Professional Development in Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjie20

The search for evidence-based features of effective teacher professional development: a critical analysis of the literature

Christa S. C. Asterhan & Adam Lefstein

To cite this article: Christa S. C. Asterhan & Adam Lefstein (20 Nov 2023): The search for evidence-based features of effective teacher professional development: a critical analysis of the literature, Professional Development in Education, DOI: 10.1080/19415257.2023.2283437

To link to this article: https://doi.org/10.1080/19415257.2023.2283437

Published online: 20 Nov 2023.

PROFESSIONAL DEVELOPMENT IN EDUCATION
https://doi.org/10.1080/19415257.2023.2283437

Routledge
Taylor & Francis Group

ARTICLE

OPEN ACCESS Check for updates

The search for evidence-based features of effective teacher professional development: a critical analysis of the literature

Christa S. C. Asterhan and Adam Lefstein

The Seymour Fox School of Education, The Hebrew University of Jerusalem, Jerusalem, Israel

ABSTRACT

Scholarly efforts to identify core design features for effective teacher professional development have grown rapidly in the last 25 years. Many concise lists of design principles have emerged, most of which converge on a consensus of 5–7 presumably 'effective' features (e.g. collaborative tasks, active learning, focus on content). The proliferation and convergence of reviews create the impression that this consensus is based on strong evidence from large-scale, replicated, and rigorously controlled research studies. We critique the empirical foundation on which conclusions about evidence-based design features for teacher professional development have been based, by the same evidential standards that have been adopted within this field of scholarly work. We conclude that the empirical foundations for these lists are problematic and that claims to methodological rigour are misleading as they are based on flawed inferences. We further argue that the ambition to identify general features of effective professional development is also problematic, and reflect on why, despite its weaknesses and potentially adverse consequences for research and practice, we as a field continue to herald this consensus. We call for greater focus on the development, testing, and refinement of theories about teacher professional learning in order to advance understanding, policy, and practice in the field.

ARTICLE HISTORY
Received 26 March 2023
Accepted 29 October 2023

KEYWORDS
Teacher professional development; critical review; evidence-based policy; effective design

Policy-makers appreciate simple answers to complex questions. In the field of teacher professional development (PD), at least, the research community has tended to oblige them. Specifically, educational researchers have been producing lists of the core features of effective PD designs for a quarter century at least. Guskey (2003) reviewed 13 such lists published between 1995 and 2001, noting how claims about a consensus among researchers, professional development specialists, and policymakers already surfaced before the turn of the century (e.g. Hawley and Valli 1999). More recently, Desimone (2009) synthesised findings from the available literature at the time by concluding that 'there is enough empirical evidence to suggest that there is in fact a consensus on a core set of features' (p. 183) for effective teacher professional development efforts: (a) a focus on subject matter content and how students learn that content, (b) collaboration and interaction with colleagues, (c) engagement in active learning tasks for teachers, (d) coherence with existing curricula and policies, and (e) extended duration of PD programs.

Many other systematic literature reviews and large-scale survey studies have arrived at similar conclusions (e.g. Kennedy 1998, Garet *et al.* 2001, Desimone *et al.* 2002, Penuel *et al.* 2007,

CONTACT Christa S. C. Asterhan asterhan@huji.ac.il The Seymour Fox School of Education, The Hebrew University of Jerusalem, Jerusalem, Israel

Timperley *et al.* 2007, Van Driel *et al.* 2012, Darling-Hammond et al. 2009, 2017, Scher and O'Reilly 2009, Van Veen *et al.* 2012, Cordingley *et al.* 2015, Dunst *et al.* 2015, Ciesielski and Creaghead 2020, Hubers *et al.* 2022). These lists of core features have been adopted, expanded, and reiterated in numerous scholarly publications, reports and PD programmes, both as a rationale for designing new PD efforts, as well as a starting point for subsequent research. Indeed, it is rare to *not* find these core features mentioned as an evidence-based starting point in the introductory section of a report, thesis, dissertation, or research publication related to PD. To date (14 November 2023), the Desimone (2009) paper has been cited 6,688 times according to Google Scholar and recently garnered the prestigious SAGE publishing house 10-year impact award. Garet *et al.* (2001) was cited 9,311 times, and the Darling-Hammond *et al.* (2017) report has received 4,435 citations in the 5.5 years since its publication. The numerous citations and reiterations of these lists of effective PD features have further cemented this sense of consensus in the field.

However, in spite of this oft-cited consensus, some scholars have critiqued the empirical research on which it is based (Wayne *et al.* 2008, Hill et al. 2013, 2022, Sims and Fletcher-Wood 2021) or have failed to find such distinctive features (e.g. Guskey 2003, Yoon *et al.* 2007). Moreover, recent studies in which PD has been purposefully designed according to these features and then compared to control groups have not produced the anticipated results (e.g. Garet et al. 2008, 2011, 2016, Yang *et al.* 2020).

In the present essay, we aim to explore this apparent discrepancy by critically examining the empirical research base from which these sets of core features for effective PD have been derived. We critique this literature from within its own parameters and criteria: A general characteristic of this scholarly literature is the aspiration to rely on the strongest empirical evidence possible. To this end, it has prioritised research methods that allow causal inference, that are based on large data sets, that use quantifiable student (and/or teacher) outcomes as the main measure of success, and that have been tested across multiple studies. We show how, within these parameters, common claims about the methodological rigour of these lists are misleading, as they are based on problematic inferences. In the second part of this essay, we problematise the very project of attempting to identify universal PD design features as a feasible, or even desirable, goal.

Our argument in these sections about the limits of what we do and can know may seem minor, even pedantic. However, in the social life of research and policy, such minor issues can have major consequences for policy, practice, and research. We conclude this essay with reflections on why, despite their weaknesses and potentially adverse consequences, we as a field continue to produce such lists of effective PD features.

The research base underlying the lists of effective PD features

To date, most empirical efforts to identify effective PD design features are based on comprehensive literature reviews of primary research employing (quasi-)experimental comparisons of PD vs. no-PD conditions or large-scale, correlational studies of variance in outcomes of existing PD programmes. Although not entirely mutually exclusive, we discuss these two types of research separately, and then consider a third and less frequent research design, namely controlled experimental studies directly comparing PD design features.

Literature reviews of (quasi-)experimental research

The first category contains literature reviews of collections of (quasi-)experimental research (e.g. Kennedy 1998, 2016, Hawley and Valli 1999, Timperley *et al.* 2007, Yoon *et al.* 2007, Blank and de las Alas 2009, Desimone 2009, Scher and O'Reilly 2009, Walter and Briggs 2012, Gersten *et al.* 2014, Darling-Hammond *et al.* 2017, Maandag *et al.* 2017, Lynch *et al.* 2019, Hubers *et al.* 2022). These include different types of literature reviews, such as systematic and narrative reviews, best-evidence syntheses, rapid reviews, and meta-analyses. Many (but not all) of these were conducted at the behest of government agencies, think tanks, policy institutes, and non-governmental organisations. Their overall goal is to stipulate not only *whether* PD has an impact but also, and more importantly, to identify the design features that set successful PD programmes apart. Some limit their focus to specific areas of teaching (e.g. STEM or mathematics) while others lump all content areas together.[1] Selection criteria for including primary research are stipulated by the What Works Clearinghouse (WWC) standards of rigorous research, prioritising randomised controlled experiments that contain some form of systematic, controlled comparisons between quantified classroom outcomes (teacher behaviour and/or student achievement) at scale. On its face, this endeavour seems commendable. Moreover, the fact that different literature reviews conducted by teams from different countries, including at least three meta-reviews (Cordingley *et al.* 2015, Dunst *et al.* 2015, Cirkony *et al.* 2022), converge on similar sets of core PD design features seems promising.

However, the quality of any systematic review or meta-analysis depends upon the quality of the primary studies on which they are based and on the rigour and relevance of the inclusion criteria used by reviewers (Davies 2000). In a recent critique of the literature, Sims and Fletcher-Wood (2021) already highlighted the many cases in which authors failed to report on their search and inclusion criteria. They also show that considerable chunks of the primary research base used in many reviews do not approximate WWC standards of rigour, despite claims to the contrary.

Here, we would like to highlight an additional and basic caveat that is often overlooked, even in existing critiques of this literature and even in the most sophisticated and meticulously executed meta-analyses (e.g. Lynch *et al.* 2019). We argue that the primary research on which the systematic literature reviews and meta-analyses are based are problematic grounds for drawing inferences about the relative effectiveness of PD design features. This problem is not related to whether the primary research studies adhered to WWC standards or not (some did, others did not). The problem arises from the fact that in the vast majority of these studies effective PD was not the *object* of study itself, but rather a *precondition* to study another object of interest.

To be able to infer that a particular PD *feature* is more effective than another requires that research designs enable comparisons between two (or more) PD programmes that differ in terms of select features (e.g. with or without engagement in active learning tasks) but are otherwise identical (e.g. with regard to content, duration, PD facilitators' background etc.). However, until recently, the educational research community has not demonstrated much interest in experimental, large-scale research on PD design as an object of inquiry in and of itself (Lynch *et al.* 2019). Thus, such direct, controlled comparison studies have been rare (we discuss specific exceptions in section 1.3).

There is, however, abundant educational research that includes elements of teacher PD and employs an experimental design, though the PD design features were not the focal variables examined. In the overwhelming majority of such studies, the objects of investigation are interventions that are hypothesised to improve instruction quality and student outcomes. To enable such a study, researchers typically provide PD to the participating teachers in order to assist them in implementing the intervention – a new curriculum, teaching practices, instructional materials, or some combination thereof. Researchers in such studies are careful to design their PD delivery according to the current state of the art, but their studies are designed to test the effects of the instructional improvement, not to test the features of the PD. They therefore do not include comparisons of different *PD features*, because PD design is not the object of investigation, but rather a prerequisite to create differences between the conditions that are.

Thus, in the majority of studies that make up the primary research base for the many reviews and meta-analyses, teachers in the treatment group participate in some form of PD delivery to learn new ways of teaching, or to work with newly developed materials or curricula to improve instruction, whereas teachers in the control condition receive neither. To illustrate this with a specific example, let us consider a study by Carpenter *et al.* (1989), which features in many systematic literature reviews that seek to identify effective PD design features. Carpenter *et al.* (1989) investigated how teacher knowledge about children's understanding of mathematical topics improved teaching practices and student achievement.

Teachers were randomly assigned to either participate in an 80-hour PD programme about children's mathematical thinking or to a business-as-usual control group (i.e. no PD at all). This study was indeed designed according to rigorous standards of randomised experimental field research and provides strong evidence of how teachers' pedagogical content knowledge can improve student achievement. In other words, it is a very rigorous study, which provides strong evidence for claims about the importance of teacher knowledge about children's understanding of mathematics. However, it does not provide any research evidence about how to best design the teacher PD programme to improve that knowledge (such as collaborative teacher learning tasks or extensive PD programme duration), because the control group did not receive any PD at all.

Unfortunately, this study is typical of the body of primary research upon which systematic reviews and meta-analyses on effective PD design features base their conclusions. In light of the dearth of rigorous research that directly compares different PD programmes or features and the abundance of primary research on instructional improvement (which *de facto* includes some PD in the experimental condition but no PD in control conditions), the comprehensive literature reviews are based on the latter type, either entirely (e.g. Yoon *et al.* 2007, Gersten *et al.* 2014) or largely (e.g. 24 out of 35 studies included in Darling-Hammond *et al.* 2017). Moreover, even when a few studies with direct and controlled comparisons of PD design features were included in the selected set of primary research studies, these are not distinguished from the rest, nor given any special status in the analyses (e.g. in Lynch *et al.* 2019).

Given the paucity of controlled experimental studies that directly compare PD design features, and given the pressure to find effective design principles, it is understandable that scholars would turn first to existing research in their search for evidence. However, the overreliance on a body of literature that was never designed to provide answers to questions about PD design can and has led to flawed inferences.

We would like to illustrate the implications of this common-sensical, yet surprisingly often overlooked caveat with an imaginary and purposely simplistic analogy from the field of medicine. Imagine that we want to ascertain what are the most effective methods for the packaging and delivery of medical treatments (i.e. pills) through a thorough and systematic research review. Unfortunately, no one has yet conducted rigorous controlled comparisons of the different methods. However, we have a wealth of primary research studies that are highly rigorous in the way that they study the relative effectiveness of different pills for the treatment of a vast variety of medical conditions. All of these studies employ methods for medicinal packaging and delivery. So, we survey this literature, asking what are the packaging and delivery methods used in the most rigorous studies showing the greatest degrees of effectiveness. We find that in all these studies the pills are packaged in blister packs of thermoformed plastics with aluminium foil lids. Clearly, many medical researchers and practitioners have faith in blister packs, which has indeed become standard practice in the pharmaceutical industry. However, we should not attribute the success of the medical treatments to the use of blister packs or claim that this packaging method is superior to alternatives, which were not tested.

Similarly, in the field of professional development research, many experiments that test instructional innovations use a core set of professional development methods to deliver their treatments to teachers. Clearly, most educational researchers and practitioners have faith in these PD methods, which have indeed become recognised as good practice in the field. However, when these experiments on instructional innovations are successful, i.e. lead to improvements in student achievement, we cannot necessarily attribute their success to the PD methods used or claim that these methods are superior to alternatives, which were not tested. Yet, that is exactly the type of faulty reasoning employed when 'effective PD design features' are extracted from primary research that was never designed to answer that question and does not compare different PD features at all.

Large-scale correlational studies of instructional reforms

The second category of scholarly work concerns correlational studies that accompany large-scale reform efforts in which decisions about the exact form of PD delivery and implementation have been left to local management (e.g. Cohen and Hill 1998, Garet *et al.* 2001, Desimone *et al.* 2002, Penuel *et al.* 2007, Fischer *et al.* 2018). In these cases, topical content can be assumed to be held fairly constant, since it is stipulated and provided by central administration, but PD delivery format is not. Such situations create opportunities to study the correlates of naturally occurring variance in PD format in relatively controlled settings.

However, a major drawback of these studies is that, notwithstanding the often awe-inspiring data collection efforts, the complex statistical modelling and analyses, in the end these are survey studies that rely almost entirely on self-report data, instead of objective measures of PD delivery or teacher behaviour. Participating teachers are surveyed with regard to their professional development experiences (e.g. the extent to which they focus on content, their coherence with goals and expectations, types of PD activities offered), as well as key programme outcomes, such as self-report estimations of their knowledge, their teaching capacity, and/or their instructional practices. Whereas in some cases, scholars also managed to collect external measures, such as student achievement scores (e.g. Fischer *et al.* 2018) or some rough indicators of practice (e.g. from teachers' downloads of Web-provided teaching materials, Penuel *et al.* 2007), the majority of significant findings directly related to PD effectiveness features are based on different types of teacher self-report data (either on PD features, on practice, or on both). In some cases, teachers are even asked *directly* about the effects of PD on their teaching. For example, in the highly influential study by Garet *et al.* (2001, also used in Penuel *et al.* 2007), teacher outcomes were assessed with items, such as '*Please indicate the extent to which you made changes in your teaching practices as a result of the PD programme (on a scale from 0 = none to 3 = significant changes)*'.

We should exercise caution in interpreting these and similar findings. First, research from adjacent fields has shown that learners' subjective self-reports about what works best for them do not always align with objective measures of learning and change (Kirschner and van Merriënboer 2013). In fact, a recent study comparing self-reports with direct assessments of teacher knowledge gains following PD found no correlations between them (Copur-Gencturk and Thacker 2021).

Second, teachers are by no means naïve about theories of learning and effective PD. We assume that they too are influenced by the prevailing common sense about effective PD, shaped in part by a quarter century of reports listing core features. Similarly, subjective reports about their own instructional practices are also likely to be shaped by current views on what effective PD *should* look like, especially when the programme in which they participated made these elements even more salient.

Third, since people tend to align their perceptions, attitudes, and behaviour to be consonant (Festinger 1957), the existence of an association between targeted PD features and self-reported change measures may be explained by a reverse pattern of causation, especially when surveys are administered in hindsight. Having invested a considerable chunk of time in a PD programme, teachers may retrospectively come to perceive it as beneficial – otherwise, why did they persist (Arkes and Blumer 1985)? This may explain the positive associations in some of the reports between amount of contact hours in the PD programme and self-reported instructional change. In another example, the positive correlation between coherence with local norms and standards and self-reported instructional change may be explained by the possibility that those teachers who succeeded in implementing the targeted changes perceived the programme to be more coherent with local norms and standards in hindsight – since most teachers wish to view their practice as aligned with such guidelines and expectations (see also Penuel *et al.* 2007 for a similar argument).

All research methodologies have limitations and advantages, and we do not wish to downplay the importance of correlational studies of large-scale reform initiatives or the use of self-report surveys. However, we should be aware of what we can and cannot reasonably infer from them. In this particular case, and based on the aforementioned alternative explanations for reported findings, we should exercise caution in using this collective body of research to further cement claims about the relative effectiveness of various PD design features.

Experimental studies directly comparing PD design features

Perhaps due to increasing interest in PD design and recent shifts in funding policies (Lynch *et al.* 2019), several recent publications have reported on controlled experiments specifically designed to compare different PD formats (Garet *et al.* 2008, Russell *et al.* 2009, Fisher *et al.* 2010, Powell *et al.* 2010, Penuel *et al.* 2011, Heller *et al.* 2012, Fishman *et al.* 2013, Grigg *et al.* 2013, Piasta *et al.* 2017, Taylor *et al.* 2017, Osborne *et al.* 2019, Yang *et al.* 2020).[2] Some of these experiments even targeted PD programme features selected from the aforementioned lists (e.g. Garet *et al.* 2008, Yang *et al.* 2020). With a few exceptions (Heller *et al.* 2012, Taylor *et al.* 2017), however, the different PD programme design features targeted in these studies were overall not found to lead to significant differences in PD outcomes, particularly *student* outcomes. For example, Garet *et al.* (2008) compared a standard PD programme for early reading interventions with two PD programmes specifically designed to integrate the effective PD design recommendation of 'content-focus' (both programmes), as well as continuous one-on-one coaching sessions (in only one programme). The three PD programmes achieved similar results. Similarly, different versions of a PD programme for improving argumentation in science classes produced comparably favourable outcomes in Osborne *et al.* (2019). Finally, Garet and colleagues conducted two separate studies (Garet et al. 2011, 2016) in which they compared PD programmes that were designed specifically according to effective PD design principles (especially duration and focus on content) to control conditions (PD business as usual). They found no effects on student outcomes, and only on a few teacher measures. Either the underlying instructional interventions (Desimone 2023) or the teacher PD designs were ineffective, or both.

In sum: what is the quality of the evidence behind claims about evidence-based effective PD design features?

Based on extensive literature reviews and large-scale quantitative studies, lists of design principles for effective PD programmes have been compiled. These principles corroborate with common sense, current views about meaningful PD, practitioners' experiences, and socio-constructivist, situative, and cognitive theories of learning. Moreover, these lists include designs that have been employed in countless studies in which they have been associated with measurable improvements in student outcomes. Reviews of this literature claim to privilege experimental and quasi-experimental studies (even though they differ in the extent to which they actually adhere to these criteria in a strict sense, see Lynch *et al.* 2019, Sims and Fletcher-Wood 2021). Some (but not all) of the reports and meta-analyses employ the accoutrements of rigorous, systematic reviews, including exhaustive searches based on clear criteria and using sophisticated statistical methods (e.g. Lynch *et al.* 2019).

Nevertheless, despite the impression created by the many influential publications, we conclude that, by the different authors' own evidential standards, the empirical foundations for these lists of PD design features are problematic, and the claims to methodological rigour are misleading. Not because the primary research studies on which the extensive reviews are based were weak or badly executed, but because most of them were never designed with the intent of comparing PD designs and, therefore, do not lend themselves to drawing inferences about relative effectiveness.

With the growing attention to teacher learning as a topic of empirical interest in and of itself, direct controlled comparisons are gradually becoming more frequent. Thus far, however, findings from these studies do not echo the expectations about differential effects for different PD design features, not even the ones identified in the core features lists: Even though PD often had an overall effect on student and/or teacher outcomes when compared to no PD conditions, different PD programme designs on the same topic rarely produced differences in outcomes. Null results do not prove that the design features are inconsequential of course. Yet, the overall paucity of effects from direct and controlled comparisons does raise further questions concerning the prevailing consensus with regard to existing lists of evidence-based PD design features.

Should we attempt to identify general effective PD features?

We have critiqued here the methodological shortcomings of the empirical foundation on which claims about the identification of evidence-based, effective PD design features have been based. This critique can benefit future reviews of research on teacher PD effectiveness, as it highlights crucial pitfalls that plague many previous reviews, such as only including studies with positive outcomes (e.g. in Darling-Hammond *et al.* 2017), and not differentiating between studies that include PD as a means to an end and studies that directly investigate and compare PD designs. Heeding these calls for research and attending to these issues in reviews should improve the evidence base while remaining within the current paradigm. This critique also echoes previous calls (e.g. Hill *et al.* 2013) for more research that directly targets PD design as an object of study.

Nevertheless, we wish to also problematise some of the assumptions underlying the entire programme that seeks to uncover general features for effective PD design and the adoption of RCT-driven standards of research to achieve it: First, we posit that it is unrealistic to expect to find a one-size-fits-all answer to the question of what PD approach is more or less effective. Some aspects of teaching may be easier to change than others (Borko 2004, Desimone and Garet 2015, Kennedy 2016, Desimone 2023). Effectiveness is, among other factors, likely to be a product of the interaction of means and ends (different design features for different types of knowledge and skills). For example, it is reasonable to expect that PD focusing on teaching routine skills may be more effective through direct instruction and individual exercises to improve fluency, whereas improving professional judgement may best be accomplished through hands-on simulations or collaborative sense-making and reflection on representations of practice with like-minded colleagues (Horn and Garner 2022).

Second, some PD design features are more likely than others to impact the ‘gold standard’ of outcome measures usually targeted in this field of research (i.e. standardised student achievement scores). PD programmes that target the effective teaching of specific disciplinary content, are likely to show a stronger statistical association with standardised student test scores than PD programmes that focus on more general aspects of teaching or on supporting more general student skill development (e.g. argumentation skills), since most standardised achievement tests measure student content knowledge. This issue further confounds the comparison of findings regarding core PD design features. It may also partially explain why the PD design feature of ‘focus on content’ emerges on so many different lists of effective PD features (see also Hill *et al.* 2022, for a critique of the focus on teacher content knowledge as an effective design feature).

Third, the effectiveness of PD programmes and their delivery is likely dependent upon a variety of environmental factors, such as teachers’ work conditions, incentive structures, curricular materials and other resources, school leadership, and informal teacher learning processes. For example, available evidence (e.g. Kraft and Papay 2014, Ronfeldt *et al.* 2015) suggests that school professional environments are consequential for teacher learning and effectiveness. Attending to these and other environmental factors that likely mediate PD programme effects would enhance both the study and design of PD.

Fourth, merely trying to identify a set of features will not be enough. A design’s effectiveness critically depends on how it is enacted (Patfield *et al.* 2021). Take, for example, decades of research on student-led, small group learning. Yes, group work can be effective for certain types of student learning outcomes, but its effectiveness is dependent, among other conditions, on the type of task that students are assigned, students' collaboration and communication skills, and the availability and quality of teacher facilitation (e.g. Webb 2009). Likewise, with regard to teacher PD: Research on effectiveness can usefully inform this work, but it needs to also take into account the aspects of teaching and learning targeted, the policy, professional and school environment in which teachers work, facilitators' and teachers' professional knowledge, skill, judgement, and wisdom, and how all these factors interact to shape the PD design's enactment.

Fifth, since the ultimate aim of most PD efforts is to improve student learning, it is not surprising that student test scores have become the 'gold standard' for PD effectiveness. However, it is also a very ambitious standard as it is the most distal variable in a long causal chain of effects: The PD programme is expected to impact teachers' skills, beliefs, and/or knowledge, which translate into differences in participants' classroom practices, which affect cognitive, motivational, and/or affective aspects of student action, which, eventually, translate into individual student test scores (Kennedy 2016).

Finally, focusing exclusively on associations between PD features on one end of this chain and student outcomes on the other overlooks the importance of better understanding the processes and mechanisms of teacher professional learning (Kennedy 2016). Design features do and do not work for reasons that are partially rooted in our theoretical understandings about how teachers learn and improve their practice. Improving designs requires improving theory (Horn and Garner 2022). Hence, if we as a field are serious about improving returns on teacher PD efforts, then we should prioritise the development, testing, and refinement of theories about teacher professional learning and move beyond the process – product logic that has dominated the literature (Opfer and Pedder 2011, Hill *et al.* 2013, Boylan *et al.* 2018, Strom and Viesca 2021). This effort involves not only asking how well a professional development practice works but also why it does and does not work and under which conditions (Opfer and Pedder 2011). For example, Hill and Papay (2022) note that some successful PD programmes employ teacher-driven follow-up sessions in which participants share with peers their experiences enacting the instructional practices they are learning. They offer a number of conjectures about why such sessions may be effective, including providing support in meeting challenges of implementation and functioning as a form of social accountability. Hill and Papay suggest that research could vary the forms and functions of these meetings in order to better understand how they operate.

Randomised controlled research designs could (and perhaps even should) be a part of such a research agenda but are likely to be more powerful if they focus on associations between less distal factors. Using large scale RCTs for studying the effect of selected PD design features on student outcomes (two very distal variables) is theoretically possible, but not necessarily the most optimal use of limited resources. It requires very large samples to obtain adequate power to detect the differential impact of a small number of features, and even then, the chances of finding significant differences are small. This difficulty may partially explain the null effects found in the studies reviewed in section 1.3 above.

Why do we keep making lists of core PD design features even though their evidence base is problematic?

Many of the issues we have raised here are not new, yet somehow, despite the criticisms, these lists of effective PD features continue to proliferate. In closing, we wish to reflect on the gravitational forces that appear to be pulling the field to identify 'effective' features and to present them with greater confidence than the evidence warrants. Perhaps, our own experience may be instructive in this regard. We initially began reviewing the research on effective teacher professional development as members of an Academy of Sciences Consensus Panel commissioned by the Ministry of

Education in Israel to study ways of improving teacher professional development. We noted in our report chapter the claims of a consensus about core effective features, and also the problematic evidence base upon which this consensus rests. Much to our dismay, the list of five core features appeared prominently in the draft executive summary and short animated film produced by the Academy administrative staff, without any of our reservations. Policy-makers want the bottom line, they explained, without all the hedging and qualifications. A bullet-pointed list of features, concise enough to fit on one slide or one frame of an animated film, is exactly what they are looking for.

We sense that we are not alone in this experience. A recent National Academy of Sciences Consensus Study Report (National Academies of Sciences, Engineering and Medicine 2020) included in its summary and conclusion chapter a statement that the evidence regarding the impact of professional development on student outcomes is 'mixed', but that 'there is better evidence that in-service, content-specific professional development programmes with the following characteristics can have a positive impact on student learning' (p. S5). The summary and conclusion then list four features that do not appear in that format in the relevant chapter of the report, which presents a nuanced and critical discussion of the relevant research. Likewise, Desimone *et al.* (2002) note that, though a consensus about effective PD characteristics frequently appears in the scholarly literature, 'there is little direct evidence on the extent to which these characteristics are related to better teaching and increased student achievement' (p. 82, similar disclaimers appear in Desimone 2009, Desimone and Garet 2015). However, in the many thousands of references to these studies, the core features frequently appear as a definitive and rigorous evidence-based finding.

How can we explain this pressure to erase the nuance and present solid evidence about effective PD features? We speculate that an important source of pressure is the wide-spread sense of dissatisfaction with current PD programmes and practices, bolstered in part by high profile reports on the ineffectiveness of existing practice (e.g. TNTP 2015), on teacher dissatisfaction with PD (e.g. Boston Consulting Group 2014) and on uneven returns on the considerable resources that governments invest in PD (Jacob and Lefgren 2004). As a result, policymakers may be uneasy about continuing to invest in what are believed to be ineffective practices. Reports on evidence-based, rigorously researched PD design features may help alleviate their concerns.

Likewise, researchers also have a vested interest in participating in the identification and circulation of lists of 'effective' PD features. Recall that most research involving PD does not investigate it as its primary object, but rather an intervention on curriculum or instruction. We researchers need to convince policymakers, funders, and journal reviewers that our interventions are based on solid evidence. Basing our designs on 'the current consensus' about effective PD absolves us of the need to actually test these designs, thereby allowing us to focus on the curriculum, instructional strategy, or learning materials we have developed.

The lists of features are not inherently flawed. Indeed, they make a lot of sense theoretically, even though they are rather general and open to different interpretations. The primary problem of attributing to these lists greater certainty than they deserve is that it creates the impression that the issue has been settled. As a result, researchers are more likely to base their PD programmes on these features, rather than directly studying them, and this dynamic may keep us as a field from breaking new ground in the study of teacher learning and professional development. We hope that this critical review will contribute to awakening the field from our dogmatic slumber.

Notes

1. Some reviews are dedicated to selected formats of PD, such as teacher coaching programmes (e.g. Kraft *et al.* 2018). These reviews are beyond the scope of our discussion here, which focuses on the many efforts to extract general PD design features without further specifications.

2. A few studies in which different forms and duration of PD were compared were omitted from this list as they contained study-specific confounds or other specific issues that render the interpretations of their results equivocal (e.g. Landry *et al.* 2009, Roth *et al.* 2011, Vernon-Feagans *et al.* 2015).

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Adam Lefstein http://orcid.org/0000-0002-9686-2662

References

Arkes, H.R. and Blumer, C., 1985. The psychology of sunk cost. *Organizational behavior and human decision processes*, 35 (1), 124–140. doi: 10.1016/0749-5978(85)90049-4

Blank, R.K. and de las Alas, N., 2009. *The effects of teacher professional development on gains in student achievement: how meta-analysis provides scientific evidence useful to education leaders*. Washington, DC: Council of Chief State School Officers.

Borko, H., 2004. Professional development and teacher learning: mapping the terrain. *Educational Researcher*, 33 (8), 3–15. doi:10.3102/0013189x033008003.

Boston Consulting Group, 2014. *Teachers know best: teachers' views on professional development*. Seattle, WA: Bill and Melinda Gates Foundation.

Boylan, M., *et al.*, 2018. Rethinking models of professional learning as tools: a conceptual analysis to inform research and practice. *Professional development in Education*, 44 (1), 120–139. doi: 10.1080/19415257.2017.1306789

Carpenter, T.P., *et al.*, 1989. Using knowledge of children's mathematics thinking in classroom teaching: an experimental study. *American Educational research journal*, 26 (4), 499–531. doi:10.3102/00028312026004499.

Ciesielski, E.J. and Creaghead, N.A., 2020. The effectiveness of professional development on the phonological awareness outcomes of preschool children: a systematic review. *Literacy research and instruction*, 59 (2), 121–147. doi:10.1080/19388071.2019.1710785.

Cirkony, C., *et al.*, 2022. Beyond effective approaches: a rapid review response to designing professional learning. *Professional development in Education*, 1–22. doi:10.1080/19415257.2021.1973075.

Cohen, D.K. and Hill, H.C., 1998. *Instructional policy and classroom performance: the mathematics reform in California (RR-39)*. Philadelphia, PA: Consortium for Policy Research in Education. doi:10.1037/e382712004-001.

Copur-Gencturk, Y. and Thacker, I., 2021. A comparison of perceived and observed learning from professional development: relationships among self-reports, direct assessments, and teacher characteristics. *Journal of teacher Education*, 72 (2), 138–151. doi: 10.1177/0022487119899101

Cordingley, P., *et al.*, 2015. *Developing great teaching: lessons from the international reviews into effective professional development*. Teacher Development Trust. Available from: https://tdtrust.org/wp-content/uploads/2015/10/DGT-Full-report.pdf [Accessed 28 Feb 2023]

Darling-Hammond, L., *et al.*, 2009. *Professional learning in the learning profession* 12. Washington, DC: National Staff Development Council.

Darling-Hammond, L., Hyler, M.E., and Gardner, M., 2017. *Effective teacher professional development*. Palo Alto, CA: Learning Policy Institute. doi:10.54300/122.311.

Davies, P., 2000. The relevance of systematic reviews to educational policy and practice. *Oxford review of Education*, 26 (3–4), 365–378. doi: 10.1080/713688543

Desimone, L.M., *et al.*, 2002. Effects of professional development on teachers' instruction: results from a three-year longitudinal study. *Educational Evaluation and policy analysis*, 24 (2), 81–112. doi:10.3102/01623737024002081.

Desimone, L.M., 2009. Improving impact studies of teachers' professional development: toward better conceptualizations and measures. *Educational Researcher*, 38 (3), 181–199. doi: 10.3102/0013189x08331140

Desimone, L.M., 2023. Rethinking teacher PD: a focus on how to improve student learning. *Professional development in Education*, 49 (1), 1–3. doi: 10.1080/19415257.2023.2162746

Desimone, L.M. and Garet, M.S., 2015. Best practices in teachers' professional development in the United States. *Psychology, Society, & Education*, 7 (3), 252–263. doi:10.25115/psye.v7i3.515.

Dunst, C.J., Bruder, M.B., and Hamby, D.W., 2015. Metasynthesis of in-service professional development research: features associated with positive educator and student outcomes. *Educational research & reviews*, 10 (12), 1731–1744. doi:10.5897/ERR2015.2306.

Festinger, L., 1957. *A theory of cognitive dissonance*. Stanford University Press. doi: 10.1515/9781503620766.

Fischer, C., *et al.*, 2018. Investigating relationships between school context, teacher professional development, teaching practices, and student achievement in response to a nationwide science reform. *Teaching and teacher education*, 72, 107–121. doi:10.1016/j.tate.2018.02.011

Fisher, J.B., *et al.*, 2010. Effects of a computerized professional development program on teacher and student outcomes. *Journal of teacher Education*, 61 (4), 302–312. doi:10.1177/0022487110369556.

Fishman, B., *et al*., 2013. Comparing the impact of online and face-to-face professional development in the context of curriculum implementation. *Journal of teacher Education*, 64 (5), 426–438. doi:10.1177/0022487113494413.

Garet, M.S., *et al*., 2001. What makes professional development effective? Results from a national sample of teachers. *American Educational research journal*, 38 (4), 915–945. doi:10.3102/00028312038004915.

Garet, M.S., *et al.*, 2008. *The impact of two professional development interventions on early reading instruction and achievement (NCEE 2008-4030)*. Washigngton, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M.S., *et al.*, 2011. *Middle school mathematics professional development impact study findings after the second year of implementation (NCEE 2011-4024)*. Washinton, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M.S., *et al.*, 2016. *Focusing on mathematical knowledge: the impact of content-intensive teacher professional development (NCEE 2016-4010)*. Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gersten R., *et al.*, 2014. *Summary of research on the effectiveness of math professional development approaches (REL 2014-74010)*. Washington, D.C.: Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Grigg, J., *et al*., 2013. Effects of the scientific inquiry professional development interventions on teaching practice. *Educational Evaluation and policy analysis*, 35 (1), 38–56. doi:10.3102/0162373712461851.

Guskey, T.R., 2003. Analyzing lists of the characteristics of effective professional development to promote visionary leadership. *NASSP Bulletin*, 87 (637), 4–20. doi: 10.1177/019263650308763702

Hawley, W.D. and Valli, L., 1999. The essentials of effective professional development: a new consensus. *In*: L. Darling-Hammond and G. Sykes, eds. *Teaching as the learning profession: handbook of policy and practice*. San Fransisco, CA: Jossey-Bass, 127–150.

Heller, J.I., *et al*., 2012. Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of research in science teaching*, 49 (3), 333–362. doi:10.1002/tea.21004.

Hill, H.C., Beisiegel, M., and Jacob, R., 2013. Professional development research: consensus, crossroads, and challenges. *Educational Researcher*, 42 (9), 476–487. doi: 10.3102/0013189x13512674

Hill, H.C. and Papay, J.P., 2022) *Building better PL: How to strengthen Teacher learning*. Research Partnership for Professional Learning, Annenberg Institute. Available from: https://annenberg.brown.edu/rppl/what-works [Accessed 28 Feb 2023]

Hill, H.C., Papay, J.P., and Schwartz, N., 2022) *Dispelling the myths: What the research says about Teacher professional learning*. Research Partnership for Professional Learning, Annenberg Institute. Available from: https://annenberg.brown.edu/rppl/dispelling-the-myths [Accessed 28 Feb 2023]

Horn, I. and Garner, B., 2022. *Teacher learning of ambitious and equitable mathematics instruction: a sociocultural approach*. New York, NY: Routledge. doi:10.4324/9781003182214.

Hubers, M.D., Endedijk M, D., and Van Veen, K., 2022. Effective characteristics of professional development programs for science and technology education. *Professional development in Education*, 48 (5), 827–846. doi: 10.1080/19415257.2020.1752289

Jacob, B.A. and Lefgren, L., 2004. The impact of teacher training on student achievement: quasi-experimental evidence from school reform efforts in Chicago. *Journal of human resources*, 39 (1), 50–79. doi:10.3386/w8916.

Kennedy, M.M., 1998. The role of preservice teacher education. *In*: L. Darling-Hammond and G. Sykes, eds. *Teaching as the learning profession: handbook of policy and practice*. San Francisco, CA: Jossey Bass, 54–85.

Kennedy, M.M., 2016. How does professional development improve teaching? *Review of Educational research*, 86 (4), 945–980. doi: 10.3102/0034654315626800

Kirschner, P.A. and van Merriënboer, J.J.G., 2013. Do learners really know best? Urban legends in education. *Educational Psychologist*, 48 (3), 169–183. doi:10.1080/00461520.2013.804395.

Kraft, M.A., Blazar, D., and Hogan, D., 2018. The effect of teaching coaching on instruction and achievement: a meta-analysis of the causal evidence. *Review of Educational research*, 88 (4), 547–588. doi:10.3102/0034654318759268.

Kraft, M.A. and Papay, J.P., 2014. Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and policy analysis*, 36 (4), 476–500. doi: 10.3102/0162373713519496

Landry, S.H., *et al.*, 2009. Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational psychology*, 101 (2), 448–465. doi:10.1037/a0013842.

Lynch, K., *et al.*, 2019. Strengthening the research base that informs STEM instructional improvement efforts: a meta-analysis. *Educational Evaluation and policy analysis*, 41 (3), 260–293. doi:10.3102/0162373719849044.

Maandag, D.W., *et al.*, 2017. *Features of effective professional development interventions in different stages of teacher's careers. A review of empirical evidence and underlying theory*. Groningen: Lerarenopleiding Rijksuniversiteit Groningen.

National Academies of Sciences, Engineering and Medicine, 2020. *Changing expectations for the K-12 teacher workforce: policies, Preservice Education, professional development, and the workplace*. Washington, DC: The National Academies Press. doi:10.17226/25603

Opfer, V.D. and Pedder, D., 2011. Conceptualizing teacher professional learning. *Review of educational research*, 81 (3), 376–407. doi:10.3102/0034654311413609.

Osborne, J.F., *et al.*, 2019. Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational research journal*, 56 (4), 1067–1112. doi:10.3102/0002831218812059.

Patfield, S., Gore, J., and Harris, J., 2021. Shifting the focus of research on effective professional development: insights from a case study of implementation. *Journal of Educational change*, 24 (2), 345–363. doi:10.1007/s10833-021-09446-y.

Penuel, W.R., *et al.*, 2007. What makes professional development effective? Strategies that foster curriculum implementation. *American Educational research journal*, 44 (4), 921–958. doi:10.3102/0002831207308221.

Penuel, W.R., Gallagher, L.P., and Moorthy, S., 2011. Preparing teachers to design sequences of instruction in earth systems science: a comparison of three professional development programs. *American Educational research journal*, 48 (4), 996–1025. doi: 10.3102/0002831211410864

Piasta, S.B., *et al.*, 2017. Effectiveness of large-scale, state-sponsored language and literacy professional development on early childhood educator outcomes. *Journal of research on Educational effectiveness*, 10 (2), 354–378. doi:10.1080/19345747.2016.1270378.

Powell, D.R., *et al.*, 2010. Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational psychology*, 102 (2), 299–312. doi:10.1037/a0017763.

Ronfeldt, M., *et al.*, 2015. Teacher collaboration in instructional teams and student achievement. *American Educational research journal*, 52 (3), 475–514. doi:10.3102/0002831215585562.

Roth, K.J., *et al.*, 2011. Videobased lesson analysis: effective science PD for teacher and student learning. *Journal of research in science teaching*, 48 (2), 117–148. doi:10.1002/tea.20408.

Russell, M., *et al.*, 2009. Face-to-face and online professional development for mathematics teachers: a comparative study. *Journal of asynchronous learning networks*, 13 (2), 71–87. doi:10.24059/olj.v13i2.1669.

Scher, L. and O'Reilly, F., 2009. Professional development for K–12 math and science teachers: what do we really know? *Journal of research on Educational effectiveness*, 2 (3), 209–249. doi: 10.1080/19345740802641527

Sims, S. and Fletcher-Wood, H., 2021. Identifying the characteristics of effective teacher professional development: a critical review. *School effectiveness and school improvement*, 32 (1), 47–63. doi: 10.1080/09243453.2020.1772841

Strom, K.J. and Viesca, K.M., 2021. Towards a complex framework of teacher learning-practice. *Professional development in Education*, 47 (2–3), 209–224. doi: 10.1080/19415257.2020.1827449

Taylor, J.A., *et al.*, 2017. The effect of an analysis-of-practice, videocase-based, teacher professional development program on elementary students' science achievement. *Journal of research on Educational effectiveness*, 10 (2), 241–271. doi:10.1080/19345747.2016.1147628.

Timperley, H., *et al.*, 2007. *Teacher professional learning and development. Best evidence synthesis iteration (BES)*. Wellington, New Zealand: Ministry of Education.

TNTP, 2015. *The mirage: confronting the hard truth about our quest for teacher development*. Brooklyn, NY: TNTP.

Van Driel, J.H. *et al.*, 2012. Current trends and missing links in studies on teacher professional development in science education: a review of design features and quality of research. *Studies in science education*, 48 (2), 129–160. doi:10.1080/03057267.2012.738020.

Van Veen, K., Zwart, R., and Meirink, J., 2012. What makes teacher professional development effective? A literature review. *In*: M. Kooy and K. Van Veen, eds. *Teacher learning that matters*. New York, N.Y.: Routledge, 23–41.

Vernon-Feagans, *et al.*, 2015. The targeted reading intervention: Face-to-face vs. webcam literacy coaching of classroom teachers. *Learning Disabilities research and practice*, 30 (3), 135–147. doi:10.1111/ldrp.12062.

Walter, C. and Briggs, J., 2012. What professional development makes the most difference to teachers? A report sponsored by Oxford University Press. Available from: https://clie.org.uk/wp-content/uploads/2011/10/Walter_Briggs_2012.pdf

Wayne, A.J., *et al.*, 2008. Experimenting with teacher professional development: motives and methods. *Educational Researcher*, 37 (8), 469–479. doi:10.3102/0013189x08327154.

Webb, N.M. 2009. The teacher's role in promoting collaborative dialogue in the classroom. *The British journal of educational psychology*, 79 (1), 1–28.

Yang, R., *et al.*, 2020. Curriculum-based teacher professional development in middle school science: a comparison of training focused on cognitive science principles versus content knowledge. *Journal of research in science teaching*, 57 (4), 536–566. doi:10.1002/tea.21605.

Yoon, K.S., *et al.*, 2007. *Reviewing the evidence on how teacher professional development affects student achievement (REL 2007-033)*. Washington, DC: Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest.