Is implementation of evidence-based interventions in schools related to pupil outcomes? A systematic review

Paul Killerby and Sandra Dunsmuir

Dr Paul Killerby, Educational Psychologist, Achieving for Children
Achieving for Children EPS, The Moor Lane Centre, Chessington KT9 2AA
paul.killerby@achievingforchildren.org.uk

Prof Sandra Dunsmuir, Director, Educational Psychology Group,
University College London, 26 Bedford Way, London WC1H 0AP
s.dunsmuir@ucl.ac.uk

Is implementation of evidence-based interventions in schools related to pupil outcomes? A systematic review

Aims: The growing influence of implementation science has resulted in educational researchers exploring what occurs within schools to support intervention effectiveness. This paper provides an overview of existing research so that practitioners can understand the extent to which measures of implementation are associated with the outcomes of school-based interventions.

Method: This paper systematically identified studies which correlated or directly compared the implementation of school-based interventions with pupil outcomes. Effect-sizes are reported and the strength of evidence appraised using a weight-of-evidence framework.

Findings: The 13 studies reviewed reported 32 quantified effect sizes which represented the strength and direction of the relationship between measures of implementation and intervention outcomes in schools. The review also identified gaps in current evidence which have implications for further research and practice.

Limitations: This review did not explore factors which supported staff to implement interventions effectively. As such, this review focusses on the effects of implementation, rather than detailed practices.

Conclusions: This review found that educational researchers rarely measured fidelity of programme implementation. When fidelity is measured, there are indications that proper execution and co-ordination of evidence-based interventions is positively related to pupil outcomes. However, the measurement of implementation fidelity can be undermined when

data is transformed into arbitrary categories, such as ‘good’ and ‘bad’. The practicalities of effectively transporting evidence-based interventions into school settings are discussed.

Keywords: Implementation; schools; evidence-based; intervention; outcomes

Educational psychologists (EPs) have traditionally observed how the medical and allied professions have responded to new scientific movements before considering how these ideas will be applied within EP practice (Kratochwill & Stoiber, 2002). The application of implementation science to educational psychology has followed this trend. Implementation science has received increased attention in the wake of the perceived failure of the evidence-based practice movement to produce consistent results in real world settings – the oft-cited research-practice gap (Ogden & Fixsen, 2015). This resulted in some researchers calling for an increased focus upon intervention effectiveness rather than intervention efficacy (Kessler & Glasgow, 2011). Efficacy is the investigation of beneficial effects under controlled conditions, whereas effectiveness considers beneficial effects under ‘real world’ conditions (Flay et al., 2005). This is an important distinction for EPs, who are well-placed to investigate how practitioners can effectively transport evidence-based interventions (EBIs) into a range of school settings so that all children can access effective mental health and special educational needs (SEN) provision.

Clinical psychologists have attempted to bridge the research-practice gap by evaluating the real world effectiveness of EBIs in school settings (Schaeffer et al., 2005). More recently, educational researchers have started to explore school-based implementation effects for EBIs (Sutherland, Conroy, Vo, & Ladwig, 2015). These studies have considered how variability across the eight implementation dimensions outlined by Durlak and Dupre (2008; see Table

1) influence school-based intervention outcomes. Implementation dimensions can vary with regard to programme design and practical delivery and across different intervention groups.

Table 1
Implementation Dimensions

Implementation Dimensions (from Durlak and Dupre, 2008)	Definition
Fidelity	The degree to which the actual intervention delivered resembles the original programme content.
Dosage	The extent to which the total intervention content is delivered to participants.
Quality	How effectively an intervention has been delivered to participants.
Participant Responsiveness	The extent to which an intervention sustains the attention of participants and promotes active engagement.
Program Differentiation	The extent to which the content of an intervention can be discriminated from other interventions.
Monitoring of control/comparison conditions	The extent to which those not accessing an intervention are accessing alternative provisions or services.
Program Reach	The extent to which an intervention is delivered to its target population.
Adaptation	Alterations made to an intervention to ensure effective delivery.

There has been increasing interest in the delivery of effective school-based EBIs in educational research and policy making contexts in the UK and beyond (Fox, 2003). However, there is growing criticism that a historical focus upon efficacy has neglected the opportunity to understand what practitioners need to do to deliver effective interventions (Biesta, 2007; Chorpita et al., 2011). Schools have thus faced a research-practice gap, dividing researchers prioritising EBI efficacy and practitioners prioritising EBI effectiveness. Educational psychologists are capable of closing this gap by using implementation science to explore EBI delivery in schools, through questioning the extent to which there is adherence to both the proper execution of specific intervention elements and whether practices are co-ordinated effectively. Despite increased attention to this issue, EPs do not currently have

access to systematic reviews which compare naturally occurring EBI implementation variance with pupil outcomes in schools.

The first question for this review therefore examined whether there is evidence that implementation of EBIs is positively associated with EBI pupil outcome measures. The second question considered how school-based research has measured and analysed EBI implementation in schools, including the choice of measurements and methodology used, the clarity of reporting and statistical treatment of data.

Method

Selection of studies

A Boolean search of three databases (PsychINFO, ERIC and Web of Science) was undertaken in August 2014 using the following keyword search criteria:

- (meas* adj implemen*) OR (meas* adj fidelity) OR (meas* adj dosage) OR (meas* adj quality) OR (meas* adj dissemination) OR (meas* adj integrity) OR (meas* adj adherence) OR (meas* adj accept*) OR (commitment adj2 intervention)

AND

- school

AND

- (evidence-based prog*) OR (evidence-based treat*) OR (evidence-based strat*) OR (evidence-based interven*) OR (empirically supported treat*) OR (empirically based treat*)

Studies that were published before the year 2000 were excluded from the review. This was due to legislative changes influencing the use of EBIs in schools, including the 2001 National Curriculum in the UK and the 2000 No Child Left Behind Act in the USA. The review was limited to peer-reviewed journals published in English. Ancestral searches of the identified studies were conducted using the existing inclusion criteria.

Inclusion and exclusion criteria

To be included a study had to meet three criteria:
First, it needed to include an EBI in a school setting delivered to school-age pupils (ages 4-18). Second, it needed to provide a continuous measure of an EBI implementation dimension(s) (see Table 1) and a continuous measure of a child outcome(s). Third, it had to directly compare measures of EBI implementation with measures of pupil outcomes through correlation or group comparisons.

The search yielded 695 studies including duplicates. An initial screening of titles and abstracts was carried out and the majority were excluded for failing to meet the inclusion criteria. The remaining 51 studies were examined and checked against the inclusion/exclusion criteria, after which a further 41 were excluded. Ancestral searches of the 10 identified studies identified a further 3 papers which met the existing inclusion criteria, thus a total of 13 studies were included in this review.

Coding of the studies

Each study was weighted according to an adaptation of Gough's Weight of Evidence (WoE) Framework (Gough, 2007). This stipulates that weight of evidence is considered according to four criteria:

- WoE A, methodological quality: generally accepted criteria for evaluating evidence by those who use and produce it.
- WoE B, methodological relevance: review-specific judgement about the appropriateness of the design and measures to answer the review question.
- WoE C, relevance of evidence: review-specific judgement about the relevance of evidence to the review question.
- WoE D, overall weighting: combination of WoE A, B and C.

As the literature search identified studies using correlational and single participant designs, WoE A was determined using recommendations specific to these designs: correlational studies and single participant studies were appraised using criteria from Thompson, Diamond, McWilliam, Snyder, & Snyder, (2005) and Horner et al., (2005) respectively. WoE B considered the reliability of measures, while WoE C addressed the ecological validity of implementation measurement. Ecological validity is the extent to which the findings of a study can be generalised to real world settings. Therefore studies which controlled for self-report bias or observer effects were positively weighted in WoE C. Results are summarised in Table 2.

Table 2
Weighting of included studies

Study	WoE A*	WoE B	WoE C	WoE D
Balfanz, MacIver, and Byrnes, (2006)	Low	Low	Very Low	*Low*
Black, (2007)	Very Low	Very Low	Medium	*Low*

Cross, Gottfredson, Wilson, Rorie, and Connell, (2010)	Very Low	Very Low	Low	*Low*
Domitrovich et al., (2010)	Low	Medium	High	*Medium*
Dufrene et al., (2012)	High	Medium	High	*High*
Kam, Greenberg, and Walls, (2003)	Very Low	Low	Low	*Low*
Kupzyk, Daly, and Andersen, (2012)	High	Low	Medium	*High*
Lillehoj, Griffin, and Spoth, (2004)	Medium	Medium	Medium	*Medium*
Pas & Bradshaw, (2012)	Medium	Medium	Medium	*Medium*
Spoth et al., (2002)	Very Low	Very Low	Very Low	*Low*
Stein et al., (2008)	Low	Low	High	*Medium*
Stormshak, Dishion, Yasui, & Light, (2005)	Very Low	Medium	High	*Medium*
Taylor, Pearson, Peterson, & Rodriguez, (2005)	Very Low	Low	Low	*Low*

Results

Review question one: Is compliant implementation of evidence-based interventions in schools related to pupil outcomes? If so what is the strength of this association?

As this review was concerned with the 'real world' relationship between EBI implementation and pupil outcomes, included studies had to correlate or compare continuous measures of EBI implementation variance with pupil outcomes. Included studies used regression analysis, analysis of (co)variance and descriptive statistics to examine the strength of the relationship between compliance of intervention implementation (as measured by the dimensions in Table 1) and pupil outcomes. Standardised effect sizes were reported in four of the thirteen studies, three of which identified the statistic used. Raw data was used to calculate standardised effect sizes in a further four studies and unstandardised effect sizes in an additional two studies (Table 3).

Six of the thirteen studies used regression analysis to predict the effect of EBI implementation upon pupil outcomes (Domitrovich et al., 2010; Balfanz et al., 2006; Stein et

al., 2008; Taylor et al., 2005; Lillehoj et al., 2004; Pas and Bradshaw, 2012). Kam et al. (2003) and Spoth et al., (2004) used analysis of covariance to predict EBI implementation effects. Analysis of variance was used by Stormshak et al. (2005) to compare pupil outcomes across independent groups according to programme dosage. The inclusion of standardised measures of effect size (e.g., standardised betas or partial eta-squared) allowed the comparison of these effects between studies and was positively weighted in WoE B.

No statistical significance tests were used by Cross et al. (2010), Black (2007), Dufrene et al. (2012) or Kupzyk et al. (2012). These papers reported descriptive statistics and based subsequent analysis on comparisons between the reported raw data. For example, Dufrene et al. (2012) compared the average number of disruptive child behaviours with an average measure of teacher praise and effective instruction delivery.

Table 3 outlines the relationships between implementation dimensions and EBI outcomes across the reviewed studies. Reported and calculated effect sizes have been included alongside effect size interpretation. Effect size interpretation guidelines were taken from Cohen (1992) and Richardson (2011). Effect direction is positive (indicating a positive relationship between EBI implementation and pupil outcomes) unless it is labelled as a negative effect in parentheses.

Table 3
Implementation-Pupil outcome effect sizes within reviewed studies

Study	Name of EBI Implemented (pupil outcome targeted/type of study)	Implementation Dimension Measured	Effect Size (ES)	Effect Size Interpretation	WoE D
Balfanz, MacIver, and Byrnes, (2006)	Talent Development Middle School Mathematics Program (Maths achievement)	Quality	Implementation index-Maths achievement: *b*=1.9 β=0.15	*Small*	*Low*
Black, (2007)	Olweus Bullying Prevention Program (Bullying levels)	Fidelity	**d*=-0.5 (High/Low Fidelity groups x bullying incident density) **d*=1.03 (High/Low Fidelity groups x student reported bullying victimisation)	*Medium* *Large* (negative effect)	*Low*
Cross, Gottfredson, Wilson, Rorie, and Connell, (2010)	All Star Prevention Curriculum After School Program (Drug use and violence prevention)	Dosage Quality Engagement	Average Student Experience score for each school site correlated with 3ximplementation measures; **r*=0.86 (Management)	*Large*	*Low*
Domitrovich et al., (2010)	Head Start Research-Based, Developmentally Informed REDI (Social emotional competence, language and literacy skill)	Fidelity Generalisation Engagement Dosage	Unstandardised betas reported (n100, p.292) (unable to calculate standardised ES)	n/a	*Medium*
Dufrene et al., (2012)	Direct Behavioural Consultation (Disruptive classroom behaviour)	Fidelity	**r*=-.752 (Fidelity/Praise-Disruptive Behaviour) **r*=-.79 (Fidelity/Effective Instruction Delivery-Disruptive Behaviour)	*Large* *Large*	*High*

Kam, Greenberg, and Walls, (2003)	Promoting Alternative Thinking Strategies: PATHS (Delinquent behaviour)	Fidelity	Secondary implementation effects (effect of interaction between implementation-principal support): *η^2p=0.07 (aggression) *η^2p=0.08 (behavioural dysregulation) *η^2p=0.05 (social emotional competence) *η^2p=0.06 (on task behaviours)	 *Medium* *Medium* *Small* *Small*	*Low*
Kupzyk, Daly, and Andersen, (2012)	Parent Oral Reading Training (Oral reading fluency)	Fidelity	(unable to calculate standardised ES)	n/a	*High*
Lillehoj, Griffin, and Spoth, (2004)	Life Skills Training/Strengthening Families Programme (Substance related knowledge and behaviour)	Fidelity	β=0.02 Self-report implementation – multiple outcomes (average across 12 drug attitude and norm outcomes) β=0.06 Observer rating implementation – multiple outcomes (average across 12 drug attitude and norm outcomes)	*Null* *Null*	*Medium*
Pas & Bradshaw, (2012)	School Wide Positive Behaviour Intervention and Supports (Reading and maths achievement, attendance)	Fidelity	Implementation Phases Inventory – pupil outcomes: β=0.146 (Maths Achievement) β=0.171 (Reading Achievement) β=-0.088 (Truancy) β=-0.015 (Suspensions) Benchmarks of Quality (BOQ) – pupil outcomes: β=0.038 (Maths Achievement) β=-0.003 (Reading Achievement) β=-0.115 (Truancy)	 *Small* *Small* *Null* *Null* *Null* *Null* *Small*	*Medium*

			β=-0.115 (Suspensions) School-wide Evaluation Tool (SET)-pupil outcomes: β=-0.001 (Maths Achievement) β=-0.003 (Reading Achievement) β=-0.014 (Truancy) β=0.054 (Suspensions)	*Small* *Null* *Null* *Null* *Null*	
Spoth et al., (2002)	Life Skills Training (LST) Program (Substance related knowledge)	Fidelity	**d*=-0.054 (post-test) **d*=-0.16 (1.5 years post baseline)	*Null* *Null*	*Low*
Stein et al., (2008)	Kindergarten Peer Assisted Learning Strategies (Reading achievement)	Fidelity Engagement	Average fidelity rating - Reading Achievement *b*=0.16 (unable to calculate standardised ES)	n/a	*Medium*
Stormshak, Dishion, Yasui, & Light, (2005)	Family Resource Centre (FRC) School Provision (Problem behaviour)	Dosage	Program dosage – teacher perception of risk behaviour: *$\eta^2 p$=0.012	*Small*	*Medium*
Taylor, Pearson, Peterson, & Rodriguez, (2005)	CIERA School Change Framework (Reading achievement)	Fidelity	School reform effort – two reading measures: *b*=1.34, (**ES=0.29) Reading comprehension NCE *b*=4.87, (**ES=0.38) Reading fluency, words per min	 *Small*** *Medium***	*Low*

* Calculated by the reviewer from reported data.

** Effect Statistic unspecified; calculation using the method reported may have inflated the reported statistic.

Review question two: How has school-based research measured and analysed EBI implementation in schools?

All of the reviewed studies compared EBI implementation dimensions with pupil outcomes. The strength of these studies attending to real world implementation variance and potential effects should be acknowledged alongside common difficulties measuring implementation variation and analysing its effects. Some of these difficulties are considered below.

Conversion of interval data to ordinal data

All reviewed studies used implementation measures which yielded continuous data, for example the number of sessions delivered (EBI dosage) or the percentage of content delivered (EBI fidelity). All studies retained continuous data for use in their analyses with five exceptions (Black, 2007; Cross et al., 2010; Spoth et al., 2002; Stormshak et al., 2005; Kupzyk et al., 2012). These studies frequently converted continuous implementation measures into 'high', 'medium' and 'low' categories. No rationale was provided for the application of these criteria. Conversion of continuous data to ordinal categories negatively influenced WoE B weighting for these studies. This is because this practice discards score variability and distorts variable distributions, resulting in analyses that are less ecologically valid (Thompson et al., 2005).

Assessing the reliability of implementation measures and pupil outcome measures

Reliability is defined as the ability of a measure to yield consistent results when the same entities are measured under different conditions (Field, 2013). The reviewed studies used observation and self-report methods to measure EBI implementation. These methods can be undermined by observation effects and report bias (Spoth et al., 2002). For example, staff

members may behave differently when watched by a researcher or may feel pressured into reporting high intervention compliance. To understand the extent of these effects, some studies compared different implementation measures to see whether they correlated. For example, Lillehoj et al. (2004) calculated the percentage agreement between two different observers who measured programme fidelity (interrater reliability). When different measures of the same thing are correlated, this produces a reliability coefficient. Thompson et al. (2005) advise reporting the reliability coefficients for all measures used. For the purposes of this review, this included coefficients for all measures of implementation and pupil outcomes. Two of the included studies did not report reliability coefficients from internal data (Black, 2007; Stormshak et al., 2005). This was partially responsible for the very low WoE A weightings ascribed to these studies.

Reporting standardised and unstandardised effect sizes

Effect sizes represent the magnitude of an observed effect by quantifying the relationship between multiple groups or multiple variables (Field, 2013). For example, a relationship between teacher praise and attentive behaviour could be illustrated using a correlation coefficient (*r*). Unstandardised effect sizes are represented in their original units (e.g., on average, when a teacher praised a student four times, their attentive behaviour increased by one extra minute, r=0.25). Standardised effect sizes express the relationship between different variables using standard deviation as a unit, rather than the original measures. This supports readers to compare effects between studies, particularly if different measures are used which are difficult to associate (Field, 2013). Out of the reviewed studies, Balfanz et al., (2006), Lillehoj et al. (2004) and Pas and Bradshaw (2012) were the only ones to report standardised (β) and unstandardised (*b*) effect sizes. This was positively weighted in WoE A weightings, as it enabled the reviewers to compare implementation effects with other studies.

Discussion: Implications for research and practice

Review question one: Is compliant implementation of evidence-based interventions in schools related to pupil outcomes? If so what is the strength of this association?

This review yielded 32 quantified effect sizes which represented the strength of relationships between the degree of compliance of EBI implementation with the published protocols and pupil outcomes. Using interpretations from Cohen (1992) and Richardson (2011), eighteen of these effects were deemed positive, thirteen null (indicating no directional relationship) and one negative. This provides some support for the hypothesis that compliant EBI implementation in schools is positively associated with programme outcomes (Battistich, Schaps, & Wilson, 2004).

Of the eighteen positive effect sizes reported in Table 3, nine effects were rated as small, four as medium and five as large (Cohen, 1992; Richardson, 2011). This suggests that different aspects of EBI implementation, including programme and contextual variables, influenced programme outcomes. However, when WoE weightings were considered alongside these relationships it became apparent that the majority of large and medium effects were observed in studies which were rated as Low on overall WoE. In contrast, studies rated Medium and High predominantly observed null or small relationships between EBI implementation and child outcomes.

The findings represented in Table 3 suggest that studies of higher quality were less likely to observe strong EBI implementation effects in school settings. One explanation for this finding is that 'high quality' studies were more likely to use continuous implementation

measures (e.g., fidelity percentage), while 'low quality' studies were more likely to use categorical measures (e.g., 'good' or 'bad' fidelity labels) in their analyses. The findings of this review suggest that this practice may artificially increase effect sizes and implementation effects, undermining the purpose of measuring 'real world' implementation. This possibility has implications for EPs, whose role includes weighing evidence when considering which SEN and mental health interventions can be effectively transported into schools. Educational psychologists need to critically attend to the way implementation is measured when appraising the evidence-base for school-based EBIs, paying particular attention to the use of categorical implementation measures. It is preferable for researchers to report continuous measures of EBI implementation and retain these measures for analysis. This will support readers to compare achieved implementation between studies.

Although this review found that studies of higher quality were less likely to observe strong EBI implementation effects, there was one notable exception: Dufrene et al. (2012) reported a strong negative correlation between teachers' use of components from a 'compliance training package' and disruptive classroom behaviours. More specifically, increased use of praise statements and 'effective instruction delivery' (modifying instructions to include a waiting period for children) was associated with reduced demonstration of noncompliance and aggression. One explanation for these effects is that the 'compliance training package' targeted small groups of 10 pupils who were more likely to show positive outcomes compared to the larger populations targeted in some other studies. This is consistent with the argument that targeted interventions can demonstrate greater effectiveness if they are matched with a targeted population (Horowitz & Garber, 2006). A second explanation is that teachers were trained in the 'compliance training package' within a 'direct behavioural consultation' framework. This is a collaborative framework which emphasised joint problem

solving between the consultant trainers and consultee teachers. It is possible that staff members were increasingly motivated to adhere to the 'compliance training package' because they had established a relationship with the trainer or had increased awareness of programme components. Another consideration is that this was the only reviewed paper to use constant observations of staff implementation across all phases of the study. If a constant observer presence facilitated a more open relationship between researchers and practitioners, this may have resulted in participants adhering to protocols to a greater extent that they would if they had been observed intermittently or recorded fidelity using self-report measures, which are highly prone to response bias. This suggests that, when monitoring implementation, routine observation needs to be built in and arrangements for communication and open dialogue between EPs and staff members should be included. This will provide an important foundation from which the causes underpinning report-bias and observer effects can be explored and addressed.

Review question two, how is implementation measured and analysed during evaluation of school-based EBIs?

A range of implementation dimensions were measured across the 13 reviewed studies. Fidelity was the most commonly measured implementation dimension (measured by 10 studies), followed by dosage and engagement (n=3), quality (n=2) and generalisation (n=1). One measurement concern was the lower reliability and predictive power of teacher self-report measures. Domitrovich et al. (2010) observed low correlations between teacher and observer fidelity measures, while Lillehoj et al.'s (2004) regression ascribed greater predictive power to observational measures of fidelity (e.g., checklist completion by an independent observer) compared to teacher self-report. It is possible that report bias or social

desirability effects influenced self-report implementation ratings. Some studies attempted to account for these confounding effects by using observation checklists and comparing these with self-report implementation measures. An additional option would be to work with school staff to understand why social desirability effects may be present and how researchers and practitioners can overcome these. This would require partnership with school staff at an early stage in the research process so that both parties can identify the social pressures which may influence self-report implementation measures.

This review suggests that self-report measures of EBI implementation may be influenced by cultural, contextual and political factors within schools. However, there is evidence to suggest that there is potential for educational researchers to be similarly influenced. The five reviewed studies which categorised implementation measures reported predominantly high levels of EBI fidelity (Balfanz et al., 2006; Black, 2007; Cross et al., 2010; Kupzyk et al., 2012; Spoth et al., 2002). These studies collectively described 'high' implementation occurring 40% of the time, 'moderate' implementation occurring 38% of the time and 'low' implementation levels being achieved 22% of the time. This is congruent with Kratochwill and Shernoff's (2004) and Durlak and Dupree's (2008) claims that compliant implementation rarely occurs.

One reason for simplifying implementation compliance into categories (high, moderate and low) is that it supports understanding about the nature of programme delivery (Cross et al., 2010). However this review observed considerable variation in the use of semantic labels to describe EBI implementation. For example, an EBI could achieve 79% fidelity and be described as 'low' according to Spoth et al.'s (2002) criteria, while a 28% quality rating would be classified as 'medium-low' according to Balfanz et al. (2006). There is a need to

make these measures accessible to readers but this degree of inconsistency is concerning. If raw data suggests low implementation, then semantic descriptors should strive to represent this. When researchers use such labels they should be transparently defined alongside the original scaled measures to aid reader interpretation. The rationale for assigning labels to particular values (e.g., >80% fidelity = 'good') should be empirically justified using evaluations of existing programmes. Where these are not available, educational researchers should make the best use of available implementation research to justify the use of such terms.

With an increasing emphasis on the transportation of EBIs into schools (Department of Health & Department for Education, 2017) and the use of traded services to deliver applied psychology (Norwich, 2013) there is growing pressure on researchers and practitioners to demonstrate that EBIs can be feasibly delivered in schools. This exposes EPs, who are often associated with the interventions they are evaluating, to conflicts of interest. For example, schools may be increasingly motivated to purchase an EBI which is easy to implement, rather than one which indicates good outcomes when delivered with fidelity in evaluations. Researchers and practitioners need to anticipate these difficulties. One way to negotiate such conflicts would be to define semantic labels, such as 'high' and 'low', according to specific, measurable criteria and apply these consistently to describe implementation across the eight dimensions described by Durlak and DuPre (2008). These should be stated prior to analysis and based on existing evaluations of the EBI where possible.

Summary and Conclusions

The introduction to this article argued that schools have faced a research-practice gap dividing researchers prioritising EBI efficacy and practitioners prioritising EBI effectiveness. Kratchowill and Shernoff (2004) claim that researchers and practitioners should attempt to bridge this gap by working together to evaluate the transportability of EBIs from controlled conditions to real world school settings. More specifically, they argue that understanding EBI transportability requires educationalists to consider the feasibility of implementing EBIs in school settings.

This review found that educational researchers rarely measured EBI implementation during delivery in school settings. This oversight has prevented research consumers from understanding how EBIs can be feasibly delivered in ways that are compliant with programme protocols when transported from controlled research studies to real world school contexts. Educational psychologists therefore need to routinely monitor and report implementation dimensions (see Table 1) so that the profession can develop an understanding of what actually occurs when specific EBIs are delivered in schools. This requires EPs to select implementation dimensions which are appropriate and feasibly measured for the EBI and context in question.

The findings of this review suggest that the social context of research in schools can skew self-report implementation measures. Educational psychologists therefore need to increasingly measure EBI implementation in a way which promotes measurement validity. This could be attained through increased use of observation (e.g., fidelity/quality checklists) to monitor implementation dimensions. However, these methods are confined by the

increased resources required as well as potential observer effects. An alternative is to invite EBI implementers to identify such social pressures at the start of the research process and consider ways in which systematic reviews of implementation can be built into the evaluative process. This may include the use of video recording or training other observers within the school setting. School staff are able to consider contextual variables and this presents EPs with an opportunity to build on partnerships with schools to decide how to accurately measure EBI implementation. Collaborative questions may include: which implementation dimensions do staff think it is important to measure for this specific EBI and school setting? Which methods will allow staff to measure these dimensions feasibly?

In order to capture the full range of variation when using self-report methods, sources of additional information should be sought to triangulate self-report data where possible. Educational psychologists will also have to work with school staff to overcome the social influences which seemingly contribute to report bias. This should include normalising the difficulties of EBI delivery and debasing any expectations of 'perfect fidelity'. By creating a culture where researchers and practitioners are encouraged to work collaboratively to identify and manage sources of variation when implementing EBIs, EPs can develop better integration of objectives for researchers and practitioners, which will enhance the growth of practice-based evidence (Lucock et al., 2003). This review found that EBI implementation in schools is frequently reported as 'good' or 'high' even when raw data suggests considerable variation. It is therefore important for EPs to acknowledge that implementation of interventions frequently does not adhere to protocols and evidence suggests that this impacts negatively on pupil outcomes. However, through developing partnerships with school staff and encouraging identification of 'real world' instances of imperfect practice, EPs can open discussions about strategies to address the pressures that lead to compliance drift. This

review proposes that EPs have an important role to play in uniting researchers and practitioners when acknowledging implementation difficulties and using these to progress towards increasingly effective delivery of EBIs in school contexts, a goal in which school staff, EPs and pupils have a strong stake.

References

Balfanz, R., Mac Iver, D. J., & Byrnes, V. (2006). The implementation and impact of evidence-based mathematics reforms in high-poverty middle schools: A multi-site, multi-year study. *Journal for Research in Mathematics Education, 37*, 33-64.

Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of primary prevention, 24,* 243-262.

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

Black, S. (2007). Evaluation of the Olweus Bullying Prevention Program: How the program can work for inner city youth. In Hamilton Fish National Institute on School and Community Violence (Ed.), *Proceedings of Persistently Safe Schools,* (pp. 25-35).

Chorpita, B. F., Daleiden, E. L., Ebesutani, C., Young, J., Becker, K. D., Nakamura, B. J. et al. (2011). Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. *Clinical psychology: science and practice, 18,* 154-172.

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

Cross, A. B., Gottfredson, D. C., Wilson, D. M., Rorie, M., & Connell, N. (2010). Implementation quality and positive experiences in after-school programs. *American journal of community psychology, 45,* 370-380.

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly, 25,* 284-298.

Department of Health and Social Care and Department for Education (2017). *Transforming children and young people's mental health provision: a green paper.* Cm 9523. APS Group.

Dufrene, B. A., Parker, K., Menousek, K., Zhou, Q., Harpole, L. L., & Olmi, D. J. (2012). Direct Behavioral Consultation in Head Start to Increase Teacher Use of Praise and Effective Instruction Delivery. *Journal of Educational and Psychological Consultation, 22,* 159-186.

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,* 327-350.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S. et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6,* 151-175.

Fox, M. (2003). Opening Pandora's Box: Evidence-based practice for educational psychologists. *Educational Psychology in Practice, 19*, 91-102.

Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Research papers in education, 22,* 213-228.
Horowitz, J.L. & Garber. (2006). The prevention of depressive symptoms in children and adolescents: a meta-analytic review. *Journal of consulting and clinical psychology, 74,* 401-415.
Horner, R. H., Carr, E. G., Halle, J., Mcgee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179.
Kam, C. M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4,* 55-63.
Kessler, R. & Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. *American journal of preventive medicine, 40,* 637-644.
Kratochwill, T. R. & Shernoff, E. S. (2004). Evidence-based practice: Promoting evidence-based interventions in school psychology. *School Psychology Review, 33,* 34-48.
Kratochwill, T. R. & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17,* 341-389.
Kupzyk, S., Daly III, E. J., & Andersen, M. N. (2012). Preparing Teachers to Train Parents to Use Evidence-Based Strategies for Oral Reading Fluency with Their Children. *Contemporary School Psychology, 16,* 129-140.
Lillehoj, C. J. G., 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,* 242-257.
Lucock, M., Leach, C., Iveson, S., Lynch, K., Horsefield, C., & Hall, P. (2003). A systematic approach to practice-based evidence in a psychological therapies service. *Clinical Psychology & Psychotherapy, 10,* 389-399.
Norwich, B. (2013). Understanding the profession of educational psychology in England: Now and in the future. *The Educational and Developmental Psychologist, 30,* 36-53.
Ogden, T. & Fixsen, D. L. (2015). Implementation science: a brief overview and look ahead. *Zeitschrift fur Psychologie, 222,* 4-11.
Pas, E. T. & Bradshaw, C. P. (2012). Examining the Association Between Implementation and Outcomes. *The Journal of Behavioral Health Services & Research, 39,* 417-433.
Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 135-147.
Schaeffer, C. M., Bruns, E., Weist, M., Stephan, S. H., Goldstein, J., & Simpson, Y. (2005). Overcoming challenges to using evidence-based interventions in schools. *Journal of Youth and Adolescence, 34,* 15-22.
Spoth, R., Guyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a communit-university collaboration context. *Journal of Community Psychology, 30,* 499-518.
Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, L., Yen, L. et al. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis, 30,* 368-388.

Stormshak, E., Dishion, T. J., Yasui, M., & Light, J. (2005). Implementing family-centered interventions within the public middle school: Linking service delivery to change in student problem behavior. *Journal of Abnormal Child Psychology, 33,* 723-733.

Sutherland, K. S., Conroy, M. A., Vo, A., & Ladwig, C. (2015). Implementation integrity of practice-based coaching: Preliminary results from the BEST in CLASS efficacy trial. *School Mental Health, 7,* 21-33.

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2005). The CIERA school change framework: An evidenc-based approach to professional development and school reading improvement. *Reading Research Quarterly, 40,* 40-69.

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*(2), 181-194.