in a different way than originally envisaged (see above). Finally, pilot study data may also provide useful insights relating to required contextual and/or compositional characteristics and foundations for change, which may lead to modification of the perceived 'optimal conditions' for the intervention. This in turn can inform the design of a subsequent efficacy trial.

One view of the piloting stage is that it is best used to establish the *social validity* of the intervention. Social validity refers to the value and social importance attributed to a given innovation by those who are direct or indirect consumers of it (Hurley, 2012; Luiselli & Reed, 2011). In the EEF context, these are likely to be teachers and other school staff, parents, pupils, other education professionals, and policy-makers. Adapting Wolf's (1978) classic taxonomy, social validity can be thought of in terms of *acceptability* (e.g. are the intended outcomes of the intervention wanted, needed, and/or socially significant?), *feasibility* (e.g. are the procedures employed to achieve the intended outcomes of the intervention considered to be acceptable?) and *utility* (e.g. are the

outcomes of the intervention considered to satisfactory[6], and worth the effort required to achieve them?). Thus, intervention piloting can seek answers to these critical questions. It is, however, important to be aware that the importance of social validity does not begin and end with a pilot study – it should continue to be central throughout the intervention development and evaluation cycle. It is also important to acknowledge that some aspects of social validity are, by definition, confounded in a pilot study that involves a school or other institution that has actively volunteered to participate. In such schools there is presumably some latent perception of need (and benefit) relating to the problem that the proposed intervention is designed to address. Thus, focused pilot evaluation may also be supplemented with broader, scoping research at this stage (e.g. a larger scale 'market research' exercise in which views about a proposed innovation are sought).

Fraser et al (2009) argue for an emphasis on *process* as opposed to *outcomes* at the pilot stage. The predominance of process in the pilot stage therefore means that IPE is central. Key implementation-related questions may include: *can implementers deliver the intervention in the time allotted?*; *does any sequencing of content and other aspects of intervention design make sense to implementers and recipients?*; *are suggested activities congruent with the context of delivery (e.g. target population, setting)?*; *are recipients engaged?* (Feeley et al., 2009; Fraser & Galinsky, 2010). However, this does not mean that useful outcome data cannot be generated in parallel. Indeed, in health-related fields the tradition of conducting an 'exploratory trial' – in which both process and outcome data can be generated – is well established, and features as a distinct stage in Campbell et al's (2000) framework for the design and evaluation of complex interventions. Although this may seem inherently problematic (e.g. why would one conduct a randomised trial that is highly likely to be statistically underpowered and lacking other features that enhance the security of its findings?), exploratory trials can provide critical, albeit tentative, evidence about outcomes (e.g. *Are there indications of impact on intended outcomes? Of what kind of magnitude? For whom?*) that can help to inform decisions about whether the significant undertaking of a definitive efficacy trial is warranted.

Thus, the pilot stage of intervention development can take various forms, use a range of methodologies and serve multiple purposes. By way of illustration, consider the studies outlined in Box 4 below:

[6] It is important to note that the pilot stage of evaluation also provides an opportunity to identify intervention outcomes that were previously unanticipated.

Kendal, Callery, and Keeley (2011) report on the piloting of 'The Change Project', a self-referral, guided self-help intervention to promote emotional wellbeing in adolescents. Their qualitative study, set in three urban secondary schools and drawing upon interview and focus group data from school staff and students, was designed primarily to explore process and develop theory to inform a future trial. The authors' primary focus was on the social validity of the intervention, and in particular its acceptability and feasibility. Perceived barriers to and facilitators of uptake and implementation of the intervention as planned are also explored. Kendall, Callery and Keeley (2011) draw a number of conclusions from the data generated that have direct implications for the refinement of the intervention ahead of an efficacy trial (e.g. providing a greater focus on time management when training staff as a means of improving the acceptability of the intervention).

Bluth et al (2015) conducted a randomised pilot study of a school-based mindfulness programme known as 'Learning to BREATHE'. In an active control group design, 27 'at-risk' adolescents were randomly assigned to receive the intervention or attend a substance abuse course. In addition to reporting on the preliminary impact of Learning to BREATHE on a range of hypothesized outcomes (such as perceived stress), the authors also assessed its feasibility and acceptability using both quantitative (attendance records and a perceived credibility scale) and qualitative (interviews with participating students) methods. Bluth et al (2015) derive a number of implications for future implementation from their findings, such as the importance of delivery context (e.g. *where* in school the intervention is delivered) and allowing curriculum adaptations to improve fit to local needs. Although only one statistically significant intervention effect was found (for depression), effect sizes associated with the mindfulness programme across a range of outcome variables suggested that an efficacy trial was warranted.

Mishna, Muskat, and Wiener's (2010) study details the development and pilot implementation of a manualized school-based group intervention for middle school students with learning disabilities, the goals of which included increasing self-advocacy skills and improving social adjustment. The authors' qualitative evaluation, which drew upon interviews conducted with teachers, students and their parents, concentrated primarily upon these stakeholders' perceptions of the intervention, and in particular the extent to which they felt that it was able to achieve its intended outcomes. There was also a specific focus on the acceptability of the intervention manual. From their analysis of this data, Mishna, Muskat and Wiener (2010) highlight a number of proposed 'practice principles' to inform future iterations of the intervention, alongside specific adaptations to content that were recommended by teachers (e.g. the alteration of sequencing such that a session on bullying was brought forward).

Box 4. Diverse examples of intervention pilot studies.

A salient reminder of how pilot studies can be so useful in allowing developers and evaluators to clarify assumptions about the processes underpinning an intervention and/or highlighting flaws in programme theory can be seen in Evans, Scourfield, & Murphy's (2014) qualitative case studies of the Student Assistance Programme (SAP) in four Welsh secondary schools, a targeted intervention for students exhibiting challenging behaviour. The authors identified notable iatrogenic processes associated with participation (e.g. negative labeling). Whilst not entirely unanticipated – iatrogenic processes and effects have been identified in other interventions – the authors' findings exposed assumptions regarding the process of change in SAP and the need to reconsider aspects of the intervention design in order to avoid unintended consequences.

2.6 Chapter summary

- Interventions may be defined as, "purposively implemented change strategies" (Fraser & Gallinsky, 2010, p.459).
- The 'lifecycle' of intervention development and research generally follows a sequence involving attempts to define and understand the problem to be addressed, designing and describing a proposed solution (including the articulation of an intervention theory), piloting and refinement, establishment of intervention efficacy, demonstration of intervention effectiveness, and finally scale-up and dissemination.
- Interventions can be characterised by a number of basic features, including (but not limited to) their form and function. The TIDieR framework represents an attempt to improve and standardise how interventions are described by developers and evaluators.
- Intervention theory may take the form of a logic model and/or theory of change
- The pilot stage may be used to assess the social validity (acceptability, feasibility, and utility) of an intervention.

- Pilot studies may be used as an opportunity to refine intervention design, materials and theory, but can also be used to optimise evaluation design.

3 IPE in randomised trials

It is perhaps inevitable that evaluators' main focus will be on outcomes when an intervention is examined in an RCT. However, IPE remains critical at this stage. Indeed, it is possible to argue the ability to make causal inferences about the intervention's influence on outcomes - one of the principal hallmarks of RCTs – is significantly weakened without the insights provided through attention to implementation and processes. As Bruhn, Hirsch and Lloyd (2015) state, "Changes in dependent variables can be attributed to the independent variable only when researchers can show accurate implementation" (p.336).

In this section we review common functions of IPE that span *both* efficacy and effectiveness trials. These are:

- documenting implementation
- assessing usual practice
- researching adaptations
- establishing differential intervention benefits

In the chapters that follow we review the literatures pertaining to additional IPE functions that are more likely to be associated with a specific trial stage (e.g. understanding the influence of contextual variation in effectiveness trials). Each of the core IPE functions noted above is examined in more detail below. As noted in the introductory chapter of this review, it has been proposed that without a comprehensive IPE, interpretation of outcomes is necessarily limited, and the probability of a Type III error (the inaccurate attribution of cause) is increased. For example, in the event of null findings (e.g. a failure to identify hypothesised intervention effects in an intention to treat analysis), lack of an IPE makes it nearly impossible to distinguish between theory failure, implementation failure and evaluation failure. Poor outcomes may thus be incorrectly attributed to an inaccurate programme theory, rather than the omission of critical intervention components during implementation, with the result that a potentially effective programme is abandoned prematurely. Even when effects are found, it cannot be assumed that these are directly attributable to a given intervention if it has not been confirmed that it was delivered as planned (Lendrum & Humphrey, 2012).

IPE in a trial context should aim to provide comprehensive coverage of the dimensions of implementation outlined in section 1.5. However, in the extant IPE literature it has been common practice to examine only a single dimension (Berkel, Mauricio, Schoenfelder, & Sandler, 2011), and consideration of fidelity/adherence or dosage have predominated to date. In Durlak and DuPre's (2008) review, for example, 63% of studies focused on fidelity, compared to only 10% that examined quality. Failing to consider the full range of implementation dimensions can lead to another Type III error. That is, if the only aspect of implementation that researchers assess is fidelity, it may be tempting to conclude that low fidelity is causal if there null intervention effects. In fact, these poor outcomes may be more accurately attributed to an aspect that has not been measured, such as quality or participant responsiveness. Similarly, if not all aspects of implementation are considered, and fidelity is measured as high, poor outcomes may be incorrectly seen as programme or theory failure (Lendrum, Humphrey, & Greenberg, 2016).

3.1 Research design in IPE

Before reviewing and discussing core IPE functions in trials, it is first necessary to establish the range of methods available to serve said functions. Our starting point in discussing these methodological considerations is that a pragmatic approach should be a given. Evaluators should embrace the methods most suitable for answering the primary research questions for their particular IPE. This is clearly evident in the range of design and methods on display on the IPE literature (contrast, for example, Boxes 5 and 6, below). Therefore, following Evans, Scourfield, and Murphy (2015) and Moore et al (2014), we argue that there is no such thing as a 'typical' IPE, and nor should there be.

One of the key merits of research in this field is the ability to draw upon a rich and diverse palette of methods.

Whatever the methods used, it is important to note at the outset of this section that the act of assessing implementation itself can influence intervention outcomes (although the same may also be said of outcome measurement - Song & Ward, 2015). For example, Smith, Schneider, and Smith's (2004) review of school-wide anti-bullying interventions found that those in which implementation was systematically monitored produced larger effect sizes than those where it was not considered. This may be the result of a 'researcher effect' on implementer behaviour. Implementers may pay greater attention to ensuring they deliver an intervention as intended by the developer when they know they are going to be observed and/or have to report on their activities during the course of implementation. Thus, gathering implementation data may have a 'sensitising' effect on implementers (Hansen, Pankratz, & Bishop, 2014). An alternative explanation may be found in certain systemic differences between programmes. That is, implementation monitoring may be standard for some interventions because of the emphasis placed on fidelity by developers (sometimes enacted through the use of an external coaching model), or because the structure and content of the programmes lend themselves to straightforward measurement (for example, the assessment of implementation may be simpler for manualised, prescriptive interventions than for those with a greater degree of flexibility built in to their design).

As a (relatively) new field, one of the major current challenges for implementation science is the establishment of rigorous methodologies for examining the various dimensions of and factors affecting implementation. As Ogden and Fixsen (2014) note, "even if the concept of implementation is not new, the idea of developing ways of measuring it certainly is. Consequently, there is a great need for the development of instruments which operationalise and standardise the measurement and analyses of implementation processes and outcomes" (p.8). The optimal balance between and integration of quantitative and qualitative methods of data generation and analysis in trial settings is a particularly salient issue in this regard. Reviews of IPE in trials have demonstrated significant growth of mixed methods approaches from around 17% of studies a decade ago (Oakley et al., 2006) to approximately 30% more recently (Borglin, 2015). This upsurge reflects the need for trial designs to mirror the complexity of the interventions being evaluated. Mixed methods designs offer considerable advantages over the use of quantitative or qualitative methods alone, including their complementary nature (e.g. the strengths of one method address the limitations of the other), increased credibility and trustworthiness (e.g. through triangulation of data), and inherent flexibility (e.g. ability to answer both explanatory and exploratory research questions) (Creswell & Piano Clark, 2011).

A number of mixed methods designs for IPE can be applied in the context of a trial. Broadly speaking, these vary in terms of the level of priority afforded to the different strands (e.g. equal, unbalanced), the amount of interaction between them (e.g. independent, interactive), their timing (e.g. concurrent, sequential, multi-phase), and the protocol for mixing them (e.g. during development, data collection, analysis, or interpretation) (Borglin, 2015; Creswell & Plano Clark, 2011). The reader is also referred to a review by Palinkas et al (2011), which describes the application of mixed-methods designs in school mental health implementation research, and delineates designs based on different structural arrangements of qualitative and quantitative methods, the diverse functions of mixed methods, and the varying ways of linking quantitative and qualitative data together.

In a *convergent parallel design*, quantitative and qualitative data are collected concurrently and with equal priority. Data strands are kept separate and are only integrated during the main interpretation of findings. By contrast, an *explanatory sequential design* utilises two distinct phases, beginning with a primary quantitative strand, and ending with a secondary qualitative strand. The findings of the first phase inform the focus of the second. So, for example, if quantitative data on implementation captured through structured observations in the first phase indicates poor intervention

fidelity/adherence and dosage, the reasons underpinning this may form the central focus of qualitative interviews with implementers in the second phase.

In a direct reversal of this approach, an *exploratory sequential design* prioritises qualitative inquiry initially, which is then followed up with a quantitative data strand. Take the case of a complex, multi-component intervention for which there is limited understanding of how to operationalize different implementation dimensions (e.g. what does fidelity look like for this intervention, and how can it be assessed?). Here, the IPE may begin with a primary, qualitative phase involving observations and interviews with key stakeholders (e.g. implementers, external support staff) and document analysis of intervention materials. The knowledge generated in this phase is then used to develop a quantitative measure of implementation that can be piloted with a view to subsequent descriptive and/or relational analyses (see 3.2.2 below). Finally, other mixed methods design options involve the blending of quantitative and qualitative strands from outset (known as *embedded design*) or the adoption of a cyclical process in which one or the other is prioritised at a given stage of evaluation, with each approach utilised providing a foundation for the next (*multi-phase design*). The latter typically comprises a series of studies carried out through the lifespan of an intervention (Borglin, 2015). An illustrative example can be seen in the work of de Visser et al (2015), who describe a multi-phase approach to the development and evaluation of an intervention to address harmful alcohol use among young people. This initially involved a large-scale quantitative survey of the correlates of alcohol use among 16-21 year olds, followed by a qualitative phenomenological analysis of interviews conducted with a purposive subsample of 25. Insights from the data generated in the first two phases were used to develop video resources for use in school-based alcohol education, the utility of which was subsequently analysed via 18 focus groups with students.

3.2 Quantitative methods in IPE

Robust measurement in IPE is just as essential as it is in trial outcome measurement, and there have been calls in the implementation science literature for the standardisation of measures and analyses (e.g. Ogden & Fixsen, 2014). However, this is ‘easier said than done’ given the sheer diversity of programme characteristics in school-based interventions (e.g. form, function, complexity and structure, prescriptiveness). In a multi-site trial, the Social and Character Development Research Consortium (2010) developed a common fidelity rating measure to assess the extent to which observed implementation met programme-specific benchmarks, allowing ratings to be compared across the seven different universal social and emotional learning interventions being evaluated. However, their analyses found no association between higher fidelity ratings and improved student outcomes, and the authors concluded that their standardised measure, “may not provide adequate information” (p.iii) to explain their findings. A subsequent attempt to develop a standardised implementation observation schedule (assessing adherence, dosage, quality, participant responsiveness and adaptations) by Bishop et al. (2014) was successfully piloted in a study of the All Stars school-based drug prevention intervention, showing good inter-observer reliability for all dimensions except for participant responsiveness. However, this instrument has yet to be tested on other interventions. A key interpretive problem posed is that standardised measures such as that used in these studies do not account for differences in implementation standards across interventions. For example, some interventions may be more difficult or complex to implement than others.

It is thus more common practice to use bespoke implementation measures that are designed with the characteristics of a specific intervention in mind (albeit based on common frameworks available in the literature – see, for example, section 1.5). For established interventions that have been subject to rigorous evaluation previously (as is the case in many EEF-funded evaluations), there are likely to be existing rubrics that can be adapted to develop observation schedules and/or implementer self-report surveys. For example, the implementation measures utilised in Humphrey et al's (2015) evaluation of the Promoting Alternative Thinking Strategies curriculum were based in part on those used in earlier trials of the intervention. In cases where the intervention is

being trialled for the first time, or where existing implementation measures are not available or are inadequate, programme materials (for example, implementation guidance documents), discussion with developer(s), and the IPE literature are likely to be the primary sources of information to inform measure development. Take the case of the evaluation of the KidsMatter school mental health initiative (Askell-Williams, Dix, Lawson, & Slee, 2013; Dix et al., 2010; Dix, Slee, Lawson, & Keeves, 2012; Slee et al., 2009) during which the evaluators developed an implementation index of fidelity, dosage and quality that was derived from relevant IPE theory (e.g. Domitrovich et al., 2008) and programme guidance materials, and populated with data from a variety of sources (teachers, parents and KidsMatter project officers).

Where bespoke measures are developed or adapted, standard methodological criteria should be applied in order to ensure that they are fit for purpose. For instance, data generation instruments should be piloted and refined ahead of their full-scale use in a trial[7], and psychometric properties for measures reported, including satisfactory inter-observer reliability in the case of observational methods. As with other methodological issues noted in this review, there is room for improvement in the field. For example, in their aforementioned review, Schoenwald and Garland (2014) found that only around one third of papers provided any data on the psychometric properties of the instruments used to generate implementation data. In terms of the process of measure development, guidance is available in the literature, although this tends to focus quite narrowly upon assessing and establishing criteria for fidelity as opposed to the full range of implementation dimensions (e.g. Mowbray, 2003; O'Donnell, 2008). Hansen's (2014) recent review is an exception in this regard[8]. Collectively, these outputs explore the relative value of different data sources and assessment frequencies (see next subsection), processes in determining acceptable thresholds of implementation (as might then be used as a filter for *on treatment* analysis (Sedgwick, 2015); see section 4.1), and the variety of data generation options available to evaluators. By way of example, Hansen (2014) discusses five different methods commonly used to assess dosage in school-based interventions (the total number of subject areas taught, students' exposure across the complete intervention based on records of attendance, time spent completing the programme, sessions implemented, and completeness of implementation). Finally, we note the use and potential of the Internet as a platform for data generation in IPE, which is sure to increase in the coming years. For example, in the aforementioned GBG trial, an online 'scoreboard' was developed through which teachers could record implementation data, thus providing important information on dosage (e.g. number of times the game was played per week, and for what length of time) and participant responsiveness (e.g. rule infractions).

3.2.1 Data source, method and frequency of assessment in quantitative IPE

There are essentially four main sources of data regarding implementation of school-based interventions. These are (i) researchers (e.g. IPE evaluators in a trial), (ii) implementers (e.g. teachers), (iii) administrative and external support staff (e.g. coaches), and (iv) intervention participants (e.g. students). Since use of the latter is confined almost exclusively to assessing participant responsiveness (Hansen, 2014; O'Donnell, 2008), our discussion focuses on the first three sources.

The two most widely used data sources in the IPE literature are researchers and implementers. In relation to the former, direct observation is the preferred methodology, either captured *in vivo* or through video recording (Hansen, 2014). The clear advantage of direct observation lies in the increased rigour provided by this method. It is generally accepted to be the most valid approach - in studies where implementation data is regressed onto intervention outcomes, stronger associations are typically found than for implementer self-report (Domitrovich, Gest, Jones, Gill, & Sanford Derousie, 2010). However, direct observation is a much more time consuming and costly method than

[7] Of course, this may be achieved during the pilot of the intervention itself.
[8] Although entitled, 'Measuring fidelity', this chapter adopts the 'macro' view of fidelity (see section 1.5), and thus actually encompasses the full range of implementation dimensions.

implementer self-report. For some interventions there may also be concerns regarding the extent to which certain aspects of implementation can be directly observed. Finally, observation necessarily provides only a 'snapshot' of implementation activity. Repeated observations circumvent this issue to a degree (and indeed, averaged ratings from multiple time points may be more strongly associated with outcomes than a single measurement; Resnicow et al., 1998), but also increase evaluation costs and data burden. In terms of the latter, teacher resistance to multiple observations may increase the risk of attrition.

By contrast, the benefits conferred by implementer self-report rest in the lower cost and time associated with the method and the fact that implementers can provide a summative account of a specified period of implementation (e.g. a teacher can be asked to self-report his/her overall implementation activity over the course of a school year) as opposed to the limited window provided by direct observation. However, implementer self-report data may be subject to several sources of bias, including impression management, demand effects, and/or differences in understanding of key aspects of the implementation requirements of a given intervention (Humphrey, 2013). This bias may skew analyses and indeed, several studies have failed to find a relationship between self-reported implementation activity and intervention outcomes (e.g. Goldberg Lillehoj, Griffin, & Spoth, 2004). The utility of implementer self-report data also rests at least in part on the proposition that implementers fully understand what is expected of them in terms of intervention delivery. This cannot be assumed.

Studies of the relationship between observer ratings and self-reported implementation provide useful insights into this issue. In one such example, Hansen, Pankratz and Bishop (2014) examined the degree of correspondence between observers' and teachers' ratings of fidelity/adherence, quality, reach and participant responsiveness during the implementation of the All Stars preventive intervention. They found that (a) teachers consistently displayed a substantial positive bias in their self-ratings (that is, they scored themselves higher than did observers, with a mean effect size of 0.93), (b) with the exception of reach, teacher and observer ratings of implementation dimensions were essentially unrelated (with a mean correlation of 0.117, indicating only 1% shared variance), and (c) only observer ratings were found to be meaningfully related to student survey measures. Broadly similar findings (positive self-report bias; generally weak relations between observer ratings and implementer self-report; observer ratings more strongly associated with intervention outcomes) have been reported in several other studies (e.g. Little, Sussman, Sun, & Rohrbach (2013) for the Pathways to Health intervention; Domitrovich et al (2010) for the HeadStart REDI programme).

Given the general trends apparent in the above studies, it is difficult to advocate the use of implementer self-report in isolation, despite the obvious advantages this method yields in terms of reduced evaluation costs. In the ideal set of circumstances, a multi-method, multi-dimensional approach is clearly preferable (Hansen, 2014; Little et al, 2013). Where this is not possible due to budgetary or other constraints and a single method is required, repeated direct observation (limited to a small number of repetitions to minimise data burden) should be considered the optimal approach to assessing implementation in a trial. However, consideration must be also be given to the type of trial being conducted – for example, repeated direct observation may be incompatible with the 'real world' delivery context required at the effectiveness stage.

Finally, the nature of some interventions presents opportunities for the collection of 'naturally occurring' quantitative IPE data from external support staff. By way of illustration, as part of the technical support and assistance role played by external coaches in the aforementioned Good Behaviour Game, observational data on fidelity is routinely gathered as the basis for discussion with the teachers implementing the programme, but this has also been utilised for evaluative purposes (e.g. Wehby, Maggin, Moore Partin, & Robertson, 2012). Similarly, teachers themselves are required to collect data on children's adherence to the four key classroom rules noted in Box 3 both during and outside of periods of implementation, providing a proxy for participant responsiveness (Chan, Foxcroft, Smurthwaite, Coombes, & Allen, 2012). Where it is

possible to obtain for use in the IPE of a trial, the minimal data burden imposed makes this an attractive option, albeit one that is also subject to certain limitations (e.g. ethical considerations regarding informed consent for use of data; possible bias introduced as a result of coach-teacher relations; natural restrictions on the nature of the data collected).

Where quantitative data is being collected regarding factors affecting implementation as opposed to dimensions of implementation, there is a much stronger tradition of using surveys. Given this tradition and the fact the various factors outlined in section 1.6 are universal, there exists a variety of instruments that could be integrated into an IPE. A useful starting point is Chaudoir, Dugan, & Barr's (2013) systematic review of measures for factors affecting the implementation of health innovations, in which the authors identified 62 measures at structural, organisational, implementer, participant, and intervention levels.

3.2.2 Quantitative data analysis in IPE

There is no single, agreed approach regarding the 'correct' way in which to model quantitative implementation data. The approach adopted will likely vary as a function of a variety of factors, including the skills and experience of the research team, the design of the trial, the characteristics of the intervention itself, the nature and form of the implementation data collected, and of course the intended purpose of the analysis (e.g. what is the research question being addressed?). Following Onwuegbuzie and Leech (2006), we make the basic distinction between using quantitative implementation data for *descriptive, comparative* or *relational* purposes. Thus, at the most basic level, quantitative implementation data can be presented descriptively as a means to provide evidence that a given intervention was delivered. For example, Piquette, Savage and Abrami's (2014) randomised trial of the effects of a web-based reading intervention ('ABRACADABRA') on pupils' literacy outcomes reports on observed implementation fidelity purely as a means of 'validating' the impact analyses that are the main focus of study. As previously discussed, this is a helpful preliminary use of IPE data as it strengthens the claim that any effects found are attributable to the intervention.

However, quantitative IPE data may also be used relationally. Such modelling may serve a variety of valuable purposes, including establishing whether implementation variability is predictive of outcome variability, empirical validation of programme theory, and the identification of critical intervention components. The majority of studies that have sought to link quantitative IPE data to intervention outcomes have utilised regression-based analytical techniques (for example, hierarchical linear modelling). Consider the examples provided below in Box 5:

Askell-Williams, Dix, Lawson and Slee (2013) examined the relationship between implementation variability in the KidsMatter mental health initiative and changes over time in primary school students' social and emotional competence. The authors developed an 'Implementation Index' focusing on fidelity, dosage and quality of delivery that drew upon survey data from teachers, parents and Kidsmatter project officers in the 100 participating schools (c.5000 children). Survey items from these sources were subjected to Latent Class Analysis in order to identify those with the greatest discriminative power. Visual binning was then applied to the resulting implementation score in order to group schools (e.g. low, average, and high). Finally, these groupings were entered as explanatory variables in a three-level hierarchical linear model (school, student, time) in which students' social and emotional competence (SEC) was the outcome variable. The authors found a significant interaction between index grouping and time, with significantly greater improvements in SEC over time in both average and high schools when compared to low schools. The effect sizes associated with these differences were small in both cases.

Birnbaum et al (2002) conducted a component analysis of the Teens Eating for Energy and Nutrition in Schools (TEENS) health behaviour intervention using data from their cluster-randomised trial (16 schools, c.3500 students). Natural variation in the implementation of TEENS allowed comparison of outcomes among students exposed to different components of the intervention (e.g. school environment, curriculum plus school environment, peer leaders plus curriculum plus school environment), with control group outcomes used as a reference point. Poisson mixed-model regression analyses revealed that differences in exposure to the various components of TEENS were associated with differences in the degree of change in students' eating patterns over time. To wit, incremental increases in exposure to intervention components were associated with greater change in health eating behaviour, with the largest effect sizes observed among those students who received all three components of TEENS. For example, this group reported nearly a half-serving increase in daily fruit consumption, compared to a quarter-serving increase among those exposed to two intervention components, and a fifth-serving decrease in the group who were only exposed to a single component.

Rosenblatt and Elias (2008) report on the implementation of a video-based social-emotional learning intervention ('Talking with TJ') in a study designed to assess its impact on academic achievement across the transition from elementary to middle school (7 schools, 10 classrooms, 145 students). The researchers used teachers' self-reported completion of intervention sessions as a measure of dosage, classifying classrooms as high, medium or low with reference to naturally occurring gaps in the distribution of these scores. A repeated measures analysis of co-variance (controlling for demographic differences and baseline emotional intelligence) demonstrated a significant interaction between dosage group and students' transitional grade point average (GPA) scores. Students in classrooms where higher numbers of Talking with TJ lessons were delivered showed significantly smaller drops in GPA in their transition year than did their peers in lower dosage classrooms.

Box 5. Examples of studies linking implementation variability to intervention outcomes.

Such studies highlight a range of methodological and analytical considerations regarding the ways in which implementation data are treated. First, as noted elsewhere in this review, evaluators may seek to document single or multiple aspects of implementation. Second, in cases where data on different aspects of implementation are recorded, these may be combined or reduced to provide a single proxy, or treated as distinct explanatory variables. Third, the implementation data itself may be recorded and treated categorically or continuously. Fourth, analyses may use control group data as a reference point, or simply examine the relationship between different levels of implementation and outcomes in the intervention group.

Published studies that link implementation variability to intervention outcomes also highlight a number of limitations that are inherent in this strand of the IPE literature. First, in spite of the language used in many studies (e.g. 'influence', 'impact'), the fact that implementation is naturally variable as opposed to experimentally imposed means that causal inferences must be tempered with caution (Nelson et al., 2012). Of course, there are a handful of exceptions in which evaluators have used random allocation to test the impact of different intervention components (for example, the aforementioned study of the impact of coaching in school mental health interventions by Cappella et al., 2012) but by and large implementation-outcome links should be described in terms of *associations* and *relationships* to reflect the correlative nature of findings (see section 1). Second, because of the reliance on natural variation, minimal variability may create problems for relational analyses. In studies where levels of implementation are uniformly high or low, this can reduce the power available to detect any relationship to intervention outcomes. In addition, where implementation is uniformly low there is a serious confound to the interpretation of trial findings (that is, the presence or absence of an effect cannot be reliably attributed to the intervention). Third, in studies where

implementation is treated categorically, designations given are usually arbitrary. Furthermore, as they are made with reference to locally obtained data rather than established thresholds, their comparability across studies is limited (that is, what is 'high' in study *x* may not be 'high' in study *y*) (Durlak & DuPre, 2008). Finally, there is a pressing need for the adoption and inclusion of effect size metrics to increase comparability of findings across IPE studies. However, Durlak (2010) cautions against the uncritical application of standard conventions for assessing the magnitude of effects (e.g. Cohen, 1992) as these ignore context and practical significance.

3.3 Qualitative methods in IPE

While the above provides primarily quantitative data and can thus be incorporated in statistical analyses, there is also a clear and very important role for qualitative data generation and analysis in IPE. In particular, qualitative IPE can help evaluators to (i) determine stakeholder views of intervention implementation (e.g. continued documentation of social validity), (ii) document the implementation of the intervention (e.g. fidelity, adaptations, and the reasons for them), (iii) highlight contextual influences on implementation (particularly those that may have implications for the transferability and/or sustainability of the intervention when it goes to the next stages of evaluation), and (iv) validate or modify the intervention theory by illuminating the mechanisms underpinning the change process (e.g. finding out more about how and why the intervention improves outcomes) (Atkins, Odendall, Leon, Lutge, & Lewin, 2015). Qualitative data derived from the IPE in a trial can help evaluators to understand why changes in usual practice (see section 3.5) have or have not occurred. This can provide additional insights into issues that are not always amenable to quantification. Illustrating this point, teacher interview data in the aforementioned PATHS trial revealed that the intervention was typically delivered during the timetable slot allocated for PSHE (thus providing information about displacement) and that the perceived similarity of the intervention to the SEAL programme acted as both a facilitator and barrier to implementation. Thus, while teachers were generally familiar and comfortable with the content of PATHS, some felt it was insufficiently distinctive (indicating low programme differentiation) (Humphrey et al, 2015).

Atkins et al (2015) propose several key steps in planning the qualitative strand(s) of an IPE. An initial task is to establish the role of qualitative researcher(s) relative to the rest of the trial team (for example, clarifying expectations and establishing a protocol for the nature and amount of contact with other trial team members such that any blinding is not compromised). The specific research questions to be addressed through qualitative methods should then be developed, including consideration of overlap and integration with quantitative methods where appropriate. Use of programme theory (e.g. the logic model and/or theory of change), implementation theory (e.g. normalisation process theory - May & Finch, 2009) and relevant classic theory (e.g. drawn from psychology, sociology and other fields, such as social learning theory – Bandura, 1986) to inform these questions is standard. The study design follows. Here, decisions are made in regarding the timing and fit of the qualitative IPE in relation to the trial as a whole (e.g. parallel vs. sequential - see section 3.1 on mixed methods designs), in addition to standard considerations such as overall approach (e.g. longitudinal case studies), sampling strategy (e.g. maximum variation), data generation methods (e.g. observation, interview, document analysis) and sources (e.g. teachers, external support staff, pupils, parents, school leadership), analytical level (e.g. manifest vs. latent) and orientation (e.g. inductive vs. deductive), and planned approach to integration/synthesis with quantitative findings (Saunders, 2016).

Examples of studies that have utilised qualitative IPE in a school-based trial context are included below in Box 6:

Gorard, Siddiqui, and Huat See (2015) conducted a qualitative IPE in the context of an effectiveness trial of the Philosophy for Children (P4C) programme. The authors made frequent visits to schools in the intervention arm of the trial to assess intervention fidelity, barriers to and facilitators of implementation, and participant perceptions (e.g. resentment or resistance). Observations of the initial P4C training for teachers and actual P4C implementation in schools were conducted, alongside interviews with the school lead for the intervention, teachers and pupils. Alongside this, eight participating schools volunteered for more in-depth P4C observations. Gorard and colleagues (ibid) found that P4C was delivered with variable fidelity throughout the trial (although interestingly, accompanying quantitative analysis of fidelity data failed to establish a relationship with outcomes); those schools rated poorly on this aspect of implementation were observed to be facing considerable challenges including staff turnover. Teachers reported challenges in delivering P4C at the recommended frequency due to perceived timetable inflexibility. Indeed, this was cited as the primary barrier to delivery. In spite of this, the authors reported generally positive perceptions of the intervention among teachers and pupils following initial resistance.

Humphrey, Lendrum, Wigelsworth and Kalambouka (2009) report on the implementation of the small group work element of the primary social and emotional aspects of learning programme. The authors' IPE, which comprised qualitative case studies of five primary schools, was conducted in the context of a large quasi-experimental trial (Humphrey et al., 2008) and focused on the processes and management of SEAL implementation in schools. Data was generated through interviews with school staff, children and parents, observations of intervention sessions and other settings, and document analysis, and was analysed using qualitative content analysis. The resultant process model indicated that successful implementation was contingent upon a range of factors, including the skills and experience of the group facilitator and the availability of an appropriate physical space to conduct intervention sessions. Key aspects of the delivery of small group interventions included setting achievable targets for children and providing opportunities for pupils to articulate their emotional experiences.

Chittleborough, Nicholson, Young, Bell, and Campbell (2013) conducted a qualitative IPE in the context of a major cluster randomised trial of the 'Hands Up For Max!' intervention to improve hand-washing in primary schools. In addition to basic quantitative data gathering relating to implementation (e.g. reach, dosage, fidelity) captured through log sheets, the authors also conducted interviews with staff co-ordinating intervention delivery, direct observations of intervention delivery, focus groups with pupils, and interviews with teachers in a small subsample of schools selected via stratified sampling on the basis of school size and proportion of children eligible for free school meals. The data generated was used to provide an additional source of evidence pertaining to fidelity, in addition to answering questions regarding the acceptability and sustainability of the intervention. In terms of the former, Chittleborough et al's (ibid) analysis highlighted issues relating to the perceived developmental appropriateness of the intervention materials. In terms of the latter, teachers reported that the intervention could be integrated into several areas of the curriculum and thereby taught on an on-going basis with ease.

Box 6. Examples of qualitative IPEs conducted in a trial context.

3.3.1 Putting the quality in qualitative IPE

Reporting of qualitative IPE is often rather poor. It is not unusual to see little or no description of, or justification for, the particular approach taken in analysis (e.g. thematic analysis, grounded theory, interpretative phenomenological analysis). Furthermore, there is often very little, if any, evidence of attempts to establish the quality, validity and credibility of qualitative IPE, with only limited information presented regarding the rationale for key aspects of research design (e.g. sampling approach). Although there are a small number of exceptions (see for example, Evans, Murphy, & Scourfield, 2015; Evans et al., 2014), this represents a serious and significant gap that means the true potential of qualitative IPE is not currently realised. This may be reflective of the lack of presence of qualitative specialists in trial teams (Atkins et al, 2015) and/or the priority afforded to quantitative data generation and analysis.

However, just as there are established standards for the design and conduct of quantitative intervention research (e.g. American Psychological Association, 1996; Torgerson & Torgerson, 2013), there exist equivalent guidance documents for qualitative evaluation (e.g. Henwood & Pidgeon, 1992; Spencer, Ritchie, Lewis, & Dillon, 2003) that could be adapted for use in IPE to inform design and/or assess quality. In this context, credit is given for criteria such as an appropriate research design, clear sampling rationale, analysis close to the data, evidence of explicit reflexivity, comprehensiveness of documentation, and negative case analysis. The nature of these and other quality factors in qualitative evaluation suggests a more subjective approach than would be the case in examining equivalent indicators in quantitative research (for

example, a trial used a randomised design or it didn't, and in a well written report this should be obvious to even the lay reader), but such is the nature of the beast. The various indicators in the aforementioned quality guidance documents may therefore be best thought of as a matter of degree as opposed to being binary (e.g. presence/absence). Integrated with the IPE literature, one can envisage a more substantive and informative role for qualitative inquiry in efficacy trials.

3.4 Documenting implementation

At the most basic level, IPE in a trial may provide descriptive information about the implementation of a given intervention that serves to increase the internal validity of the trial (e.g. to demonstrate that the intervention actually took place). However, despite the increased emphasis placed in IPE in recent years, even this basic function is often excluded in the reporting of school-based trials. For example, Bruhn et al's recent (2015) review of 'treatment integrity' in school-wide interventions reported that fewer than half measured and reported data on implementation. Similar findings were reported by Sanetti, Gritter, and Dobey (2011) in their review of interventions in the field of school psychology. This reveals a critical flaw in the reporting of most trials: "In the absence of objective and documented specification of an independent variable (i.e., operationally defined), as well as accurate independent variable application (i.e., treatment integrity), incontrovertible conclusions regarding the relation between a treatment and behaviour change are not possible" (Gresham, 2009, p.533). Of those studies that do provide data on implementation, a smaller proportion report analyses examining the relationship between implementation variability and outcomes. A recent estimate of 10% was offered by Schoenwald and Garland (2014) following their systematic review of over 300 studies reporting on psychosocial treatments for child or adult mental health problems. Examples of studies that have documented implementation-outcome relationships are provided in Box 5, in the section on approaches to analysis.

In terms of examining *why* implementation varies in a trial, it is necessary to return to the factors affecting implementation outlined in section 1.6. Given that, on the whole, reporting of implementation in school-based trials is inadequate (see above), it should come as no surprise that there is not a plethora of empirical literature on the factors affecting implementation in school-based interventions. This is in spite of an abundance of theory and conjecture (see, for example, Domitrovich et al's (2008) aforementioned implementation quality determinant framework). Thus, Durlak (2010) cautions that the relative importance of the numerous factors affecting implementation (at least 23, according to Durlak and DuPre, 2008) remains unknown. Exemplar studies that have sought to understand the influences on implementation variability can be seen in Box 7:

Adopting a qualitative case study approach, Carless (2003) explored the implementation of task-based teaching in three Hong Kong primary schools. Three teachers were observed 17 times during their implementation over the course of one school year. They also participated in six focused interviews and completed an attitude scale. The authors' thematic analysis of study data revealed six key influences on implementation: teacher understanding of task-based teaching, their attitudes towards it, the time available for implementation, the role of textbooks and topics, teacher preparation and resources, and the language proficiency of pupils. Through a qualitative mapping of these factors against the extent of each teachers' implementation, Carless (2003) attempts to demonstrate how their influence varies from individual to individual (e.g. time available is deemed to be a major influence for one teacher, but not the other two).

Drawing on survey data from a nationally representative sample of over 500 schools, Payne, Gottfredson and Gottfredson (2006) used structural equation modelling to examine hypothesized relationships between various exogenous community factors (e.g. poverty), school and programme characteristics (e.g. leadership support) and the intensity of implementation of school-based preventive interventions. The authors' analysis revealed (i) few direct effects of community factors on implementation, but various *indirect* effects of operating through school factors, and (ii) a number of direct effects of school and programme characteristics on implementation. For example, the extent to which a given programme was integrated into school operations predicted the frequency with which it was delivered to students, while local programme processes such as the amount and quality of training available for an intervention was related to its level of use throughout the school (e.g. how many staff were believed to be engaging in implementation).

Kallestad and Olweus (2003) examined the influence of a range of teacher- and school-level factors on the implementation of the Olweus Bullying Prevention Programme using multi-level modelling of data from 89 teachers across 37 schools in Bergen, Norway. Two measures of implementation were used, representing activity at the level of the classroom (e.g. use of different elements of the intervention, such as class meetings) and individual student (e.g. the teachers' contact and involvement with students identified as victims or perpetrators of bullying) respectively. These were regressed onto a number of teacher- (e.g. age, gender, readiness to intervene) and school-level (e.g. school climate) factors. All data were derived from self-report surveys. In the main classroom implementation model, 53% of the variance in implementation was explained by the combination of teachers' perceptions of the level of bullying in the school, whether they had themselves been victimised as children, the level of depth with which they reported having read the intervention manual, their affective involvement in cases of bullying, and the extent to which they rated themselves and their colleagues to be important in reducing bullying. Standardised regression co-efficients revealed the latter to be the most influential contributory factor.

Box 7. Examples of studies exploring factors affecting implementation of school-based interventions.

3.5 Assessing usual practice

Being able to define and document what constitutes 'usual practice' in an efficacy trial is a critical consideration for a number of reasons. First, in terms of those schools, classrooms/teachers or pupils assigned to the control arm of the trial, establishing the counterfactual is a necessary step in establishing causal effects (Maldonado & Greenland, 2002). The counterfactual approach argues that it is not possible to state that *x* causes *y* without knowing what *y* is when *x* is manipulated, whilst also showing that *y* does not occur in the absence of *x* (Little & Rubin, 2000). Put more simply, if we say that a given intervention is efficacious in improving children's progress in reading, a natural question that should follow is, "compared to *what*?" Thus, establishing the counterfactual – what happens in the absence of a change in *x* – is crucial. Documenting usual practice in as much detail as possible is particularly important in trials of school-based interventions because the idea of establishing a true 'no treatment' control group is increasingly accepted as unrealistic (Durlak & DuPre, 2008). The core components of most educational interventions are unlikely to be completely unique and will overlap – sometimes significantly – with what schools already do on a day-to-day basis (Nelson, Cordray, Hulleman, Darrow & Sommer, 2012).

Following on from the above, a second reason for assessing usual practice in the control arm of a trial is that this may change in response to randomisation. Schools that sign up to participate in a trial do so at least in part because the intervention being evaluated is seen to meet a perceived need. Thus, it is reasonable to expect that those who do not receive it may seek other means through which to meet this need during the trial period. This is known as compensatory rivalry (Conrad & Conrad, 2005), sometimes colloquially referred to as the 'John Henry' effect, and is considered to be a threat to the validity of a randomised trial because the attempts to 'equalise' the conditions undertaken by members of the control group confound the assumptions of the counterfactual. This is particularly salient when an intervention appears to produce null results. In such a

situation it may be tempting to conclude that it is not efficacious, but without having assessed usual practice in the control arm this may lead to a Type III error. The intervention may indeed be efficacious, but its effects may have been masked by compensatory rivalry (McMillan, 2007). Thus, repeated assessment of usual practice in control group contexts over the course of an efficacy trial is desirable in order to more fully support claims of causal effects (or lack thereof).

Assessment of usual practice should not, however, be confined to the control group of a trial. The third reason for assessing it is to establish *programme differentiation* in the intervention arm (see section 1.5). Interventions do not exist in a vacuum and those schools, classrooms/teachers or pupils allocated to the treatment arm will have an established pattern of usual practice in their attempts to improve the outcomes targeted in the trial. It is vital to document this so that the *distinctiveness* of the intervention (that is, how different is it to what we were already doing?) can be established. Such information can be useful in helping evaluators to understand how existing foundations affect the successful implementation of an intervention. On the one hand, higher programme differentiation ('revolution') may be advantageous because the intervention is seen as more distinctive and thus 'adds unique value'. Conversely, lower programme differentiation ('evolution') may be desirable because the intervention then feels more 'familiar' to staff and presumably will be easier to assimilate within existing processes and practices (Humphrey, 2013). As with the control group scenario, this may change during the course of the trial, but for different reasons. Some aspects of usual practice may be discontinued because space is needed in the school timetable for implementation (a 'displacement' effect) and/or because of a prevailing view that the new intervention meets perceived needs fully (a 'panacea' effect). If the aspects of usual practice that are discontinued are equally effective as the intervention in improving specified outcomes, this may lead to the masking of effects (and thus a Type III error in the absence of this information).

Thus, documentation of usual practice over time in all participating sites should be considered a critical aspect of the IPE in an efficacy trial, and will allow estimation of what Nelson et al (2012) refer to as the 'achieved relative strength' of an intervention. Unfortunately, there is a paucity of published work from which to draw exemplar methods of data generation and analysis. Indeed, in Durlak and DuPre's (2008) review of implementation studies, none of the studies examined included consideration of programme differentiation and/or monitoring of control conditions. In reports of efficacy trials, reference to activity in the control group is typically made, but this is often limited to a very short description with no indication of how this information was sourced, how much variability there was or whether any changes in practice occurred over time that may have influenced the trial outcomes (see for example, Stallard et al., 2014).

Among the very few studies in which usual practice data has been generated and analysed, those featured in Box 8 offer useful illustrations of possible methods that can be applied:

Civetelli (2012) examined programme differentiation in the context of a trial of Project Early Reading Intervention (ERI), a small group curriculum delivered daily for 30 minutes with at-risk students in kindergarten settings, conducted in 46 classrooms and focusing on 151 students. An observational tool (Reading Intervention Observational Tool, RIOT) was developed and utilized in both intervention and control classrooms to assess instructional content and teacher behaviours relating to reading support. Data generated through this process allowed the author to assess intervention fidelity (via reference to an exemplar delivery protocol) and programme differentiation (through comparison between intervention and control settings). The author's analyses indicated that (a) teachers in the intervention group did not adhere well to the ERI curriculum, and (b) the ERI and control groups were highly similar across many dimensions assessed in the RIOT.

Humphrey et al (2015) examined usual practice in relation to social and emotional learning as part of their trial of the Promoting Alternative Thinking Strategies curriculum (45 schools, c.4500 children). School level surveys were completed that documented involvement and implementation status of universal (e.g. the primary social and emotional aspects of learning (SEAL) whole school element) and targeted (e.g. nurture groups) interventions at baseline (prior to randomisation) and at 12-month follow-up. Analysis of the data generated indicated evidence of compensatory rivalry among schools in the control arm, with reported increases in both levels of SEL provision (relative to stability in the intervention arm). Given this, the data was incorporated in subsequent outcome analyses (e.g. changes in usual practice added as a school level co-variate in intention to treat analyses), enabling an estimation of the influence of these changes on trial outcomes. The authors were able to surmise that the changes in usual practice in control schools did not substantively affect the results of the study.

In Schonert-Reichl, Smith, Zaidman-Zait and Hertzman's (2012) trial of the Roots of Empathy (ROE) programme, the usual practice of both control and intervention teachers was assessed via self-report survey. Specifically, teachers were asked to rate the frequency with which they implemented social and emotional learning activities above and beyond those included in ROE. In addition, they were asked to report on whether they had been involved in any continuing professional development relating to social and emotional learning, and their beliefs regarding the importance of this aspect of education. The authors' analyses indicated that teachers in the intervention and control arms were equivalent in terms of training and beliefs, a significant compensatory rivalry effect was identified, with control teachers implementing significantly more social and emotional learning activities than their counterparts in the intervention group. The latter data was not subsequently modelled in relation to outcomes, presumably because the researchers found a number of intervention effects in their main analyses (thus, they were able to demonstrate the effectiveness of ROE even in the presence of compensatory rivalry, rather than using it as a potential explanation for the lack of a treatment effect).

Box 8. Examples of studies that have generated and analysed data on usual practice.

Ideally, documentation of usual practice in an efficacy trial occurs prior to randomisation at the study baseline and then is repeated at least once at an appropriate future point in time (e.g. at post-test). As noted above, both observational and survey methods can be used – the strengths and limitations of which are noted in section 3.2.1. It is likely that bespoke instruments will need to be developed that are matched to the aspects of practice that relate most directly to the intervention and its intended outcomes. By way of example, in the aforementioned EEF Good Behaviour Game trial, a teacher self-report survey was designed to assess use of general behaviour management strategies, rewards systems, approaches to managing disruptive and inappropriate behaviour, and implementation of proprietary interventions (e.g. Behaviour2Learn). This was based upon both existing research (e.g. Reupert and Woodcock's (2010) survey of behaviour management approaches among trainee teachers), the knowledge of the research team, and information on the core elements of the intervention (e.g. use of rewards, group membership – see Box 3). This survey is being repeated on an annual basis during the two-year trial.

3.6 Researching adaptations

As noted earlier, adaptations are seen as inevitable during the implementation of an intervention (Lendrum & Humphrey, 2012) and can be expected at any stage of the development and evaluation process (although they are perhaps more evident in effectiveness trial contexts, wherein researcher and developer influence are inevitably reduced). In researching the adaptation process, the IPE in a trial may consider, "when adaptations are being made, why they are being made, what they consist of, and how they align with the program's goals and theory" (Moore et al., 2013, p. 150). Moore et al (ibid) propose three dimensions of adaptation: *fit*, *timing* and *valence*. In terms of fit, they

make the distinction between adaptations made for *logistical* reasons and those made for *philosophical* reasons, the former being those that are made to accommodate pragmatic constraints (e.g. delivering an abbreviated version of a lesson in order to fit a pre-set timetable), and the latter reflecting differences in underlying conceptualisation (e.g. changing intervention content where it is considered inappropriate for a given audience).

Timing of adaptations may be *pro-active* or *reactive*. Pro-active adaptations are those that occur when problems of fit to local context/need are anticipated and changes made in advance. By contrast, reactive adaptations, are made 'on the fly' in response to issues that arise during the course of implementation. Finally, adaptations may be *positive* (in keeping with the goals and theory of the intervention), *neutral* (neither aligned with nor deviating from goals and theory) or *negative* (deviating from goals and theory) in valence. Examples of studies that have examined the adaptation process in the context of school-based interventions (including Moore et al's (2013) application of the above taxonomy in a study of the implementation of a range of evidence-based prevention curricula) are provided in Box 9:

Hansen et al's (2013) quantitative study examined teacher adaptations during the delivery of All Stars, a middle school drug prevention programme. Video recordings of implementation were coded, with a focus on the number and types of adaptation made and their valence (e.g. positive, negative, neutral) in relation to their service of the programme's goals. Measures of other dimensions of implementation were recorded, in addition to student and classroom outcome measures. The authors found that all teachers made adaptations, and that they were consistent over time in the nature of adaptations they made. In terms of the relationship between adaptational style and outcomes, a higher percentage of students remained non-drug users in classes of teachers who made relatively fewer adaptations, but whose average adaptation was rated as being positive. By contrast, teachers that made relatively frequent adaptations, regardless of their valence, had lower proportions of students who remained non-drug users. Finally, teachers who were judged to make positively valenced adaptations were rated significantly higher than negative adapters on most other measures of implementation (e.g. adherence, student engagement).

Ozer, Wanis and Bazell's (2010) primarily qualitative study explored the roll-out of two preventive school-based interventions – Project Towards No Drug Abuse (TND) and Too Good for Drugs and Violence (TGDV) in two urban school districts. Data was generated through observations, student and staff interviews, and consultation with programme developers. Using this data, the researchers were able to document the type of adaptations made. As with Hansen et al's (2013) All Stars study, all teachers made adaptations, which were primarily changes to instructional format, integration of real-life experiences into sessions (e.g. sharing examples from their own life), and provision of supplementary resources. The reasons cited for these adaptations were most commonly pedagogical or classroom management issues. Unlike Hansen's study, the researchers' quantitative analyses did demonstrate a trend towards increases in some types of adaptation over time. The authors speculate that this pattern (high fidelity in initial implementation, followed by increased adaptation) reflected the teachers' desire to improve the fit of the intervention to their local context (e.g. the needs of their students, their preferred pedagogical approach).

In a mixed-methods study, Moore, Bumbarger and Cooper (2013) examined the nature and rationale for adaptations made to 10 different evidence-based preventive interventions (including four that were school-based) when delivered in natural contexts. Data was derived from an annual survey conducted as part of a grant-funding agreement in the US State of Pennsylvania. Of the 240 respondents, just less than half reported making adaptations. The most frequently reported types of adaptations were changes to procedures (e.g. modifying the location of the intervention), dosage (e.g. altering the number and/or length of sessions) and content (e.g. adding or deleting activities). The most frequently cited reasons for these adaptations were lack of time, limited resources and difficulties retaining participants, all of which were noted by more than 70% of respondents. Accompanying analysis of qualitative examples of adaptations provided by respondents revealed that they were mostly made to improve logistical rather than philosophical fit, were proactive rather than reactive, and were more likely to be of a positive valence than a negative one.

Box 9. Examples of the study of adaptation of interventions in real-world settings.

3.7 Establishing differential intervention benefits

From a technical standpoint, planned sub-group analyses form part of the outcome assessment protocol for a trial. However, conceptually they may be thought of as part of an IPE as the purpose they ultimately serve is to improve our understanding of how and why the intervention works, which includes knowing for whom it is more or less effective (Slavin, 2012). Purists may argue that subgroup analyses are, "weak at best, misleading at worst" (Petticrew et al., 2012, p.95) because of the bias they can introduce when compared to the standard 'intention to treat' approach. However, their use in trials is defensible, provided that the subgroup(s) of interest and expected direction of effect are

specified in advance (that is, there is evidence that they were in fact *planned* rather than being introduced *post-hoc*) (Petticrew, et al, ibid). Participants in interventions are not simply passive consumers and we should not expect them to respond in a uniform manner (Bonell et al., 2012). Furthermore, sub-group analyses can make an important contribution to our understanding of the extent to which an intervention works to reduce established inequalities (Petticrew et al, 2012). For example, the EEF remit means that sub-group analyses that examine differential intervention effects for pupils who are eligible for free school meals is a standard feature in trial reports.

The process of identifying relevant sub-groups for analysis may be informed by theoretical considerations, empirical precedents, and/or extant frameworks. In terms of theory, this may relate directly to the intervention (e.g. the theory of change or logic model may imply or specify subgroup effects) or be drawn from the broader literatures that underpin it (e.g. risk and resilience theory). With regard to empirical precedents, earlier studies of the intervention (or interventions that are similar to it) may provide additional information. For example, in the aforementioned GBG trial, a planned sub-group analysis relating to boys with elevated levels of conduct problems at baseline was specified in the trial protocol on the basis of both theory (e.g. the gendered socialisation of competitiveness - Gneezy, Leonard, & List, 2009) and prior research findings (e.g. differential effects of the GBG for participants at-risk for conduct problems was established by Kellam, Rebok, Ialongo, and Mayer, 1994). Finally, frameworks such as PROGRESS-Plus (Kavanagh, Oliver, & Lorenc, 2008) delineate markers of some of the social determinants of inequality that may be used to identify populations for whom an intervention may produce amplified effects. Whatever methods are used, planned subgroup analyses require a compelling case. To increase the rigour of this aspect of outcome analyses, Sun, Briel, Walter, and Guyatt (2010) propose an 11-point checklist for assessing the credibility of subgroup effects that spans trial design (e.g. was the direction of the subgroup effect specified in advance?), analysis (e.g. is the subgroup effect independent?) and context (e.g. is the size of the subgroup effect meaningful?).

3.8 Chapter summary

- A mixed-methods IPE approach involving both quantitative and qualitative data is preferable to just one or the other. A variety of designs may be utilised; these vary in terms of the level of priority afforded to the different strands (e.g. equal, unbalanced), the amount of interaction between them (e.g. independent, interactive), their timing (e.g. concurrent, sequential, multi-phase), and the protocol for mixing them (e.g. during development, data collection, analysis, or interpretation).
- Quantitative methods of data generation in IPE tend to use bespoke (e.g. intervention specific) instruments, which should be subject to the standard considerations in measure development (e.g. establishment of psychometric properties, including inter-rater reliability).
- There are three primary sources of data for quantitative IPE: researchers (e.g. independent observer), administrative and external support staff (e.g. coaches), and implementers (e.g. teachers). Each has particular strengths and limitations, although direct observation by independent researchers is generally considered to be the preferred approach because of the increased objectivity and rigour afforded by this method.
- In terms of frequency, averaged implementation ratings from multiple time points may be more strongly associated with outcomes than a single measurement; however, there are associated increases in evaluation costs and data burden.
- Quantitative IPE data may be generated for descriptive or relational purposes. Relational analyses typically utilise regression models and can be used to establish whether implementation variability is predictive of outcome variability, empirically validate programme theory, and/or identify critical intervention components.
- Qualitative IPE data can help evaluators to determine stakeholder views of intervention implementation, document the implementation of the intervention,

highlight contextual influences on implementation and/or validate or modify the intervention theory by illuminating the mechanisms underpinning the change process.

- Rigorous qualitative IPE is evidenced by an appropriate research design, clear sampling rationale, analysis close to the data, evidence of explicit reflexivity, comprehensiveness of documentation, and negative case analysis. Such evidence enhances the quality, validity and credibility of qualitative IPE.
- Documentation of implementation is a fundamental aspect of trial stage IPE. It should be done as comprehensively as possible. Data generated can be used to enhance the internal validity of trial findings (e.g. to demonstrate that the intervention took place, and/or that implementation variability is related to outcome variability).
- Assessment of usual practice through IPE is critical in terms of establishing the counterfactual, exploring the extent to which this changes through the course of a trial, and establishing programme differentiation.
- Researching the adaptation process helps us to understand, "when adaptations are being made, why they are being made, what they consist of, and how they align with the program's goals and theory" (Moore et al., 2013, p. 150).
- Adaptations may be understood in terms of fit (e.g. are changes made for logistical philosophical reasons?), timing (e.g. are changes made pro-actively, or reactively?) and valence (e.g. are changes in keeping with programme theory?). Pre-specified sub-group analyses can help evaluators to understand whether an intervention produces differential benefits for certain participants.

4 Additional IPE functions in efficacy trials

The first stage of trial-based evaluation examines whether an intervention achieves its intended outcomes under optimal conditions. Put another way, it is the stage at which evaluators ask, “Can it work?” (Haynes, 1999). This is most commonly known as an efficacy trial (Flay et al., 2005; Gottfredson et al., 2015) but is also variously referred to as *definitive* (Campbell et al, 2000) or *explanatory* (Roland & Torgerson, 1998) trials. Use of intervention theory and pilot stage findings are critical here, not least in defining what conditions are ‘optimal’. There is often an implicit assumption that implementation will be of uniformly high quality in efficacy trials (see for example Flay et al., 2005), but this is rarely, if ever the case in school-based interventions (Bruhn, Hirsch, & Lloyd, 2015). As noted at the outset of this review, we know that implementation variability can predict outcome variability. Furthermore, interventions do not happen in a vacuum, and so understanding context and social processes is crucial, even when the ultimate focus may be on outcomes. IPE therefore continues to be essential and integral to intervention research at this stage of development and evaluation.

As indicated elsewhere in this review, a high quality trial-based IPE allows researchers to ‘look inside the black box’ (Saunders, Evans, & Joshi, 2005). In addition to the four core functions noted in the previous chapter, IPE in an efficacy trial may also serve to (a) examine the relationship between ‘on treatment’ status and intervention outcomes and (b) empirically validate the intervention theory (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012).

4.1 ‘On treatment’ status and intervention outcomes

As noted in the previous chapter, quantitative IPE data can be used to establish the relationship between implementation variability and intervention outcomes. A rudimentary approach utilised in some clinical trials is ‘on treatment’ (sometimes also called ‘per protocol’ or ‘adherence to protocol’) analysis, in which only outcome data for participants known to have completed a given intervention protocol is assessed (Sedgwick, 2015). Comparison with findings from an intention to treat analysis – in which *all* potential participants are included - can enable a basic estimate of the maximum *potential* benefit of the intervention. When applying such techniques to data from a range of drug prevention trials, Derzon, Sale, Springer, and Brounstein's (2005) meta-analysis found that intervention effect sizes rose twelvefold. However, per protocol analysis can be difficult to apply in school-based efficacy trials because of the lack of established thresholds for describing the delivery of a given programme as ‘on treatment’ and the complex, multi-dimensional nature of implementation (for example, dosage may be per protocol but not fidelity/adherence). However, there are published examples in the literature. For example, Berry et al (2015) report an on-treatment analysis of the PATHS curriculum, in which they set a threshold of 80% fidelity (sourced from coach/teacher ratings) as a means to compare the intervention outcomes of children in classrooms where implementation was deemed to fall below or exceed said threshold.

4.2 Empirical validation of programme theory

An important but rarely discussed flaw in the traditional ‘black box’ RCT design is that even if significant intervention effects are identified, causal claims may still be limited (or, at least, attenuated). In the absence of good IPE data we do not know how (or, in some cases, if) the intervention was delivered and whether the means by which it was presumed to effect change are in fact valid. Thus, another function of IPE in efficacy trials can be to inform empirical validation of programme logic models and/or ToC. This may include modelling of mechanisms and processes that are thought to trigger change to demonstrate their association with outcomes, and/or exploring the temporal relationship between outcomes. The latter is particularly pertinent when the outcome of interest is distal to the intervention and change is assumed to be underpinned in some way by direct effects on more immediate, primary outcomes. This is likely to apply to a large number of interventions that receive EEF grant funding, where improvement in academic attainment may be theorised as an *indirect* consequence of change in other

outcome domains (e.g. behaviour). By way of illustration, consider the work of Banerjee, Weare, and Farr (2014), who analysed a proposed theory of change for the social and emotional aspects of learning (SEAL) programme. Using multi-level path analysis of data drawn from a large number of schools and pupils, the authors demonstrated that implementation quality ratings were significantly associated with school ethos, which in turn mediated associations with pupils' social experiences, overall school attainment, and persistent absence.

4.3 Chapter summary

- In addition to the four core functions noted in Chapter 3, the IPE in an efficacy trial may also serve to examine the relationship between 'on treatment' status and intervention outcomes and empirically validate the intervention theory.
- On treatment analysis involves comparison of overall intervention effects with those for whom the intervention protocol is known to have been completed to a pre-specified standard.
- Empirical validation of programme theory may include modelling of mechanisms and processes that are thought to trigger change to demonstrate their association with outcomes, and/or exploring the temporal relationship between outcomes.

5 Additional IPE functions in effectiveness trials

If an efficacy trial asks, “Can it work?” in relation to an intervention, an effectiveness trial (also referred to as a ‘pragmatic’ trial; Roland & Torgerson, 1998) poses the question, “Will it work?” (Haynes, 1999). The shift in emphasis between efficacy and effectiveness trials relates to moving from internal to external validity in the process of establishing whether and how an intervention improves outcomes. Thus, while efficacy trials maintain high levels of internal validity through researcher and/or developer control over the conditions of implementation, effectiveness trials focus on the increased external validity conferred by trialling the intervention in ordinary, ‘real world’ conditions. Given this shift, IPE remains a critical aspect of trial design. This is particularly the case given the observed ‘voltage drop’ in outcome effect sizes as an intervention moves from efficacy to effectiveness (see for example, Wigelsworth et al., under review).

In addition to the four core functions noted in the Chapter 3, IPE in an effectiveness trial may also serve to (a) enhance our understanding of contextual influences, and (b) distinguish between different intervention components and identify those that are critical (also known as ‘active ingredients’) through analysis of natural variation or experimental manipulation (Bonell, Fletcher, Morton, Lorenc, & Moore, 2012).

5.1 IPE challenges in effectiveness trial contexts

Integration of IPE in an effectiveness trial poses a couple of immediate challenges. First, given the reduced evaluator and/or developer control at this stage of evaluation, there is an additional ‘burden of proof’ to demonstrate that an intervention was actually delivered in real world conditions in effectiveness trials (Flay et al., 2005; Gottfredson et al., 2015). Here then, an important initial task for evaluators is to rigorously document the conditions of implementation. For example, confirmation that intervention materials, training and technical support are readily available is necessary (Gottfredson et al., 2015), alongside a clear description of the range of settings and/or populations that are being examined in the trial. Second, implementation will be considerably more variable in an effectiveness trial (Glasgow, Lichtenstein, & Marcus, 2003). Increased adaptation to improve the fit of the intervention to local context and need is particularly likely. As such, capturing this in a meaningful way should be a key focus in an effectiveness trial IPE (see section 3.6). Third, we have already noted that gathering implementation assessment data may have a ‘sensitising’ effect on implementers, thereby inadvertently changing their behaviour (Hansen et al., 2014). Thus, a comprehensive IPE in an effectiveness trial may have a confounding influence, particularly if a given intervention does not typically include systematic monitoring of implementation. This challenge is more difficult to address, and may therefore require some methodological trade-offs. For example, despite the limitations of the method noted above, assessment via implementer self-report may be preferable to direct observation when evaluating an intervention under effectiveness conditions because of the increased researcher effect that may eventuate from the latter. Finally, the shift from efficacy to effectiveness brings with it additional factors affecting implementation. For example, in an efficacy trial, issues of funding/cost and maintenance are typically not immediate considerations because intervention materials, training and technical support are made available at no cost to schools (or at least are heavily subsidised). Similarly, the increased range of settings and/or populations implied in an effectiveness trial means that greater attention needs to be paid to issues of transferability and generalisability.

5.2 Understanding contextual influences

One of the defining features of an effectiveness trial when compared to an efficacy trial – aside from the reduced level of evaluator or developer control – is that the range of contexts of implementation becomes less restricted. Indeed, one of the central aims of an effectiveness trial is to examine the extent to which the effects established in an

efficacy trial can be generalised to a broader range of contexts[9] (Gottfredson et al., 2015). The importance of establishing, exploring and explaining contextual influences on intervention effects cannot be overstated: "It is the variability introduced by these contextual differences that creates uncertainty regarding the potential of an intervention to be brought to scale" (McDonald, Keesler, Kauffman, & Schneider, 2006, p.16). It should be no surprise at this point that we make the claim that IPE is integral to understanding how an intervention works in these new contexts. An effectiveness trial IPE should be designed in such a way that relevant data at school, implementer and participant level can be collected in order to assess if, how and why contextual variation influences implementation, and where possible to establish the resultant impact on the achievement of intended outcomes. As with other stages of programme development and evaluation, the data produced can be used for summative purposes (e.g. establishing the magnitude of contextual moderation effects on implementation) or may be formative (e.g. exploring what modifications are required in order to successfully replicate previously established intervention effects in a new context) (McDonald et al, 2006).

By way of an empirical example, consider Pas, Waasdorp, and Bradshaw's (2015) 4-year study of the influence of teacher, classroom, and school characteristics on the implementation of the Positive Behaviour Support (PBS) intervention. Utilising multi-level modelling of data collected as part of an effectiveness trial, the authors identified a range of school-level contextual factors (e.g., school size, behavioural disruptions) and teacher-level factors (perceptions of school organisational health and grade level taught) that were significantly associated with variability in PBS implementation quality. Although this study did not report on whether these associations moderated intervention outcomes, an earlier review of 16 studies of school violence interventions by Ozer (2006) suggest an, "interdependence of context and implementation factors" (p.315) was likely to moderate intervention effectiveness. Although the examples cited here are quantitative in nature, there is a clear and essential role for qualitative inquiry in this process. For example, qualitative data drawn from key intervention stakeholders may yield powerful insights into the mechanisms underpinning any context-implementation-outcomes relationships (e.g. how and why do different contextual influences affect implementation and/or outcomes?), or indeed aid in the process of mapping out the myriad theorised contextual influences at the school and classroom levels ahead of a quantitative investigation. Such research may also signpost broader contextual factors that support or inhibit implementation efforts, such as shifts in education policy: "Implementation does not occur within a vacuum, but rather reflects and is strongly influenced by school reforms and initiatives that occur within a shifting landscape of sociopolitical priorities and policies at [multiple] levels" (Han & Weiss, 2005, p.666)

5.3 Critical component analysis (CCA)

Distinguishing between different components of an intervention and identifying those that are necessary in order to trigger change (e.g. those that may be considered 'active ingredients') is a critical IPE-related task, for several reasons. First, given the inevitability of adaptation, understanding *what actually makes a difference* – particularly in a complex intervention – will help to ensure that implementers make adaptations that do not compromise the achievement of intended outcomes (Blase & Fixsen, 2013). Historically, this issue has not received enough attention. Dane and Schneider (1998) note that intervention manuals and materials typically do not specify essential intervention components, leaving implementers unsure about what can (and cannot) be adapted. Indeed, Durlak (2010) acknowledges that, "the core elements of most psychosocial and educational interventions are unknown" (p.351). Second, CCA can be useful in streamlining the design of an intervention such that it is more simple/feasible and cost-effective to implement. For example, consider a school-based intervention delivered by teachers in which external coaching is utilised as part of the implementation support system. This is likely to be a costly element of the intervention, and also one

[9] In referring to a broader range of contexts we mean not just to the specific sites at which the intervention is delivered, but also the individuals who deliver it and the participants who receive it (Gottfredson et al, 2015).

that may become a barrier to its maintenance when it moves to effectiveness and/or scale-up conditions (e.g. there may be an insufficient number of suitably qualified coaches, or schools may object to the cost of accessing their support) (Elliott & Mihalic, 2004). If a programme developer considers external coaching to be necessary for change to occur, CCA offers an opportunity to establish this in a rigorous manner (whilst also providing evidence as to why this is the case). For example, a randomised trial by Cappella et al (2012) demonstrated that the combination of training plus on-going coaching support for teachers implementing a school mental health intervention produced better outcomes than training alone. In the case of an intervention in which multiple, discrete components have been identified, this approach can be extended with the use of more complex factorial experimental designs. Of particular note here is the multiphase optimisation strategy (MOST) methodology proposed by Collins et al (2010), which offers considerable promise as a tool for CCA. In this approach, different components are systematically experimentally manipulated such that an optimal version of the intervention can be determined.

Operationalising and validating critical intervention components can, however, be complex. Interventions are often conceived of and evaluated in their entirety, making it difficult to isolate and manipulate individual components (Elliott & Mihalic, 2004). A useful framework to help developers and evaluators work through this challenge is offered by Blase and Fixsen (2013). They suggest that core components may be cast as theory-based and empirically derived principles, operationalised as contextual factors (e.g. the particular setting in which the intervention occurs), structural elements (e.g. the obligatory number and/or sequence of sessions) and specific practices (e.g. reinforcing appropriate behaviour). It is in the consideration of structural elements and specific practices that there is considerable overlap with the aforementioned implementation-outcome analyses (see previous subsection). If there is a distinction, it is that where standard implementation-outcome analyses aim to elucidate a relationship between a given aspect(s) of implementation and a specified outcome(s), CCA seeks to extend this by establishing a clear threshold for what may be deemed acceptable in terms of level of implementation.

Using this framework, CCA could be carried out in a number of ways. For example, a CCA focusing upon structural elements of an intervention may analyse the relationship between the number of sessions delivered and the achievement of intended outcomes as a means to establish the 'minimum effective dose' (Liu, 2010), or compare outcomes when defined components are present/absent. For example, Bloomquist, August, and Ostrander (1991) examined the effects of a multi-component school-based cognitive behavioural intervention for children with ADHD, directly comparing outcomes among children exposed to a streamlined version (teacher component only) to those in receipt of the full intervention package (teacher, parent and child components), with a waitlist control group used as a reference point. However, such examples highlight that CCA is perhaps better suited to simpler interventions that involve a high degree of prescription (e.g. those that are modularised/manualised) than those that are inherently more complex, flexible and/or where the identification of 'discrete' elements is problematic. For these more complex interventions, CCA fails to account for the fact that different intervention elements are intended to interact and produce synergistic effects that are greater than those of each in isolation (Bonell et al., 2012). An added difficulty is that what is 'critical' may depend upon the particular needs of participants. For example, the minimum effective dose of an intervention may vary as a function of the level of risk exposure among individual children.

5.4 Chapter summary

- Effectiveness trials pose additional challenges for IPE, including greatly increased implementation variability, and the introduction of additional factors affecting implementation.
- IPE in effectiveness trials may also be used to establish, explore and explain contextual influences on intervention effects.

- Critical component analysis (CCA) can help evaluators to distinguish between different components of an intervention and identify those that are necessary in order to trigger change.

6 The role of IPE in taking interventions 'from science to service'

At the point in the development and evaluation cycle in which both efficacy and effectiveness have been established, the focus shifts to developing knowledge and understanding of how the now 'proven' programme may successfully be brought to scale. In August, Gewirtz, and Realmuto's (2010) terms, evaluators ask how we can take an intervention, "from science to service" (p.72). This is an issue that warrants considerable attention, not least to provide a better understanding of what makes some interventions (or intervention features) more 'scalable' than others. There is a broad consensus that the education community does not have a strong record in terms of adopting empirically validated interventions, and even where these are used, the 'evidence to routine practice' lag can be up to 20 years (Walker, 2004). Two related issues are central to IPE at this stage – *scaling up* (bringing the intervention to a wider audience) and *sustainability* (maintaining effective use and impact of the intervention) (Forman, 2015). Collectively, this *Type II translational research*, "examines factors associated with the adoption, maintenance and sustainability of science-based interventions at the practice level" (Greenberg, 2010, p.37)[10]. At this stage, implementation becomes the outcome of interest. IPE can be used to understand the factors that influence intervention engagement and reach (e.g. who takes on the intervention and why?), implementation quality (e.g. when the intervention is delivered well, what supports this?) and sustainability over time (e.g. what is sustained? How?) (Greenberg, 2010). Given the likelihood for further drift and dilution at the scale-up stage, IPE can also serve an important function in documenting how and why the intervention evolves as it goes to scale (e.g. what changes and why?).

6.1 Theories, models and frameworks to inform IPE at the dissemination stage

A number of models exist that can inform the focus of IPE at this stage in the intervention development and research cycle (Rabin, Brownson, Haire-Joshu, Kreuter, & Weaver, 2008). Among these, the following offer considerable utility:

- Rogers' (2003) diffusion of innovations theory
- Glasgow, Vogt, and Boles' (1999) RE-AIM framework
- Fixsen, Naoom, Blase, Friedman, and Wallace's (2005) five-stage model of the implementation of evidence-based interventions
- Proctor et al's (2011) taxonomy of implementation outcomes

Rogers (2003) proposed that the spread and adoption of new interventions could be best understood through consideration of (1) the innovation itself, (2) the channels through which it is communicated, (3) the passage of time, and (4) the social system in which diffusion takes place. So, for example, in relation to the intervention itself, IPE at this stage may examine the extent to which the rate of adoption varies as a function of its *relative advantage* (e.g. is it viewed as being 'better' than that which precedes it?), *compatibility* (e.g. does it fit with adopters' values, experiences and needs?), *complexity* (e.g. is it simple to understand and use?), *trialability* (e.g. can it be experimented with?) and *observability* (e.g. are the outcomes of the intervention readily visible?). By contrast, Glasgow and colleagues' (1999) RE-AIM framework emphasises a five-step approach through which the process of scaling-up a scientifically validated intervention can be best understood, as follows: (1) Reach – the extent to which an intervention is truly generalisable through consideration of the representativeness of samples and the specific populations from which they are drawn; (2) Efficacy and effectiveness – the evidence of the intervention's impact on outcomes of interest under both optimal and real-world conditions; (3) Adoption – understanding the characteristics of adopters and the settings they inhabit; (4) Implementation – how the intervention is delivered following adoption; and (5) Maintenance - the extent to which an intervention becomes part of routine practice over time.

[10]Strictly speaking, Type II translational research also encompasses the effectiveness stage of intervention development and evaluation (Rabin et al, 2008).

Fixsen and colleagues (2005) propose five stages through which an evidence-based intervention passes during its dissemination. First, *exploration and adoption* occur – someone at an organisation (e.g. a school) becomes aware of an intervention and a process of evaluation occurs in which its fit, price and other important factors are considered ahead of an adoption decision. Second, *programme installation* begins, including alignment of resources, training and orientation of staff to encourage them to engage with the intervention. Third, *initial implementation* commences, during which the integration of the intervention into the organisation occurs. At this stage, the fit of the programme to the organisational culture and climate is a primary determinant of continued progress. The fourth stage is *full operation*, at which point the intervention has become fully embedded as part of usual practice, and is therefore no longer considered to be 'new' or 'innovative'. Finally, the process of ensuring *sustainability* takes over, during which staff are most likely to make adaptations to fit local conditions. Given the passage of time implied and the organisational changes that may have occurred up to this point, the potential for programme drift also increases. In applying this model, an IPE may, for example, focus on the fourth stage and examine the processes through which an intervention becomes fully embedded and 'disappears' into routine practice.

Finally, Proctor et al (2011) propose a taxonomy of eight distinct implementation outcomes, the achievement of which would constitute success in the context of scale-up and dissemination. Bringing us full circle to the concept of social validity (see section 2.5), the authors argue for the need to assess perceptions of *acceptability*, *appropriateness*, and *feasibility* of the intervention in diverse range of contexts implied during large-scale dissemination efforts. Alongside this, examination of *adoption*, *cost*, and *penetration* are proposed in order to provide insights regarding the uptake of the intervention. Finally, a return to the assessment of *fidelity* is proposed as a means to examine the extent to which the intervention is delivered as intended as it is brought to scale, along with attention to *sustainability*, in order to determine whether implementation is maintained and/or fully 'institutionalised'.

Examples of papers that have examined issues relating to the scaling-up and dissemination of interventions are included below in Box 10. Accompanying notes illustrate how the theories outlined above may be applied to enhance understanding of the processes at work:

In Askell-Williams, Slee, and Van Deur's (2013) study, the authors followed up schools that had been involved in a national pilot evaluation of the aforementioned KidsMatter mental health initiative. In the context of the national scale-up that followed the success of the pilot, Askell-Williams and colleagues were interested in which of the 100 pilot trial schools continued to implement the programme, to what degree, and why. Their analysis indicated that a significant minority (20%) were no longer implementing Kidsmatter, and among those that were, a number were selectively implementing certain components and only using resources that were thought to be useful. The key issues that challenged sustainability and the maintenance of quality implementation included: changing and competing priorities, leadership change, structural transformation (e.g. school mergers), and external factors (e.g. lack of continued external support or promotion of the programme). Interestingly, some schools reported that since they were now routine, they no longer identified certain practices as being 'Kidsmatter', providing an indication that they had reached the fourth stage in Fixsen et al's (2005) model (see above).

Olweus and Limber (2010) report on the dissemination of the Norwegian Olweus Bullying Prevention Programme (OBPP) in the United States. They report a number of challenges, including resistance from staff, school leadership and/or parents that was indicative of a lack of perceived need and/or benefit, incompatibility of existing norms and practices with OBPP, and the tendency of schools to selectively implement those aspects of the programme that were seen as 'easy', while neglecting those that required more effort. These issues highlight the utility of the various theoretical frameworks that are available to the IPE researcher. So, from a social validity perspective (Wolf, 1978), the resistance of staff illustrates the lack of traction gained when an intervention is not seen as socially significant (e.g. Olweus and Limber report that some staff viewed bullying as a rite of passage that could even be a positive learning experience, and thus not worthy of investment in intervention). Viewed through Rogers' diffusion of innovation lens, we can see how potentially critical intervention components may be omitted if aspects of the intervention are perceived to be too complex or difficult to implement. Finally, there is also direct evidence in support of Fixsen et al's (2005) proposal that a lack of fit with organisational culture and climate can significantly hinder progress during the initial implementation stage.

Denton, Vaughn and Fletcher's (2003) review explores the factors influencing the sustained high-quality implementation of reading interventions which have been demonstrated to be effective, focusing in particular on two that have been successfully disseminated: Reading Recovery and Success for All. Among the various factors identified, the authors noted the importance of opportunities for on-going mentoring and assistance for teachers to enable them to achieve 'practice mastery', teachers' acceptance and commitment to the programme, the presence of an influential programme leader in school, support from school leadership, and an appropriate programme scope that has a high degree of specificity and is perceived as relevant and useful to teachers.

Box 10. Examples of research and reviews focusing on issues affecting the scaling-up, dissemination and sustainability of evidence-based interventions.

6.2 IPE and the 'core challenge' of achieving population impact

Having explored and established the central role of IPE at different stages of the intervention lifecycle (e.g. development and piloting, efficacy, effectiveness, scale-up and dissemination), it seems fitting to return to where we began as a reminder of why those involved in intervention development and evaluation engage in this lengthy, complicated and resource-intensive process. In section 2.2 we made reference to 'defining and understanding the problem', using the example of the substantial academic achievement gap between students from disadvantaged and more affluent backgrounds. The ultimate goal – indeed, the holy grail – for those involved in this kind of work (e.g. intervention developers, evaluators, policy-makers, and organisations such as the EEF) is to be able to address such major problems at a societal level; that is, to achieve (and evidence) population impact. However, it is at this final stage where our knowledge is ultimately at its most limited. Spoth et al (2013) highlight the general failure of evidence-based interventions (e.g. those that have passed successfully through tests of efficacy and effectiveness) to penetrate and gain traction in delivery systems such as schools, and lament the extremely limited amount of Type II translational research that has been conducted to date in order to help us understand how such 'proven' approaches may be more successfully integrated into policy and practice on a large scale. The authors propose that two major challenges need to be addressed in order for progress to be made. First, they propose the need to develop significant infrastructure and capacity to support both implementation and accompanying research efforts during the scaling up process. Second, they suggest that significant advances are required in terms of the research agenda, design and methods associated with Type II research. It is our contention that this may be tied to achieving shifts in policy-level thinking, including changing forms of accountability which are accepting of the need to achieve process-based change and interim outcomes before primary outcomes can be considered; which

accept that *both* qualitative and quantitative IPE data are integral to developing a comprehensive understanding of interventions; and which understands that evaluation design must mirror the complexity of the intervention appropriately. Addressing these challenges is a necessary precursor to the core challenge of achieving population impact with evidence-based interventions.

6.3 Chapter summary

- When an intervention is deemed ready for scale-up and dissemination, IPE may be used to understand the factors that influence intervention engagement and reach, implementation quality, sustainability over time, and how and why the intervention evolves as it goes to scale.
- Various theories and frameworks may be utilised in IPE at the scale-up and dissemination phase, including Rogers' (2003) diffusion of innovations theory, and Glasgow, Vogt and Boles' (1999) RE-AIM framework.
- IPE can be used to inform Type II translational research, through which we may improve our understanding of how to improve the process by which we bring effective interventions from 'science to service'.

7 Conclusion

This research was commissioned in order to draw together existing knowledge regarding the aims, functions and methods of implementation and process evaluation (IPE) in relation to educational interventions, with a view to subsequently informing guidance for researchers akin to that already produced on the design and conduct of randomised controlled trials (RCTs) (e.g. Torgerson & Torgerson, 2013). This guidance has subsequently been written by the authors and is available here.

Through a process of thematic review involving 12 focused literature searches on a different topics (e.g. logic models and theories of change; adaptation; factors affecting implementation), we found compelling evidence of the utility of IPE at all stages of intervention development and evaluation. At the pilot stage, IPE can aid the process of articulating the nature of a given intervention and theorising the mechanisms through which it may achieve impact on outcomes. Furthermore, it can help to determine the social validity (e.g. acceptability, feasibility and utility) of the intervention. At the trial stage, IPE can be used to document implementation, assess usual practice, examine the adaptation process, and establish differential intervention benefits among subgroups of the school population. Additionally, in the context of efficacy trials, it may be used to examine the relationship between on-treatment status and outcomes, and empirically validate the intervention theory. At the effectiveness trial stage, IPE may serve to enhance our understanding of contextual influences, and identify the 'active ingredients' of the intervention. Finally, at the scale up and dissemination stage, it can be used to understand the factors that influence intervention engagement and reach, implementation quality, and sustainability over time. All of the above functions are ultimately about enhancing our understanding of how interventions create change. IPE thus enables us to 'get inside the black box' (Saunders, Evans & Joshi, 2005) and as such is a vital accompaniment to the RCT evaluation model.

8 References

American Psychological Association. (1996). Evidence based practice in psychology. *American Psychologist*, *61*, 271–285.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 445-455.

Askell-Williams, H., Dix, K. L., Lawson, M. J., & Slee, P. T. (2013). Quality of implementation of a school mental health initiative and changes over time in students' social and emotional competencies. *School Effectiveness and School Improvement*, *24*, 357–381.

Askell-Williams, H., Slee, P. T., & Van Deur, P. (2013). Social and emotional well-being programme: The nexus between sustainability and quality. *Psychology of Education Review*, *37*, 48–56.

Atkins, S., Odendall, W., Leon, N., Lutge, E., & Lewin, S. (2015). Qualitative process evaluation for complex interventions. In D. Richards & I. Hallberg (Eds.), *Complex interventions in health: An overview of research methods* (pp. 239–247). New York, NY: Routledge.

August, G. J., Gewirtz, A., & Realmuto, G. M. (2010). Moving the field of prevention from science to service: Integrating evidence-based preventive interventions into community practice through adapted and adaptive models. *Applied and Preventive Psychology*, *14*, 72–85.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Banerjee, R., Weare, K., & Farr, W. (2014). Working with "Social and Emotional Aspects of Learning" (SEAL): Associations with school ethos, pupil social experiences, attendance, and attainment. *British Educational Research Journal*, *40*, 718–742.

Barlow, A., Humphrey, N., Lendrum, A., Wigelsworth, M., & Squires, G. (2015). Evaluation of the implementation and impact of an integrated prevention model on the academic progress of students with disabilities. *Research in Developmental Disabilities*, *36*, 505–525.

Barrish, H. H., Saunders, M., & Wolf, M. M. (1969). Good behavior game: effects of individual contingencies for group consequences on disruptive behavior in a classroom. *Journal of Applied Behavior Analysis*, *2*, 119–24.

Basch, C. E., Sliepcevich, E. M., Gold, R. S., Duncan, D. F., & Kolbe, L. J. (1985). Avoiding type III errors in health education program evaluations: a case study. *Health Education Quarterly*, *12*, 315–31.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: an integrated model of program implementation. *Prevention Science*, *12*, 23–

Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R., Tobin, K., Jones, C. & Bywater, T. (2015). The effectiveness and micro-costing analysis of a universal, school-based, social-emotional learning programme in the UK: A cluster randomised controlled trial. *School Mental Health, Online First*.

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

Bishop, D. C., Pankratz, M. M., Hansen, W. B., Albritton, J., Albritton, L., & Strack, J. (2014). Measuring fidelity and adaptation: Reliability of a instrument for school-based prevention programs. *Evaluation and the Health Professions*, *37*, 231–57.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*, 253–268.

Blamey, A., & Mackenzie, M. (2007). Theories of change and realistic evaluation: peas in a pod or apples and oranges? *Evaluation*, *13*, 439–455.

Blase, K., & Fixsen, D. (2013). Core intervention components: Identifying and operationalizing what makes programs work. *ASPE Research Brief*, *2*, 1–21.

Bloomquist, M. L., August, G. J., & Ostrander, R. (1991). Effects of a school-based cognitive-behavioural intervention for ADHD children. *Journal of Abnormal Child Psychology*, *19*, 591–605.

Bluth, K., Campo, R. A., Pruteanu-Malinici, S., Reams, A., Mullarkey, M., & Broderick, P. C. (2015). A school-based mindfulness pilot study for ethnically diverse at-risk adolescents. *Mindfulness*, Online First.

Bonell, C., Fletcher, A., Morton, M., Lorenc, T., & Moore, L. (2012). Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Social Science & Medicine*, *75*, 2299–306.

Bopp, M., Saunders, R. P., & Lattimore, D. (2013). The tug-of-war: fidelity versus adaptation throughout the health promotion program life cycle. *The Journal of Primary Prevention*, *34*, 193–207.

Borglin, G. (2015). The value of mixed methods for researching complex interventions. In D. A. Richards & I. R. Hallberg (Eds.), *Complex interventions in health: An overview of research methods* (pp. 29–45). New York, NY: Routledge.

Bronfenbrenner, U. (2005). *Making human beings human: bioecological perspectives on human development*. London: Sage Publications.

Bruhn, A. L., Hirsch, S. E., & Lloyd, J. W. (2015). Treatment integrity in school-wide programs: A review of the literature (1993–2012). *The Journal of Primary Prevention*, *36*, 335–349.

Campbell, M., & Fitzpatrick, R. (2000). Framework for design and evaluation of complex interventions to improve health. *BMJ, 321,* 694-696.

Cappella, E., Hamre, B., Kim, H., Henry, D. B., Frazier, S. L., Atkins, M., & Schoenwald, S. K. (2012). Teacher consultation and coaching within mental health practice: Classroom and child effects in urban elementary schools. *Journal of Consulting and Clinical Psychology*, *80*, 597–610.

Carless, D. R. (2003). Factors in the implementation of task-based teaching in primary schools. *System*, *31*, 485–500.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*, 40–49.

Centers for Disease Control and Prevention. (1999). *Framework for program evaluation in public health*. Atlanta, GA: CDCP.

Chan, G., Foxcroft, D., Smurthwaite, B., Coombes, L., & Allen, D. (2012). *Improving child behaviour management: An evaluation of the Good Behaviour Game in UK primary schools*. Oxford: Oxford Brooked University.

Chaudoir, S. R., Dugan, A. G., & Barr, C. H. I. (2013). Measuring factors affecting implementation of health innovations: a systematic review of structural, organizational, provider, patient, and innovation level measures. *Implementation Science*, *8*, 22-42.

Child Poverty Action Group. (2015). Child poverty facts and figures. Retrieved from http://www.cpag.org.uk/child-poverty-facts-and-figures, August 2015.

Chittleborough, C. R., Nicholson, A. L., Young, E., Bell, S., & Campbell, R. (2013). Implementation of an educational intervention to improve hand washing in primary schools: process evaluation within a randomised controlled trial. *BMC Public Health*, *13*, 757.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Collins, L.M., Barker, T. B., Mermelstein, R. J., Piper, M. E., Jorenby, D. E., Smith, S.S., Christansen, B. A., Schlam, T. R., Cook, J. W. & Fiore, M. C. (2011). The multiphase optimization strategy for engineering effective tobacco use interventions. *Annals of Behavioral Medicine*, *41*, 208-226.

Connell, J., & Klem, A. (2000). You can get there from here: Using a theory of change approach to plan urban education reform. *Journal of Educational and Psychological Consultation*, *11*, 93–120.

Connell, J. P., & Kibusch, A. C. (1998). Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects and problems. In Fulbright-Andersen, K., Kubisch, A. C., & connell, J. P. (Eds.), *New approaches to evaluating community initiatives. Volume 2: theory, measurement and analysis*. Queenstown: The Aspen Institute.

Conrad, K. M., & Conrad, K. J. (2005). Compensatory rivalry. In B. S. Everitt & D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. New York, NY: Wiley and Sons.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). *Developing and evaluating complex interventions: New guidance*. London: MRC.

Creswell, J. A., & Piano Clark, V. L. (2011). *Designing and conducting mixed methods research*. London: Sage.
Damschroder, L. J., & Hagedorn, H. J. (2011). A guiding framework and approach for implementation research in substance use disorders treatment. *Psychology of Addictive Behaviors : Journal of the Society of Psychologists in Addictive Behaviors*, *25*, 194–205.
Dane, A. V, & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45.
de Visser, R. O., Graber, R., Hart, A., Abraham, C., Scanlon, T., Watten, P., & Memon, A. (2015). Using qualitative methods within a mixed-methods approach to developing and evaluating interventions to address harmful alcohol use among young people. *Health Psychology, 34*, 349–60.
Denton, C. A., Vaughn, S., & Fletcher, J. M. (2003). Bringing research-based practice in reading intervention to scale. *Learning Disabilities Research & Practice*, *18*, 201–211.
Department for Children Schools and Families. (2009a). *Achievement for All: Guidance for schools*. Nottingham: DCSF.
Department for Children Schools and Families. (2009b). *Getting to grips with Assessing Pupils' Progress*. Nottingham: DCSF.
Derzon, J. H., Sale, E., Springer, J. F., & Brounstein, P. (2005). Estimating intervention effectiveness: synthetic projection of field evaluation results. *The Journal of Primary Prevention*, *26*, 321–43.
Dix, K. L., Keeves, J. P., Slee, P., Lawson, M. J., Russell, A., Askell-williams, H., … Spears, B. A. (2010). *Kidsmatter primary evaluation: technical report and user guide*. Adelaide: Flinders University.
Dix, K. L., Slee, P. T., Lawson, M. J., & Keeves, J. P. (2012). Implementation quality of whole-school mental health promotion and students' academic performance. *Child and Adolescent Mental Health*, *17*, 45–51.
Domitrovich, C. (2008). Maximizing the implementation quality of evidence-based preventive interventions in schools: A conceptual framework. *Advances in School Mental Health Promotion, 1*, 6–28.
Domitrovich, C. E., Bradhsaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., … Ialongo, N. S. (2008). Maximising the implementation quality of evidence-based preventive interventions in schools: a conceptual framework. *Advances in School Mental Health Promotion*, *1*(3), 6–28.
Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & Sanford Derousie, R. M. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, *25*, 284–298.
Domitrovich, C., & Greenberg, M. T. (2000). The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation*, *11*, 193–221.
Durlak, J. (2010). The importance of doing well in whatever you do: a commentary on the special edition, "Implementation research in early childhood education." *Early Childhood Research Quarterly*, *25*, 348–357.
Durlak, J. a, Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: a meta-analysis of school-based universal interventions. *Child Development*, *82*, 405–32.
Durlak, J. A. (1995). *School-based prevention programs for children and adolescents*. Thouand Oaks, California: Sage.
Durlak, J. A. (2015). Studying program implementation is not easy but it is essential. *Prevention Science*. Online First.
Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–50.
Dusenbury, L. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237–256.
Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*, 237–56.
Elliott, D., & Mihalic, S. (2004). Issues in disseminating and replicating effective

prevention programs. *Prevention Science*, *5*, 47–53.
Elswick, S., & Casey, L. (2011). The Good Behavior Game is no longer just an effective intervention for students: An examination of the reciprocal effects on teacher behaviors. *Beyond Behavior*, *21*, 36–46.
Evans, R., Murphy, S., & Scourfield, J. (2015). Implementation of a school-based social and emotional learning intervention: Understanding diffusion processes within complex systems. *Prevention Science*, *16*, 754–764.
Evans, R., Scourfield, J., & Murphy, S. (2014). The unintended consequences of targeting: young people's lived experiences of social and emotional learning interventions. *British Educational Research Journal*, *41*, 381-397.
Evans, R., Scourfield, J., & Murphy, S. (2015). Pragmatic, formative process evaluations of complex interventions and why we need more of them. *Journal of Epidemiology and Community Health*, *69*, 925–6.
Feeley, N., Cossette, S., Cote, J., Heon, M., Stremler, R., Martorella, G., & Purden, M. (2009). The importance of piloting an RCT intervention. *Canadian Journal of Nursing Research*, *41*, 84–99.
Fixsen, D., Naoom, S. F., Blase, K. a, Friedman, R. M., & Wallace, F. (2005). *Implementation research: a synthesis of the literature*. Tampa: University of South Florida.
Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of Evidence: Criteria for Efficacy, Effectiveness and Dissemination. *Prevention Science*, *6*, 151–175.
Forman, S. G. (2015). *Implementation of mental health programs in schools: a change agent's guide*. Washington, DC: American Psychological Association.
Forman, S. G., Shapiro, E. S., Codding, R. S., Gonzales, J. E., Reddy, L. A., Rosenfield, S. A., … Stoiber, K. C. (2013). Implementation science and school psychology. *School Psychology Quarterly*, *28*, 77–100.
Forman, S., Olin, S., Hoagwood, K., & Crowe, M. (2009). Evidence-based interventions in schools: developers' views of implementation barriers and facilitators. *School Mental Health*, *1*, 26–36.
Foxcroft, D. R. (2014). Can prevention classification be improved by considering the function of prevention? *Prevention Science,15*, 818–22.
Fraser, M. W., & Galinsky, M. J. (2010). Steps in intervention research: Designing and developing social programs. *Research on Social Work Practice*, *20*, 459–466.
Fraser, M. W., Richman, J. M., Galinsky, M. J., & Day, S. H. (2009). *Intervention research: developing social programs*. Oxford: Oxford University Press.
Glasgow, R. E., Lichtenstein, E., & Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health*, *93*, 1261–1267.
Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health*, *89*, 1322–7.
Gneezy, U., Leonard, K. L., & List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, *77*, 1637–1664.
Goldberg Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: agreement and relation to youth outcomes. *Health Education & Behavior, 31*, 242–57.
Gorard, S., Siddiqui, N., & Huat See, B. (2015). *Philosophy for Children: Evaluation report and executive summary*. London: EEF.
Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science, 16*, 893-896.
Gottfredson, D. C., & Gottfredson, G. D. (2002). Quality of school-based prevention programs: Results from a national survey. *Journal of Research in Crime and Delinquency*, *39*, 3–35.
Greenberg, M., Domitrovich, C., Graczyk, P., Zins, J., & Services, C. for M. H. (2005). *The study of implementation in school-based preventive interventions: Theory, research, and practise.* Rockville: CMHS.
Greenberg, M. T. (2010). School-based prevention: current status and future challenges.

Effective Education, *2*, 27–52.
Gresham, F. (2009). Evolution of the treatment integrity concept: Current status and future directions. *School Psychology Review*, *38*, 533–540.
Han, S. S., & Weiss, B. (2005). Sustainability of teacher implementation of school-based mental health programs. *Journal of Abnormal Child Psychology*, *33*, 665–679.
Hansen, W. (2014). Measuring fidelity. In Z. Sloboda & H. Petras (Eds.), *Defining prevention science* (pp. 335–359). New York, NY: Springer.
Hansen, W. B., Pankrantz, M. M., Dusenbury, L., Giles, S. M., Bishop, D., Albritton, J., … Strack, J. (2013). Styles of adapation: the impact of frequency and valence of adaptation on preventing substance abuse. *Health Education*, *113*, 345–363.
Hansen, W. B., Pankratz, M. M., & Bishop, D. C. (2014). Differences in observers' and teachers' fidelity assessments. *Journal of Primary Prevention*, *35*, 297–308.
Hattie, J. (2009). *Visible learning*. London: Routledge.
Hattie, J. (2012). *Visible learning for teachers: maximising impact on learning*. London: Routledge.
Haynes, B. (1999). Can it work? Does it work? Is it worth it? *BMJ*, *319*, 652–653.
Haynes, L., Service, O. Goldacre, B. & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. London: Cabinet Office
Henwood, K. L., & Pidgeon, N. F. (1992). Qualitative research and psychological theorising. *British Journal of Psychology*, *83*, 97–111.
Hoffmann, T. C., Erueti, C., & Glasziou, P. P. (2013). Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials. *BMJ (Clinical Research Ed.)*, *347*, 3755-3765.
Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., … Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Clinical Research Ed.)*, *348*, 1687-1699.
Humphrey, N. (2013). *Social and emotional learning: a critical appraisal*. London: Sage Publications.
Humphrey, N., Barlow, A., Wigelsworth, M., Lendrum, A., Pert, K., Joyce, C., … Turner, A. (2015). *Promoting Alternative Thinking Strategies (PATHS): Evaluation report*. London: EEF.
Humphrey, N., Kalambouka, A., Bolton, J., Lendrum, A., Wigelsworth, M., Lennie, C., & Farrell, P. (2008). *Primary social and emotional aspects of learning: evaluation of small group work*. Nottingham: DCSF Publications.
Humphrey, N., Lendrum, A., Wigelsworth, M., & Kalambouka, A. (2009). Implementation of primary social and emotional aspects of learning (SEAL) small group work: a qualitative study. *Pastoral Care in Education*, *27*, 219–239.
Humphrey, N., & Squires, G. (2011). *Achievement for All national evaluation: final report*. Nottingham: DFE.
Hurley, J. J. (2012). Social validity assessment in social competence interventions for preschool children: A rReview. *Topics in Early Childhood Special Education*, *32*, 164–174.
Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus Bullying Prevention Program: A multilevel study. *Prevention and Treatment*, *6*, 1–29.
Kavanagh, J., Oliver, S., & Lorenc, T. (2008). Reflections on developing and using PROGRESS-Plus. *Equity Update*, *2*, 1–2.
Kellam, S. G., Mackenzie, A. C. L., Brown, C. H., Poduska, J. M., Wang, W., Petras, H., & Wilcox, H. C. (2011). The Good Behavior Game and the future of prevention and treatment. *Addiction Science & Clinical Practice*, *6*, 73–84.
Kellam, S. G., Rebok, G. W., Ialongo, N., & Mayer, L. S. (1994). The course and malleability of aggressive behavior from early first grade into middle school: results of a developmental epidemiologically-based preventive trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *35*, 259–81.
Kendal, S., Callery, P., & Keeley, P. (2011). The feasibility and acceptability of an approach to emotional wellbeing support for high school students. *Child and Adolescent Mental Health*, *16*, 193–200.
Lamb, B. (2009). *Lamb inquiry - special educational needs and parental confidence*. Nottingham: DCSF Publications.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*, 635–652.
Lendrum, A., Humphrey, N., & Greenberg, M. T. (2016). Implementing for success in school-based mental health promotion: The role of quality in resolving the tension between fidelity and adaptation. In R. Shute & P. Slee (Eds.), *Mental health and wellbeing through schools: The way forward.* London: Taylor and Francis.
Little, M., Sussman, S., Sun, P., & Rohrbach, L. (2013). The effects of implementation fidelity in the Towards No Druge Abuse dissemination trial. *Health Education*, *113*, 281-296.
Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, *21*, 121–145.
Liu, J. (2010). Minimum effective dose. In *Encyclopedia of Biopharmaceutical Statistics* (pp. 799–800). London: Informa.
Luiselli, J. K., & Reed, D. D. (2011). Social valdity. In *Encylopedia of Child Behavior and Development*. New York: Springer.
Maldonado, G., & Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology*, *31*, 422–9.
Maughan, B. (2013). Editorial: "Better by design" - why randomized controlled trials are the building blocks of evidence-based practice. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *54*, 225–6.
May, C., & Finch, T. (2009). Implementing, embedding, and integrating practices: An outline of Normalization Process Theory. *Sociology*, *43*, 535–554.
McDonald, S.-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, *35*, 15–24.
Mclaughlin, J., & Jordan, G. (1999). Logic models: a tool for telling your program's performance story. *Evaluation and Program Planning*, *22*, 65–72.
McMillan, J. H. (2007). Randomised field trials and internal validity: Not so fast, my friend. *Practical Assessment, Research and Evaluation*, *12*, 1–6.
Mishna, F., Muskat, B., & Wiener, J. (2010). "I'm not lazy; it's just that I learn differently": Development and implementation of a manualized school-based group for students with learning disabilities. *Social Work With Groups*, *33*, 139–159.
Moore, G., Audrey, S., Barker, M., Bond, L., Bonell, C., Cooper, C., … Baird, J. (2014). Process evaluation in complex public health intervention studies: the need for guidance. *Journal of Epidemiology and Community Health*, *68*, 101–2.
Moore, G. F., Audrey, S., Barker, M., Bond, L., Bonell, C., Hardeman, W., … Baird, J. (2015). Process evaluation of complex interventions: Medical Research Council guidance. *BMJ (Clinical Research Ed.)*, *350*, 1258-1265.
Moore, J. E., Bumbarger, B. K., & Cooper, B. R. (2013). Examining adaptations of evidence-based programs in natural contexts. *The Journal of Primary Prevention*, *34*, 147–61.
Mowbray, C. T. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340.
Naylor, P. J., Nettlefold, L., Race, D., Hoy, C., Ashe, M. C., Wharf Higgins, J., & McKay, H. a. (2015). Implementation of school based physical activity interventions: A systematic review. *Preventive Medicine*, *72*, 95–115.
Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research*, *39*, 374–396.
Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science*, *10*(1), 53-66.
O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*, 33–84.
Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *BMJ (Clinical Research Ed.)*, *332*, 413–6.
Odom, S. L., Fleming, K., Diamond, K., Lieber, J., Hanson, M., Butera, G., … Marquis, J. (2010). Examining different forms of implementation in early childhood curriculum

research. *Early Childhood Research Quarterly*, *25*, 314–328.
Ogden, T., & Fixsen, D. L. (2014). Implementation science: A brief overview and a look ahead. *Zeitschrift Für Psychologie*, *222*, 4–11.
Olweus, D., & Limber, S. P. (2010). Bullying in school: Evaluation and dissemination of the olweus bullying prevention program. *American Journal of Orthopsychiatry*, *80*, 124–134.
Onwuegbuzie, A. J., & Leech, N. L. (2006). Linking research questions to mixed methods data analysis procedures. *The Qualitative Report*, *11*, 474–498.
Ozer, E. J. (2006). Contextual effects in school-based violence prevention programs: A conceptual framework and empirical review. *The Journal of Primary Prevention*, *27*, 315–340.
Palinkas, L. A., Aarons, G. A., Horwitz, S., Chamberlain, P., Hurlburt, M., & Landsverk, J. (2011). Mixed-method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 44–53.
Pas, E. T., Waasdorp, T. E., & Bradshaw, C. P. (2015). Examining contextual influences on classroom-based implementation of Positive Behavior Support Strategies: Findings from a randomized controlled effectiveness trial. *Prevention Science*, *16*, 1096-1106
Pawson, R., & Tilley, N. (2004). Realistic evaluation. In S. Matthieson (Ed.), *Encyclopedia of Evaluation* (pp. 359–367). Thousand Oaks, CA: Sage Publications.
Payne, A. A., Gottfredson, D. C., & Gottfredson, G. D. (2006). School predictors of the intensity of implementation of school-based prevention programs: results from a national study. *Prevention Science*, *7*, 225–37.
Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., & Welch, V. (2012). Damned if you do, damned if you don't: subgroup analysis and equity. *Journal of Epidemiology and Community Health*, *66*, 95–8.
Powers, J. D. (2010). Ecological risk and resilience perspective: a theoretical framework supporting evidence-based practice in schools. *Journal of Evidence-Based Social Work*, *7*, 443–51.
Proctor, E., Silmere, H., Raghavan, R., Hovmand, P., Aarons, G., Bunger, A., … Hensley, M. (2011). Outcomes for implementation research: Conceptual distinctions, measurement challenges, and research agenda. *Administration and Policy in Mental Health and Mental Health Services Research*, *38*, 65–76.
Rabin, B. A., Brownson, R. C., Haire-Joshu, D., Kreuter, M. W., & Weaver, N. L. (2008). A glossary for dissemination and implementation research in health. *Journal of Public Health Management and Practice*, *14*, 117–123.
Raffo, C., Dyson, A., Gunter, H., Hall, D., Jones, L., & Kalambouka, A. (2009). Education and poverty: mapping the terrain and making the links to educational policy. *International Journal of Inclusive Education*, *13*, 341–358.
Raffo, C., Dyson, A., Gunter, H., Hall, D., Jones, L., & Kalambouka, A. (2010). *Education and poverty in affluent countries*. London: Routledge.
Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, a, Baranowski, T., Baranowski, J., … Wang, D. T. (1998). How best to measure implementation of school health curricula: a comparison of three measures. *Health Education Research*, *13*, 239–50.
Reupert, A., & Woodcock, S. (2010). Success and near misses: Pre-service teachers' use, confidence and success in various classroom management strategies. *Teaching and Teacher Education*, *26*, 1261–1268.
Rogers, E. M. (2003). *Diffusion of innovations*. New York, NY: Free Press.
Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. *Evaluation*, *14*, 29–48.
Rohrbach, L. A. (2014). Design of prevention interventions. In Sloboda, Zili & H. Petras (Eds.), *Defining prevention science* (pp. 275–292). New York, NY: Springer.
Roland, M., & Torgerson, D. J. (1998). Understanding controlled trials: What are pragmatic trials? *BMJ*, *316*, 285–285.
Rosenblatt, J. L., & Elias, M. J. (2008). Dosage effects of a preventive social-emotional learning intervention on achievement loss associated with middle school transition. *The Journal of Primary Prevention*, *29*, 535–555.
Rothman, J., & Thomas, E. J. (1994). *Intervention research: design and development for human services*. New York, NY: Haworth Press.
Rubenstein, L. V, & Pugh, J. (2006). Strategies for promoting organizational and

practice change by advancing implementation research. *Journal of General Internal Medicine*, *21 Suppl 2*, S58–64.
Sanetti, L. M., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review*, *40*, 72–84.
Saunders, R. P. (2016). *Implementation monitoring and process evaluation*. London: Sage Publications.
Saunders, R. P., Evans, M. H., & Joshi, P. (2005). Developing a process-evaluation plan for assessing health promotion program implementation: a how-to guide. *Health Promotion Practice*, *6*, 134–47.
Schoenwald, S. K., & Garland, A. F. (2014). A review of treatment adherence measurement methods. *Psychological Assessment*, *25*, 146–156.
Schonert-Reichl, K. A., Smith, V., Zaidman-Zait, A., & Hertzman, C. (2012). Promoting children's prosocial behaviors in school: Impact of the "Roots of Empathy" program on the social and emotional competence of school-aged children. *School Mental Health*, *4*, 1–21.
Sedgwick, P. (2015). Intention to treat analysis versus per protocol analysis of trial data. *BMJ (Clinical Research Ed.)*, *350*, 681-683.
Slavin, R. E. (2012). Foreword. In B. Kelly & D. F. Perkins (Eds.), *Handbook of implementation science for psychology in education* (p. xv). Cambridge: Cambridge University Press.
Slee, P., Lawson, M., Russell, A., Askell-Williams, H., Dix, K., Owens, L., … Spears, B. (2009). *Kidsmatter primary evaluation final report*. Adelaide: Flinders University.
Smith, J., Schneider, B., & Smith, P. (2004). The effectiveness of whole-school antibullying programs: A synthesis of evaluation research. *School Psychology Review*, *33*, 547–560.
Social and Character Development Research Consortium. (2010). *Efficacy of school-wide programs to promote social and character development and reduce problem behavior in elementary school children*. Washington, DC: SCDRC.
Song, M.K., & Ward, S. E. (2015). Assessment effects in educational and psychosocial intervention trials: an important but often-overlooked problem. *Research in Nursing & Health*, *38*, 241–7.
Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Quality in qualitative evaluation: A framework for assessing research evidence*. London: Cabinet Office.
Spoth, R., Rohrbach, L. a, Greenberg, M., Leaf, P., Brown, C. H., Fagan, A., … Hawkins, J. D. (2013). Addressing core challenges for the next generation of type 2 translation research and systems: the translation science to population impact (TSci Impact) framework. *Prevention Science, 14*, 319–51.
Stallard, P., Skryabina, E., Taylor, G., Phillips, R., Daniels, H., Anderson, R., & Simpson, N. (2014). Classroom-based cognitive behaviour therapy (FRIENDS): A cluster randomised controlled trial to Prevent Anxiety in Children through Education in Schools (PACES). *The Lancet Psychiatry*, *1*, 185-192.
Steckler, A., & Linnan, L. (2002). *Process evaluation for public health interventions and research*. San Francisco: Jossey-Bass.
Stewart-Brown, S., Anthony, R., Wilson, L., Winstanley, S., Stallard, N., Snooks, H. & Simkiss, D. (2011). Should randomised controlled trials be the 'gold standard' for research on preventive interventions for children? *Journal of Children's Services*, *6*, 228-235.
Stinchcomb, J. (2001). Using logic modeling to focus evaluation efforts. *Journal of Offender Rehabilitation*, *33*, 47–56.
Sun, X., Briel, M., Walter, S. D., & Guyatt, G. H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ (Clinical Research Ed.)*, *340*, 117.
Tingstrom, D. H., Sterling-Turner, H. E., & Wilczynski, S. M. (2006). The Good Behavior Game: 1969-2002. *Behavior Modification*, *30*, 225–53.
Torgerson, C. J., & Torgerson, D. J. (2013). *Randomised trials in education: an introductory handbook*. London: EEF.
Torgerson, C., & Torgerson, D. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies, 49*, 24–26.
Tymms, P., Merrell, C., & Coe, R. (2008). Educational policies and randomised

controlled trials. *The Psychology of Education Review*, *32*, 3–7 & 26–29.
US Department of Health and Human Services. (2002). *Finding the balance: program fidelity and adaptation in substance abuse prevention*. Washington, DC: SAMHSA.
Walker, H. M. (2004). Commentary: Use of evidence-based intervention in schools: Where we've been, where we are, and where we need to go. *School Psychology Review*, *33*, 398–407.
Wehby, J. H., Maggin, D. M., Moore Partin, T. C., & Robertson, R. (2012). The impact of working alliance, social validity, and teacher burnout on implementation fidelity of the Good Behavior Game. *School Mental Health*, *4*, 22–33.
Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., Ten-Bokkel, I., Tate, K., & Emery, C. (under review). The influence of trial stage, developer involvement and international transferability on the outcomes of universal social and emotional learning outcomes: A meta-analysis. *Cambridge Journal of Education*.
Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: update of a meta-analysis. *American Journal of Preventive Medicine*, *33*(2 Suppl), S130–43.
Wolf, M. M. (1978). Social validity: the case for subjective measurement. *Journal of Applied Behavior Analysis*, *11*, 203–214.
Woolf, S. H. (2008). The meaning of translational research and why it matters. *Journal of the American Medical Association*, *299*, 211–213.
Wright, M., Masten, A. S., & Narayan, A. (2013). Resilience processes in development: Four waves of research on positive adaptation in the context of adversity. In S. Goldstein & R. B. Brooks (Eds.), *Handbook of resilience in children* (pp. 15–37). New York, NY: Springer.
Yampolskaya, S., Nesman, T. M., Hernandez, M., & Koch, D. (2004). Using concept mapping to develop a logic model and articulate a program theory: A case example. *American Journal of Evaluation*, *25*, 191–207.
Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology*, *49*, 156–67.