Characteristics of effective teacher professional development: what we know, what we don't, how we can find out

Sam Sims[1], Harry Fletcher-Wood[2]

Abstract

Several influential reviews and one meta-review have converged on the position that teacher professional development (PD) is more effective when it is: sustained, collaborative, subject-specific, draws on external expertise, has buy-in from teachers and is practice-based. This consensus view has now been incorporated in government policy and official guidance in several countries. Despite this, several recent PD programmes incorporating these characteristics have failed to have any detectable impact on pupil attainment. This article reviews the evidence underpinning the consensus, arguing that the reviews on which it based are methodologically flawed because they employ inappropriate exclusion criteria and depend on an invalid inference method. The consensus view is therefore likely to be inaccurate. Researchers would make more progress on identifying characteristics of effective professional development by looking for alignment between evidence from basic research on human skill acquisition and features of rigorously-evaluated PD interventions.

Keywords

Professional Development, Teachers

Corresponding Author: Sam Sims, UCL Institute of Education, 20 Bedford Way, London.
Email: s.sims@ucl.ac.uk

[1] UCL Institute of Education
[2] Institute for Teaching

1. Introduction

International surveys suggest that the average teacher spends 10.5 days per year engaged in courses, workshops, conferences, seminars, observation visits or in-service training for the purposes of continuing professional development (Sellen, 2016). In countries such as Mexico, Brazil and Shanghai, teachers report spending between 24 and 40 days per annum on such professional development (PD). The motivation for this substantial investment in PD is clear: improved pupil attainment is associated with improvements in income, happiness and health (Lance, 2011) and improved teacher quality has a relatively strong relationship with improved pupil attainment (Hanushek, 2011; Chetty *et al*., 2014). How this time should be spent however, is somewhat less clear. While research has identified a few programmes or interventions for which there is persuasive evidence of impact on pupil attainment (e.g. Allen *et al*., 2011; Allen *et al*., 2015), most leaders of professional development do not have access to these programmes due to either cost or location. School leaders and teacher educators need instead to know which *characteristics* of professional development are important (Hill *et al*., 2013) to help them design or commission such PD for their own schools.

In recent years, a number of influential reviews have converged on the position that PD is more likely to improve pupil attainment if it is sustained, collaborative, has teacher buy-in, is subject-specific, draws on external expertise and is practice-based (Timperley *et al*., 2007; Wei et al., 2009; Desimone, 2009; Walter, 2012). The conclusions of these reviews has been explicitly referred to as a consensus by several authors (Van Driel *et al*., 2012; Wei *et al*., 2009; Desimone, 2009; Caena, 2011). The findings of such reviews have themselves recently been summarised in a meta-review (Cordingley *et al*., 2015) which, among other things, provides a particularly clear statement of the consensus view. Indeed, this position has become so widely accepted that it has been incorporated into government policy and official guidance for teachers in the UK, US and EU (see DfE, 2016; Menter, 2010; Caena, 2011; Desimone, 2009; Wei, 2009; Combs & Silverman, 2016). It has also begun to influence the way in which research on PD is designed and conducted (e.g. Desimone, 2009; Rutkowski *et al*., 2013; Penuel *et al*., 2007).

In this review paper, we argue that this consensus view is based on flawed methodological foundations and is likely to be misleading. Our argument begins with the observation that three recent programmes which incorporate many or all of the characteristics recommended

by the consensus view have had no detectable effect on pupil attainment. We therefore reconsider the consensus, investigating the methods employed in the papers on which the consensus view is based. We show that all but one of the reviews employ inadequate inclusion criteria, drawing on studies which are inappropriate for supporting the conclusions they reach. Moreover, we argue that, even where such reviews employ appropriate inclusion criteria, the inference process used to identify characteristics of effective professional development is logically flawed, because it provides no way of distinguishing the 'active ingredients' of such programmes from the causally redundant features which have no effect on teachers' practice and/or pupil learning. This offers one plausible explanation for the ineffectiveness of some PD programmes designed around the consensus view.

The second part of the paper considers alternative methods by which we could identify the characteristics of effective professional development. Instead of simply seeking recurring features of effective professional development, we argue that it is necessary to look for both 1) evidence of correlation between specific interventions and pupil attainment *and* 2) evidence of mechanisms from basic research (the study of fundamental processes of human learning or behaviour) which can explain why and how the characteristics of these interventions work. When combined, these two types of evidence are greater than the sum of their parts (Clarke *et al*., 2014). Evidence from well-designed evaluations cannot tell us about the effectiveness of specific features of a PD intervention, because any individual feature of an effective PD intervention may be causally redundant. Conversely, evidence of mechanism from basic research on its own cannot tell us whether a given characteristic will be effective when embedded as one component amongst many in a particular PD intervention. However, where a feature of professional development finds support from both basic research and evaluations of specific interventions, there is greater warrant for concluding that it is indeed characteristic of effective PD. We illustrate our proposed approach with reference to the literature on instructional coaching. The paper concludes with a discussion of the implications of our argument for the design of PD, focused on aspects of the consensus view which we believe are inaccurate and misleading and should therefore be revised.

Our paper is not the first to criticise the consensus view. Kennedy (2016), for example, has shown that stricter inclusion criteria lead to different conclusions about the relationships between consensus view characteristics and pupil attainment. However, Kennedy still attempts to draw conclusions about the characteristics of effective professional development by looking for recurring features of effective interventions. Our line of reasoning suggests

that this approach is likely to lead to erroneous conclusions because it does not incorporate a method for distinguishing the active ingredients of these interventions from the causally redundant components. Kennedy (2016), Opfer & Pedder (2011) and Sztjan et al. (2011) have all previously called for better integration of empirical findings with theoretical insights, providing typologies of teaching or school systems around which reviews should be organised. We extend and formalise this line of reasoning by explicitly stating which types of theory (evidence of mechanism) can help in this respect and make explicit the way in which this combines with evidence of correlation to help isolate characteristics of effective professional development. This allows us to make positive claims about which parts of the consensus view should be retained, as well as identifying and critiquing the parts which should be dropped. Our paper therefore goes beyond the existing literature in several ways.

2. The Consensus View on Characteristics of Effective Professional Development

Several researchers have identified an apparent consensus (Van Driel *et al*., 2012; Wei *et al*., 2009; Desimone, 2009; Caena, 2011) regarding the characteristics of effective PD. This consensus has grown from several influential literature reviews which have reached similar conclusions; although it is important to note that each review reaches slightly different conclusions and does more than endorse the five principles set out above (Timperley *et al*., 2007; Wei *et al*., 2009; Desimone, 2009; Walter, 2012; Cordingley *et al*., 2015). Briefly, the consensus view is that PD which is sustained, collaborative, has buy-in from teachers and school leaders, is subject-specific, draws on external expertise and is practice-based is more effective than PD which is not. Although there is some disagreement at the margins between these reviews, they overlap to a great extent.

Sustained: PD is claimed to be more effective if it is sustained over time (Timperley *et al*., 2007; Blank & Alas, 2009; Wei *et al*., 2009; Desimone, 2009; Walter, 2012). Some of the reviews develop this point by claiming that PD should be organised in a cycle or rhythm in which the content is revisited or iteratively developed. The justification for this is usually that it takes time for teachers to assimilate new knowledge or practise new techniques. By contrast, the single, one-day session is often cited as being particularly ineffective.

Collaborative: PD is claimed to be more effective if teachers take part in it as a group (Timperley *et al*., 2007; Wei *et al*., 2009; Desimone, 2009; Walter, 2012). Most often the requirement for collaboration is formulated as the need to work with multiple peers or 'community of practice'. The justification for this is usually that it gives teachers the chance to challenge each other and clarify misunderstandings. The transfer of information directly from a course leader to an individual participant is often contrasted as being particularly ineffective.

Buy-in: PD is claimed to be more effective if teachers identify with and endorse taking part in it (Timperley *et al*., 2007; Walter, 2012). This is often framed as the claim that voluntary PD is more effective than obligatory PD. However, some researchers make the more nuanced point that there can be strong buy-in for obligatory PD if the purpose and benefits of the PD are clearly explained to participants, so that they can see the value of taking part (Timperley *et al*., 2007).

Subject specific: PD is claimed to be more effective when it involves training in subject knowledge (Blank & Alas, 2009; Wei *et al*., 2009; Desimone, 2009). This is often contrasted with PD that only involves training in general pedagogical techniques, divorced from the content that they would be used to deliver. Indeed, the two are often argued to be complementary and PD is therefore most effective when *both* training on subject knowledge and general pedagogical techniques are delivered together.

Outside expertise: PD is claimed to be more effective when it involves outside expertise (Timperley *et al*., 2007; Wei *et al*., 2009; Walter, 2012). In general, outside expertise is used to mean input from people that do not work in the same school as the teachers receiving the training. The justification for this is generally that this is needed to provide challenge or fresh input, as opposed to recycling existing expertise from inside the school, with which teachers may already be familiar.

Practice/application: PD is claimed to be more effective when it involves opportunities to use, practise or apply what has been learned (Timperley *et al*., 2007; Blank & Alas, 2009; Wei *et al*., 2009; Desimone, 2009; Walter, 2012). Again, the justification for this is often that it helps teachers apply what they have learned in real classroom situations. This approach is often contrasted with lectures in which teachers receive new information passively, but do not apply it.

The precise nature of the claims being made about these six characteristics is not always clear. They are sometimes conceptualised as necessary or sufficient conditions (e.g. Cordingley *et al*., 2015), sometimes as critical features (e.g. Desimone, 2009) suggesting that they must be present for PD to be effective, and sometimes simply as important (e.g. Timperley *et al.*, 2007). Generally, no claims are made about the relative importance of different characteristics, which means that a programme containing more of the six characteristics cannot necessarily be assumed to be better than one containing fewer. In any case, it is generally implicit that each of these characteristics is a *good thing*.

The apparent consensus developed through these reviews has directly influenced and become embedded in official guidance in the UK (DfE, 2016; Menter, 2010) and the EU (Caena, 2011). It has also influenced federal and state policy in the US (see Desimone, 2009) including in the Every Student Succeeds Act, which requires professional development to be sustained, collaborative and practice-based in order to attract federal funding (see Combs & Silverman, 2016). Checklists and other "tools" have been created for educators to check whether their PD sessions conform with the consensus view (Wei *et al*., 2009; Main & Pendergast, 2015). Calls have also been made for research on PD to use the consensus view as a common organising framework for data collection and analysis (Desimone, 2009). To some extent, this has begun to happen, as the consensus view can clearly also be seen reflected in the design of questionnaires for teacher surveys (e.g. Rutkowski *et al*., 2013; Penuel *et al*., 2007).

Indeed, there have now been so many reviews of the literature on what constitutes effective PD that a meta-review of these reviews has recently been conducted (Cordingley *et al*., 2015). This meta-review offers much beside articulating this consensus, but it is perhaps the clearest expression of the consensus view we have found and endorses all six of the consensus principles. It has directly informed the development of England's recently-issued Standards for Teacher Professional Development: these provide recommendations about how post-qualification teacher training should be conducted "to ensure effective professional development" based on the "best available research" (DfE 2016, p.3-4). Because this meta-review contains a particularly clear and complete expression of the consensus view and has led directly to the implementation of official government guidance, we conduct a detailed analysis of the evidence underpinning it in Section 4 of this article. First, however, we review three recent empirical studies which suggest limitations of the consensus view in practice.

3. Empirical problems with the consensus view

In this section, we discuss three rigorous, well-designed experimental studies of PD interventions which incorporated all, or almost all, of the characteristics of professional development recommended by the consensus view, but did not find positive effects on pupil attainment. Reviewing three studies does not demonstrate that the consensus view is incorrect; rather, it illustrates its limitations and motivates the methodological examination of the consensus view that follows.

Garet et al. (2016) designed a yearlong PD programme which incorporated all six characteristics recommended by the consensus view. The programme was sustained: it offered eighty hours training during a summer workshop, followed by five two-hour meetings and three individual coaching sessions during the following year. It was collaborative: the summer training included opportunities for teachers to "solve mathematics problems or analyze examples of student work", individually and in small groups", while meetings during the school year were collaborative Mathematics Learning Communities, analysing student work with colleagues (p. 21). Participants had bought into the programme: both districts and teachers participated voluntarily (p. 7). The programme focused on subject knowledge: it had an "intensive and explicit focus on improving teachers' conceptual understanding of mathematics" and used collaborative meetings and coaching to "help teachers enact their mathematical knowledge in the classroom" (p. 2). The programmes were led by outside experts: Intel Math was delivered by facilitator pairs – a mathematician and a maths educator – all of whom were highly qualified and had several years' experience delivering the content (p. 21). Collaborative meetings and coaching were also led by external, highly-experienced facilitators (pp. 23-24). Teachers had extensive opportunities to practise and apply their work: Intel Math allowed "extended time for teachers to solve math problems, analyze student work, explain their solutions to math problems, share their analyses of student work, and receive feedback (p. ES-5)", while the Mathematics Learning Communities include problem-solving, analysis and discussion of student work and reflection on teachers' own learning (p. 23).

The study was implemented as intended: the treatment group received ninety-five hours more maths-related PD than the control group over the year; sessions were assessed against

measures of Mathematical Quality of Instruction and scored highly. Teacher participation in sessions was deemed to have been high. The results showed strong, significant effects on teachers' behaviour: they showed much higher 'Richness of Mathematics' in lessons. However, there were no significant effects on students' participation and students in the treatment group showed weaker achievement on state tests than the control group.

Garet at al. (2008) tested a second-grade reading PD intervention which included five of the six characteristics of the consensus view. The programme was sustained: teachers in Group A received either eight days of training across several months, those in Group B received the same training and sixty hours coaching across the year. The training days involved some teacher discussion and activities, while the coaching received by Group B was designed to "increase teachers' understanding of the content learned in the institute series and to provide ongoing practice and support for applying their new knowledge and implementing their core reading program effectively (p. xvii)." The programme focused on content knowledge: it was based on the "Language Essentials for Teachers of Reading and Spelling" programme which aligns with the "essential components of reading instruction (p.5)." External experts were commissioned: all the training days were led by trainers who had been certified specifically to facilitate them and mentored by the training designer; trainers had between six and fifteen years' experience training teachers. Coaches were selected for their knowledge and experience; they attended all the teachers' training days, a three-day coaching institute and four on-site follow up training sessions. The only characteristic which may be missing from the consensus view was buy-in: schools were randomly-selected to participate, although teachers were described as having been "invited" to attend summer training (p.28).

The authors reported the programme was implemented as intended, with the intervention group teachers receiving 93 percent of the planned training time, and teachers attending 78% of the sessions; teachers in both groups received substantially more PD than those in the control group: (39 hours, in Group A, 47 hours in Group B, compared to 13 hours in the control); while teachers receiving coaching (Group B) received an average of 62 hours of coaching the year (2 hours per week); almost all of which time was spent on the topics were spent on topics which were the focus of the PD. Teachers in both groups were found to know more about teaching reading and to have adopted one of the three teaching behaviours promoted more (explicit instruction). However, the additional coaching did not affect how much teachers' behaviours changed, and the changes in teacher knowledge and behaviour did not lead to significantly improved student learning among either group. Moreover, the

impact on teachers' knowledge and practices disappeared when researchers returned the following year.

In a third study, Jacob *et al.* (2017) studied the Math Solutions programme, choosing it "because it meets the criteria articulated in Desimone's (2009) description of effective professional development program features" (p. 380). The authors therefore detailed how the programme was subject-focused, sustained, collaborative, and incorporated outside expertise. It was:

> "Specifically intended to improve teachers' mathematical knowledge for teaching and to enable teachers to elicit more student thinking and reasoning in classroom instruction. The activities in which teachers engaged were all designed to have an active learning component. Math Solutions staff worked with the district to design that content of the PD to help ensure coherence with the districts priorities… The program was of a sufficient duration (over 40 contact hours spread out over the course of a year) and was offered over a 3 year period of time to allow sufficient time for change, and involved participation of groups of teachers from the same grade, subject and school" (p. 383).

Buy-in was obtained from school leadership by only partnering with schools offering "strong leadership support for the professional development" (p. 383); evidence of buy-in from teachers is less clear. The study led to some slight increases in teachers' mathematical knowledge for teaching, but scores for the Mathematical Quality of Instruction of their teaching "did not increase, and in many cases decreased" (p. 401). Student participation in mathematical reasoning and achievement also did not increase.

One might object that the design and implementation of these trials limited their ability to demonstrate impact. The studies used an intent-to-treat approach: if a teacher was allocated to the intervention group, their students' outcome data was analysed whether-or-not the teacher actually received the intervention. With such designs, high levels of non-compliance can dilute treatment effects and leave a trial underpowered to detect impact, a concern raised by Garet *et al.* (2011). Alternatively, high teacher turnover may have introduced bias or reduced the power of the trials. Jacob *et al*. (2017), for example, noted that impact estimates were robust for the study's first year but more doubtful in subsequent years as teachers moved schools. Other authors have suggested that the failure of these interventions to have a detectable effect on attainment may be due to poor implementation (Darling-Hammond *et al*.,

2017). It is possible that these arguments are sufficient to account for the null findings in the three studies reviewed in this section. However, repeated failed attempts to identify any impact from trials designed around the consensus view pose questions about that view. Therefore, we examine the origins of the consensus view.

4. Methodological problems with the consensus view: inclusion criteria

The consensus view is based on several literature reviews. Each has followed common steps. First, researchers have searched the literature, more or less systematically, to identify research on PD in schools. In particular, they have searched for published articles which have evaluated (in some way) specific PD interventions. Second, once a long-list of such articles has been identified, inclusion criteria have been used to remove articles of low quality or relevance to the research question. Third, researchers have sorted these articles into those that find the intervention they evaluate has had a positive impact, and those that did not. The fourth and final step has been to look for characteristics of PD which are (in some way) related to the effective interventions. The meta-review by Cordingley et al. (2015) then synthesised several of these reviews. As with all research, the validity of the conclusions of these (meta) reviews, and the consensus view which rests upon them, depends on the validity of the methods by which the reviews were conducted. In this section, we focus on one point in particular: the inclusion criteria used to include or exclude studies in step two. In the next section of this article, we discuss the inference process used in step four.

The specific selection criteria employed in a literature review can affect the conclusions of that review (McDonagh *et al.*, 2013) for at least two reasons. Firstly and fundamentally, they determine the articles used in the review. Missing important studies will give a partial and potentially inaccurate picture of the evidence. Secondly, the inclusion criteria must exclude studies which cannot answer the research question. In this case, the research question is: what are the characteristics of effective PD? Studies selected should include methods capable of identifying which PD interventions are effective in raising attainment and which are not. If either an incomplete set of studies or the wrong type of studies are included, then the findings of the review will be compromised. Hence, the PRISMA standards for reporting systematic reviews (Liberati *et al.*, 2009) states that "Knowledge of the eligibility criteria is essential in appraising the validity, applicability, and comprehensiveness of a review."

How good are the inclusion criteria in the meta-reviews and reviews on which the consensus view rest? The meta-review by Cordingley et al. (2015) found 980 reviews which were rated on a three-point scale stretching from: 1 - methodology and weighting of evidence clear; 2 – methodology clear but no weighting of evidence; and 3 – methodology unclear. All level 1 and level 2 reviews were retained. No further details were given on how clarity of methodology or weighting were judged for each review. However, Cordingley et al. (2015) do rank the reviews that they use in their meta-review in terms of quality. The review which they give the highest score to is Timperley et al. (2007), which they describe as "the only fully consistent and rigorous review" which they therefore use as "a cornerstone for the umbrella review" (Cordingley *et al*., 2015, p. 4).

In the review by Timperley et al. (2007), quantitative studies are judged on a three-point scale in three areas: sampling methods; control groups; and validity and reliability of test instruments. Qualitative studies were also judged on a three-point scale in three areas: depth of data collection and analysis; validity and reliability of assessment; and method of triangulation. Table 10.2 in Timperley et al. (2007) lists the set of eleven studies relevant to the characteristics of effective PD in secondary schools that were rated highly enough to be included (there is no equivalent section for primary schools). We now briefly review the methods employed by each of these studies. Adey (1999) employed a simple research design in which participants were matched to controls based on age and ability. Anderson (1992) employed an experimental design but only had a sample size of 20, which dropped to 16 through attrition. Bishop et al. (2005), Confrey et al. (2000) and D'oria (2004) employed no control variables at all, relying instead on unadjusted comparisons of outcomes. Huffman et al. (2003) could not be found online. The lead author was contacted to request a copy of the paper but none was forthcoming. Metcalf et al. (2000) employed only basic ANOVA methods to compare group means. Moxon (2003) and Ross (1994) and Ross et al. (1999) both employ before and after designs but neither conduct any covariate adjustment. Schober (1984) does employ regression analysis but only adjusts for a very limited range of variables: degree subject, urban location and average income. Finally, Tasker (2001) only reports qualitative findings.

What can be said about these studies from this brief review? The *What Works Clearing House* in the US has established an explicit set of standards for reviewing the quality of studies which is a three-point scale from: 'Meets Evidence Standards Without Reservation', 'Meets Evidence Standards With Reservation' and 'Does not Meet Evidence Screens'. Nine

of the ten studies reviewed above would be graded Does Not Meet Evidence Screens because they do not establish baseline equivalence of treatment and control groups in terms of outcomes and relevant covariates. The one randomised study might qualify for Meets Evidence Standards Without Reservation, but the small sample size and high rate of attrition means it would likely be disqualified. The *Education Endowment Foundation* in the UK uses a more nuanced system for ranking the quality of studies which spans from five padlocks (most secure) down to zero (least secure), again based on explicit criteria. Assessing the ten studies reviewed above against these criteria we find that: the qualitative articles receive zero padlocks, eight studies qualify for one padlock on the grounds that they have a comparison group but poor or no matching and the randomised controlled trial qualifies for three padlocks due to the low power due to small sample size. In summary, the most highly-rated review in Cordingley et al. (2015), which forms the "cornerstone" of that meta-review, relies on a set of eleven studies, of which at least nine do not meet the What Works Clearing House Standards at all, and score no more than one out of five padlocks when judged against the Education Endowment Foundation standards. The other reviews on which Cordingley et al. (2015) draw are rated by the authors as being of a lower standard than Timperley et al. (2007).

At this point, it might be objected that What Works Clearinghouse and Education Endowment Foundation evidence standards are unduly dismissive of evidence derived from studies that do not include equivalent control groups, i.e. qualitative studies. We do not claim that qualitative studies cannot provide useful insights about PD. We think they can. We only claim that identifying interventions which are and are not effective (step 3 in the literature review process outlined above) requires the use of studies that include equivalent control groups. Since the validity of step 4 is dependent on step 3 accurately identifying interventions which are and are not effective, using studies which do not establish equivalent control groups in step 3 means the conclusions reached in step 4 will likely be inaccurate. The studies that do not establish an equivalent control group, including the qualitative studies, used in step 3 of Cordingley et al. (2015) are therefore inappropriate *for the purpose for which they are employed*.

How do the inclusion criteria in the other reviews compare? We limit ourselves here to cross-subject reviews that look across different types of PD. Wei et al. (2009, p. 3) draw on studies using any methodologies including qualitative and case study methods, but note that "the inferences that can be drawn from such research should be treated as suggestive rather than

conclusive". Desimone (2009) also does not employ any explicit inclusion criteria but includes case study research. Walter and Briggs (2012) include in their review any empirical studies on PD. The standout paper in the field is Yoon et al. (2007) who use the What Works Clearing House standards to screen the papers in their review. They identify nine studies that meet these standards and also examine the common characteristics of the eight studies which show appositive effect on attainment. However, they are careful to warn that "Because of the lack of variability in form and the great variability in duration and intensity across the nine studies, discerning any pattern in these characteristics and their effects on student achievement is difficult" (p. 3). They conclude that more studies would be needed in order to test whether particular characteristics of PD are associated with a larger impact on attainment.

In summary, four of the five cross-subject reviews we looked at (Timperley *et al*., 2007; Wei *et al*., 2009; Desimone, 2009; Walter & Briggs, 2012) employed criteria which included studies inappropriate for accurately distinguishing interventions which did and did not work. This means that step 4 of their reviews may well have come to the wrong conclusions. Indeed, a separate review by Kennedy (2016) which included only experimental studies, does come to noticeably different conclusions. The one cross-subject review which did employ inclusion criteria capable of filtering out studies which were unable to answer the research question implicit in step 3 - Yoon et al. (2007) - concluded that not enough studies met these inclusion criteria to draw any inferences in step 4 about the characteristics of effective PD. We conclude that the consensus view is not supported by existing cross-subject reviews of the characteristics of effective PD. This provides one plausible explanation for the null findings in the three trials reviewed in Section 3.

5. Methodological problems with the consensus view: inference methods

Even if we did have reviews which both employed strong, appropriate inclusion criteria *and* identified a large number of evaluations which met these inclusion criteria, it is unclear that the inference process in step 4 of the review process (looking for common features of effective interventions) outlined in the previous section would yield accurate conclusions about the characteristics of effective PD. In this section, we describe how step 4 was conducted in the five reviews and one meta-review and explain why this is methodologically

flawed. In section 6 of the paper, we use the line of argument developed here to outline an alternative approach to identifying the characteristics of PD.

All four of the cross-subject reviews that conducted step 4 of the review process (Timperley *et al*., 2007; Wei *et al*., 2009; Desimone 2009; Walter & Briggs, 2012) used a thematic approach to analysing the features of the interventions which they identified as effective. That is, they looked for features that recurred among interventions which were found to be effective. For example, Timperley et al. (2007) note that all of the "core studies" which meet all their inclusion criteria involve teachers working in structured professional groups. They acknowledge, however, that some studies involving structured professional groups find neutral or negative impacts for students. These exceptions are taken as evidence that such professional learning communities are necessary but not sufficient for effective PD. Analysis of the counter-example cases is then conducted and it is concluded that the reason the intervention was not effective in this case was because no external expertise was involved. Desimone (2009) also looks for recurring features of successful interventions, adding that such regularities are more persuasive when they come from studies using a range of different research designs. Walter & Briggs (2012) and Wei et al. (2009) also look for recurring themes among effective interventions. The meta-review by Cordingley et al. (2015) then analysed the claims made by eleven different reviews and looked for agreement among them. They therefore look for regularity of claims among reviews which looked for regularity of features of apparently effective interventions.

The inference method described in the preceding paragraph is flawed. The regular occurrence of specific features of PD in effective interventions does not, in itself, warrant any inference about the effect of that feature of the intervention. To use an analogy, toothpaste has many ingredients but many of them would not be classified as *active* ingredients. An epidemiological study of the characteristics of effective toothpaste which used the methodology outlined above would almost certainly conclude that mint flavouring protected teeth from tooth decay, which is clearly incorrect. In the terminology of Mackie (1974), the mint flavouring in toothpaste is a *redundant part* of a set of conditions (regularly brushing human teeth with mint toothpaste containing fluoride) which are *collectively sufficient* for reducing tooth decay.

How likely are the consensus view characteristics of effective PD to be redundant? That is, are there reasons to expect that they would regularly occur in effective PD interventions other

than because they make a causal contribution to the effectiveness of that PD? We would expect mint flavouring to regularly occur in effective toothpaste even if it is causally redundant, for example, because it provides other benefits which consumers wish to purchase in conjunction with a product able to reduce tooth decay, i.e. fresh breath.

There are similar reasons to suspect the redundancy of some of the consensus view characteristics of effective PD. Take the requirement for PD to be collaborative, for example. Schools have limited budgets and are therefore more likely to commission or provide collective, group PD, rather than more expensive one-to-one PD. There is therefore a plausible explanation for collaboration co-occurring with effective PD even if it is causally redundant in those PD interventions. Similarly, consider the requirement for buy-in from teachers. Teachers are likely to be enthusiastic about and willing to participate in an effective PD programme because they have noticed positive impacts as a result of the programme, rather than the programme being effective because teachers have bought into it. This reverse causality provides a plausible explanation for buy-in co-occurring with effective PD even when it is a causally redundant part of the intervention, at least in the first instance.

To summarise, the inference process involved in step 4 of the reviews is likely to yield incorrect inferences about the characteristics of PD, even if (or in the case where) step 3 of the reviews had been properly conducted. This is because the inference process provides no way of distinguishing causally redundant and non-redundant (or active ingredients) of PD interventions. Moreover, there are often plausible reasons that consensus view characteristics of effective PD would regularly occur in effective PD interventions even when they are redundant components of such interventions, which suggests that the inference process employed in these articles would likely lead to incorrect conclusions.

6. Alternative approaches to identifying the characteristics of effective professional development

We began by arguing that school leaders need to be able to identify characteristics of effective PD if they are to design or commission such interventions. But given that such characteristics will always come as part of a package, how can we identify the active ingredients in effective PD? Russo & Williamson (2007) and Clarke et al. (2014) have

revived the arguments of Bradford Hill (1965) and Mackie (1974) to show one important way this can be done. This approach to identifying causes involves combining two types of evidence. The first is evidence of correlation, which they define as probabilistic dependence between two phenomena. An example of this might be epidemiological studies finding increased incidence of lung cancer in smokers. The second is evidence of mechanism, which they define as activities organised in such a way that they are responsible for the phenomenon. An example of this might be observing under a microscope the way in which tar from tobacco smoke creates mutations in cells. In social science, evidence of mechanism might come from basic research describing fundamental characteristics of human motivation or learning, which hold across many diverse contexts.

Clarke et al. (2014) argue that these two types of evidence "integrate in a special way" to become more than the sum of their parts. As we have discussed above, the weakness of correlational evidence is that it cannot distinguish redundant from non-redundant characteristics of interventions. The fact that a PD programme with a collaborative component appears to have an effect on pupil attainment does not guarantee that it was the collaborative component which affected the pupil attainment. The same intervention without the collaborative component might have been just as effective. Conversely, evidence of mechanism can help identify non-redundant components of a cause, but cannot determine whether a component will have a causal effect when implemented as part of an intervention. For example, we may know that people learn effectively from worked examples, but this does not guarantee that any PD intervention incorporating worked examples will improve teaching practice. Only where these two types of evidence converge can we be confident that a non-redundant characteristic of a collectively sufficient causal condition has been identified. If we found a specific PD intervention which had been shown to be effective *and* it incorporated worked examples – which have been shown to improve learning in a range of contexts – then we can be more confident that worked examples are a characteristic of effective PD.

The rest of this section illustrates how we can combine evidence of mechanisms from basic research with evidence of correlation from impact evaluations of PD to identify characteristics of effective professional development. First, we introduce the evidence for the effectiveness of coaching in improving teaching and student achievement. We then discuss evidence for two mechanisms which help explain why coaching works: one drawn from cognitive psychology – the distinction between novices and experts, and one drawn from behavioural psychology – the influence of habits. Evidence of these mechanisms enjoys

many of the indicators of importance detailed by Clarke et al. (2014, Table 2): for example, they have been tested by numerous methods and are reproducible across a range of conditions. We seek to show how these mechanisms help explain the success of coaching programmes in improving student achievement and in doing so, allow us to reach better-justified conclusions about the specific characteristics of these programmes which make them effective. Space limitations mean that our aim here is limited to illustrating briefly how this approach might be used.

PD interventions based on instructional coaching - sustained, one-to-one, deliberate practice with an expert mentor - show impressive effects on both teacher practice and student attainment. A recent meta-analysis, limited to causal studies, identified 44 evaluations of instructional coaching programmes. It found that coaching interventions raised student performance on standardized tests by an average of 0.15 standard deviations (Kraft *et al*., 2016). One notable example is My Teaching Partner (MTP). MTP provides teachers with fortnightly feedback from external observers, allowing them to repeatedly practice specific techniques. The first randomised-controlled trial of MTP in secondary schools found a positive, statistically significant effect on pupil attainment after two years (Allen *et al*., 2011). A second randomised-controlled trial, with a much larger sample, replicated these positive results (Allen *et al*., 2015). MTP is one of many coaching interventions with positive impacts on attainment included in the review.

Interestingly, MTP omits some of the consensus view characteristics of effective PD. For example, it includes only general pedagogical skills, and no subject knowledge. Indeed, it has been shown to be effective across different subjects and moderator analysis found no evidence that it was differentially effective across subjects (Gregory *et al*., 2017; Allen *et al*., 2011, p. 1036; see also Allen *et al*., 2015). Moreover, it involves no collaborative work, relying entirely on dyadic participant-coach interactions. So what are the characteristics of coaching that make it effective?

Our first evidence of mechanism relates to how and when people change their behaviours or practice. Longitudinal studies find that PD programmes often fail to bring about intended changes in teacher practice (Copur-Gencturk & Papakonstantinou, 2016). A meta-analysis of causal studies in a range of settings suggests that habits - behaviours cued automatically by environmental stimuli - are the most important reason that people fail to change their actions in this way (Webb & Sheeran, 2006). People begin with a goal directed behaviours which

gradually, through repetition in the presence of specific environmental cues, become automatic (Lally *et al*., 2009). Research in a very wide range of settings – car use, recycling, blood donation, voting – has shown that people maintain these habitual behaviours, even if their goals change (Wood & Neal, 2007). Further evidence of this mechanism comes from lab research in neuroscience, which has shown how behaviours which are repeated many times become governed by different regions of the brain and become more resistant to change at the same time (Seger & Spiering, 2011).

Coaching incorporates characteristics which are known to promote habit change. Most notably, coaching programmes require teachers to practise in their own classrooms. For example, teachers enrolled in My Teaching Partner submit fortnightly videos of themselves practising specific skills in their own lessons, which they then review along with their coach (Allen *et al.*, 2011). Experimental and observational research in a range of contexts, as well as evidence from neuroscientific research, shows that it is necessary to repeatedly practice new behaviours before they become automatic (see Wood and Neal, 2007). Moreover, meta-analysis suggests that repeatedly practicing the new techniques in the environment where you aim to reproduce them in future (i.e. the classroom) helps replace old habits by overwriting the established cue-response relationships (Webb & Sheeran, 2006). The repeated review and feedback incorporated in coaching models helps strengthen these new cue-response relationships even further. This evidence of mechanism for repeating a new technique in the target environment helps ingrain new practices - combined with evidence of correlation between coaching and pupil attainment - suggests that this type of practice is a characteristic of effective PD.

Our second mechanism relates to the distinction between how novices and experts think and learn. Novices work towards desired solutions, whereas those with more experience tend to have committed the desired solution to memory as a procedure (Larkin *et al*., 1980). Experimental research shows that this means novices' limited working memory can easily be overwhelmed by complex tasks (see Pass & Van Gog, 2006). Models and scaffolding can therefore help novices focus on the important features of a situation and avoid being overwhelmed. Experts, by contrast, benefit from more open problems and can be distracted by the support novices require. A wide range of experimental research shows that they are better able to learn from experience, focusing on what matters most about a situation and gaining new insights in the process (Sweller *et al*., 2003; see also Sternberg & Horvath, 1995; Deans for Impact, 2017). There is no clear point at which a teacher ceases to be a novice and

becomes an expert: attempts to distinguish such transition points in other fields have proved challenging (Kyun *et al*., 2013) and have sometimes identified intermediate stages with their own characteristics (Schmidt & Rikers, 2007; Spiro *et al*., 1988). Nonetheless, treating novices as experts can hamper their learning.

How does our understanding of the differences between novices and experts in general help us understand the effectiveness of specific coaching programmes? Coaching programmes often offer access to models of the sort which are beneficial to novices. For example, Content-Focused Coaching (Matsumura *et al*., 2013), involves coaches modelling specific techniques with a teacher's own class and My Teaching Partner involves an extensive video library exemplifying the use of specific skills. While this explicit form of modelling is most useful for novices, more skilled teacher are likely to benefit from more open-ended discussions of their practice. Because coaching is one-to-one, coaches are able to tailor their support to a teachers' level of expertise, gradually withdrawing the models and scaffolds that help novices learn and focusing instead in e.g. facilitating reflective discussion of specific cases and more open problems (e.g. Campbell & Griffin, 2017). This evidence of mechanism for how modelling needs to be provided for novices and then slowly withdrawn for experts - combined with evidence of correlation from impact evaluations of coaching interventions - suggests that the expertise-appropriate use of modelling is a characteristic of effective PD.

7. Conclusion

In recent years, a number of influential reviews have established an apparent consensus around the characteristics of effective PD. This consensus view has since become embedded in government policy and official guidance for teachers in the US, UK and Europe. It has also been incorporated into the design of PD programmes and education research itself. In this paper, we have argued that the consensus view is based on weak methodological foundations, in particular the use of inappropriate inclusion criteria and flawed inference methods. These shortcomings also help explain why a number of rigorously evaluated PD interventions, which incorporate the characteristics of PD recommended by the consensus view, were found to have no impact on pupil attainment.

Does the consensus need to be abandoned entirely, or simply revised? We conclude that some parts of the consensus view need dropping entirely. For example, as outlined in Section 5,

there are plausible reasons to think that collaboration would regularly occur in effective PD interventions even if it was a causally redundant component of these interventions. Instructional coaching interventions are effective (Kraft *et al.*, 2016) but do not incorporate collaboration between teachers. Moreover, moderator analysis of PD interventions in maths and science have found no correlation between the extent of collaboration and effectiveness (Blank & Alas, 2009). In addition, our discussion of the novice/expert distinction provides an account of why large group PD is unlikely to be effective, since teachers with different levels of skill require different types of professional development. There is therefore currently an absence of evidence for, as well as evidence against, the claim that collaboration is a characteristic of effective professional development.

Other parts of the consensus view, such as being sustained, require revision. Section 6 suggests repeated practice to change ingrained habits is more likely to be effective than the period over which the PD takes place. For example, a sustained PD programme might provide fortnightly sessions for two years, but if each part of the curriculum is covered only once, this is unlikely to change teachers' practice. This is also supported by moderator analysis from two meta-analyses which found that, among interventions which include repeated practice of specific skills, the overall duration of the PD programme shows no relationship with the impact on pupil attainment (Basma & Savage, 2017; Kraft *et al.*, 2016). This difference between these two points is substantively significant.

Other parts of the consensus view, such as requiring that PD be subject-specific, are of unknown value. Based on the arguments made in this paper, there is an absence of evidence about whether, for example, subject-specific CPD is more effective than general PD. The constructive contribution of this paper is to propose a way in which this can be established in future. Researchers should systematically review the literature for alignment between well-evidenced mechanisms and evaluations of specific PD interventions which include these characteristics. For example, a careful consideration of the literature on near- and far-transfer of skills may or may not provide relevant evidence of mechanism to support the claim that subject-specific professional development is more effective. If this sort of evidence can be provided (for any of the consensus view characteristics) then there will be far stronger warrant for them. This may require inter-disciplinary collaboration between e.g. psychologists and educationalists engaged in basic research about the way in which people learn and acquire skill and applied researchers with knowledge of the literature evaluating specific PD programmes. In the meantime, calls to organise research on CPD around the

consensus view (Desimone, 2009) should be resisted, as they may lead research on the characteristics of effective CPD further astray.

This paper has important implications for practice. In the US, the Every Student Succeeds Act currently requires PD to be both sustained and collaborative in order to qualify for federal funding. This paper provides reason to doubt that these are characteristic of effective PD. Policymakers should consider dropping the collaborative criteria and revising the sustained criteria. In the UK, the Standards for Teachers' Professional Development also recommend that PD should be collaborative and sustained (although the latter is couched in terms of allowing cycles of feedback). The PD standards in England contain much that is of value, including the recommendation that PD be practice-based. However, policymakers should consider revising the guidance in light of the evidence set out above. This is necessary in order to avoid spending scarce resources on programmes with questionable effectiveness and to avoid teacher educators redesigning existing programmes in line with the consensus view. Policymakers, school leaders and teacher educators should focus instead on commissioning and designing PD with characteristics for which there is strong evidence of both correlation and mechanism.

References

Adey, P. (1999). The science of thinking, and science for thinking: A description of Cognitive Acceleration through Science Education (CASE). Innodata Monographs 2. Geneva, Switzerland: International Bureau of Education (ED442622).

Allen, J., Pianta, R., Gregory, A., Mikami, A., Lun, J., (2011) An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement. Science. 333 (6045) 1034-1037

Allen, J., Hafen, C., Gregory, A., Mikami, A. and Pianta, R. (2015). Enhancing Secondary School Instruction and Student Achievement: Replication and Extension of the My Teaching Partner-Secondary Intervention. Journal of Research on Educational Effectiveness, 8(4), pp.475-489.

Anderson, V. (1992). A teacher development project in transactional strategy instruction for teachers of severely reading-disabled adolescents. Teaching and Teacher Education, 8 (4), 391-403.

Basma, B., & Savage, R. (2017). Teacher Professional Development and Student Literacy Growth: a Systematic Review and Meta-analysis.

Bishop, R., Berryman, M., Powell, A., & Teddy, L. (2005). Te Kotahitanga: Improving the educational achievement of Maori students in mainstream education. Phase 2: Towards a whole school approach (Progress report and planning document). Wellington, New Zealand: Ministry of Education.

Blank, R. K., & 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.

Bradford Hill, A. (1965). The environment and disease: association or causation? Proceedings of the Royal Society of Medicine, 58:295–300.

Campbell, P. F., & Griffin, M. J. (2017). Reflections on the promise and complexity of mathematics coaching. *The Journal of Mathematical Behavior*, *46*, 163-176.

Caena, F. (2011). Literature review Quality in Teachers' continuing professional development. European Commission Education and Training 2020 Thematic Working Group 'Professional Development of Teachers'.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Clarke, B., Gillies, D., Illari, P., Russo, F., & Williamson, J. (2014). Mechanisms and the evidence hierarchy. *Topoi*, *33*(2), 339-360.

Combs, E. & Silverman, S. (2016) *Bridging the Gap: Paving the pathway from current practice to exemplary professional learning*. Frontline Research & Learning Institute.

Confrey, J., Castro-Filho, J., & Wilhelm, J. (2000). Implementation research as a means of linking systemic reform and applied psychology in mathematics education. Educational Psychologist, 35 (3), 179-191.

Copur-Gencturk, Y., & Papakonstantinou, A. (2016). Sustainable changes in teacher practices: a longitudinal analysis of the classroom practices of high school mathematics teachers. *Journal of Mathematics Teacher Education*, *19*(6), 575-594.

Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B., ... & Coe, R. (2015). *Developing great teaching: lessons from the international reviews into effective professional development*. London: Teacher Development Trust.

Darling-Hammond, L., Hyler, M. E., Gardner, M. (2017). Effective Teacher Professional Development. Palo Alto, CA: Learning Policy Institute.

Deans for Impact (2017) Building Blocks. Austin, TX: Deans for Impact

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational researcher*, *38*(3), 181-199.

DfE [Department for Education. (2016). Standard for teachers' professional development.

D'Oria, T. (2004). How I improved my teaching practice in Grade 9 boys' physical education to increase students' participation and enjoyment. Unpublished Masters thesis, Nipissing University, Canada.

Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., Warner, E., (2011) Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation. Institute of Education Sciences.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Gregory, A., Ruzek, E, Hafen, C., Mikami, A., Allen, J. and Pianta, R. (2017) My Teaching Partner-Secondary: A Video-Based Coaching Model, Theory Into Practice, 56:1, 38-45.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3), 466-479.

Hill, H., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. Educational Researcher, 42(9), pp.476-487.

Huffman, D., Goldberg, F., & Michelin, M. (2003). Using computers to create constructivist learning environments: Impact on pedagogy and achievement. Journal of Computers in Mathematics and Science Teaching, 22 (2), 153-170.

Jacob, R., Hill, H., Corey, D. (2017) The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement. Journal of Research on Educational Effectiveness.

Kennedy, M. (2016). How Does Professional Development Improve Teaching?. Review of Educational Research, 86(4), pp.945-980.

Kraft, M.A., Blazar, D., Hogan, D. (2016). The effect of teaching coaching on instruction and achievement: A meta-analysis of the causal evidence. Brown University Working Paper.

Kyun, S. Kalyuga, S. and Sweller, J. (2013) The Effect of Worked Examples When Learning to Write Essays in English Literature, The Journal of Experimental Education, 81:3, 385-408.

Lally, P., van Jaarsveld, C., Potts, H. and Wardle, J. (2009). How are habits formed: Modelling habit formation in the real world. European Journal of Social Psychology, 40(6), pp.998-1009.

Lance, L. (2011). Non-production benefits of education: crime, health, and good citizenship. In E.A. Hanushek, S. Machin and L. Woessman (eds), Handbook of the economics of education (pp. 183–282). The Netherlands: Elsevier B.V.

Larkin, J., McDermott, J., Simon, D., Simon, H. (1980) Expert and Novice Performance in Solving Physics Problems. Science 208(4450):1335-42.

Liberati, A., Moher, D., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, *6*(7), e1000097.

Mackie, J. L. (1974). The cement of the universe. *London: Oxford Uni*.

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2012). The effect of content-focused coaching on the quality of classroom text discussions. *Journal of Teacher Education*, *63*(3), 214-228.

Main, K., & Pendergast, D. (2015). Core Features of Effective Continuing Professional Development for the Middle Years: A Tool for Reflection. *RMLE Online, 38*(10), 1–18.

McDonagh, M., Peterson, K., Raina, P., Chang, S., & Shekelle, P. (2013). Avoiding bias in selecting studies. Methods Guide for Effectiveness and Comparative Effectiveness Reviews [Internet].

Menter, I. et al. (2010) Literature Review on Teacher Education in the 21st Century. Edinburgh: The Scottish Government.

Metcalf, K. K., Vontz, T. S., & Patrick, J. J. (2000). Effects of Project Citizen on the civic development of adolescent students in Indiana, Latvia, and Lithuania. In T. S. Vontz & K. K.

Metcalf & J. J. Patrick (2000), "Project Citizen" and the civic development of adolescent students in Indiana, Latvia, and Lithuania (pp. 125-146). Bloomington, IN: ERIC Clearinghouse for Social Studies/Social Science Education.

Moxon, J. (2003). A study of the impact of the Restorative Thinking Programme within the context of a large multi- cultural New Zealand secondary school. MA (Education) thesis, Auckland, NZ.

Opfer, D., & Pedder, D. (2011). Conceptualizing teacher professional learning. Review of Educational Research, 81, 376–407.

Paas, F., & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, 16(2), 87-91.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*(4), 921-958.

Renkl, A., Hilbert, T. & Schworm, S. (2009). Example-Based Learning in Heuristic Domains: A Cognitive Load Theory Account. Educational Psychology Review, 21(67).

Ross, J. A. (1994). The impact of an inservice to promote cooperative learning on the stability of teacher efficacy. Teaching and Teacher Education, 10 (4), 381-394.

Ross, J. A., Roleiser, C., & Hogaboam-Gray (1999). Effects of collaborative action research on the knowledge of five Canadian teacher-researchers. The Elementary School Journal, 99 (3), 255-274.

Russo, F. and Williamson, J. (2007). Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Rutkowski, S., Rutkowski, L., Bélanger, J., Knoll, S… (2013) Teaching and Learning International Survey 2013 Conceptual Framework. OCED Publishing.

Schmidt, H. and Rikers, R. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. Medical Education, 41, pp.1133–1139.

Schober, H. M. (1984). The effects of inservice training on participating teachers and students in their economics classes. The Journal of Economic Education, 15 (4), 282-295.

Seger, C., & Spiering, B (2011). A critical review of habit learning and the Basal Ganglia. *Frontiers in Systems Neuroscience*, 5, 66.

Sellen, P. (2016). *Teacher workload and professional development in England's secondary schools: insights from TALIS*. London: Education Policy Institute.

Spiro, R. Coulson, R., Feltovich, P. and Anderson, D. (1988). Cognitive Flexibility Theory: Advanced Knowledge Acquisition in Ill-Structured Domains. Technical Report No. 441. Urbana, Ill: Center for the Study of Reading.

Sztjan, P., Campbell, M. P., & Yoon, K. S. (2011). Conceptualizing professional devel- opment in mathematics: Elements of a model. PNA, 5(3), 83–92.

Sternberg, R., Horvath, J. (1995) A Prototype View of Expert Teaching. Educational Researcher 24(6) 9-17

Sweller, J., Ayres, P. L., Kalyuga, S. & Chandler, P. A. (2003). The expertise reversal effect. Educational Psychologist, 38 (1), 23-31.

Tasker, G. (2001). Students' experience in an HIV/AIDS-sexuality education programme: What they learnt and the implications for teaching and learning in health education. Unpublished doctoral thesis, Victoria University of Wellington, Wellington, New Zealand.

Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Auckland: NZ Ministry of Education.

Valente, T. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18(1), pp.69-89.

van Driel, J. H., Meirink, J. A., van Veen, K., & Zwart, R. C. (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.

Walter, C., & Briggs, J. (2012). What professional development makes the most difference to teachers. *A report sponsored by Oxford University Press. Retrieved on July*, *20*, 2015.

Webb, T., Sheeran, P. (2006) Does Changing Behavioral Intentions Engender Behavior Change? A Meta-Analysis of the Experimental Evidence. Psychological Bulletin.

Wei, R. C., Darling-Hammond, L., Andree, A., Richardson, N., & Orphanos, S. (2009). Professional learning in the learning profession. *Washington, DC: National Staff Development Council*.

Wood, W., Neal, D., (2007) A New Look at Habits and the Habit–Goal Interface. Psychological Review 114(4) 843–863.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.