Level models of continuing professional development evaluation: a grounded review and critique

COLDWELL, Mike <http://orcid.org/0000-0002-7385-3077> and SIMKINS, Tim

Published version

COLDWELL, Mike and SIMKINS, Tim (2011). Level models of continuing professional development evaluation: a grounded review and critique. Professional development in education, 37 (1), 143-157.

Level models of CPD evaluation: a grounded review and critique

Mike Coldwell* and Tim Simkins

Centre for Education and Inclusion Research, Sheffield Hallam University, Sheffield, United Kingdom

*Mike Coldwell, Centre for Education and Inclusion Research, Unit 7, Science Park, Sheffield Hallam University, Howard Street, Sheffield S1 1WB, UK. Email: m.r.coldwell@shu.ac.uk

Continuing professional development (CPD) evaluation in education has been heavily influenced by' level models', deriving from the work of Kirkpatrick and Guskey in particular, which attempt to trace the processes through which CPD interventions achieve outcomes. This paper considers the strengths and limitations of such models, and in particular, the degree to which they able to do justice to the complexity of CPD and its effects. After placing level models within the broader context of debates about CPD evaluation, the paper reports our experience of developing such models heuristically for our own evaluation practice. It then draws on positivist, realist and constructivist traditions to consider some more fundamental ontological and epistemological questions to which they give rise. The paper concludes that level models can be used in number of ways and with differing emphases, and that choices made about their use will need to reflect both theoretical choices and practical considerations.

Keywords: professional development; evaluation; level models; evaluation theory

Introduction

The evaluation of continuing professional development (CPD) in education provides major practical challenges to those commissioning such evaluations, those undertaking them and those who use them. Underlying these challenges is a further one: theorising the nature and effects of CPD in ways which both do justice to the complexity of the CPD world and generate practical possibilities for programme evaluation. Our judgement is that this issue has been addressed at best unevenly. Many attempts have been made to theorise CPD but few of these seem to have influenced evaluators and where evaluation of CPD has been theory-based, such theories are often implicit, ill-specified or overly reductive. This is despite the enormous literature that exists on policy and programme evaluation generally.

The purpose of this paper is to begin to address this issue by focusing on what are often called 'level' models for evaluating development and training. Such models

draw on the hugely influential work of Kirkpatrick and Guskey, and the ideas of these writers have helped to inform much of our own work of evaluating a range of CPD (including, especially, leadership development) programmes for a number of government agencies in England. Our experience has been an evolutionary one, with a constant interplay between our theorising and the practicalities of delivering evaluations on time and to budget. This paper tries to reflect this, by locating our thinking both temporally in terms of our own learning and in relation to evaluation models developed by others. The aims of the paper are threefold: first, to consider the ways in which level models have been articulated and critiqued; second, to explain how our own evaluation work has been influenced by these models and critiques; and third, to stand back from these models' use in practice to consider some more fundamental ontological and epistemological questions to which they give rise but which are not often discussed in evaluation reports. We conclude that the complexity of CPD processes and effects and, crucially, of the social world requires a range of approaches, and that – therefore – an approach based on any single model is not enough.

Approaches to evaluation

Many attempts have been made to categorise different approaches, theories or models of evaluation. Some go back to early seminal contributors to the field (e.g. House 1978, Stufflebeam and Webster 1980, Stake 1986, Guba and Lincoln 1989); others are more recent (e.g. Alkin 2004, Hansen, 2005, Stufflebeam and Shinkfield 2007). Such classifications vary widely in their focus and underpinning rationales. At the risk of oversimplifying a complex and ever-growing field, it is useful to distinguish among three inter-related dimensions of the evaluation 'problem'. These concern respectively the 'what', the 'how' and the 'who' of evaluation processes. In relation to the 'what', a

core distinction is often made (for example Bennett 2003) between the 'classical' evaluation tradition deriving from the work of Tyler (1942) with its emphasis on specification and measurement of outputs from later approaches which present much wider perspectives such as Stufflebeam's (1983) CIPP (context-input-process-product) and Cronbach's (1982) utos (units of focus, treatments, observations/outcomes, settings) frameworks. In terms of 'how', discussion traditionally draws on wider discussions of methodology to contrast quantitative approaches, particularly experimental and quasi-experimental designs (Campbell 1975, Cook, et al, 2010) with approaches that seek to explore the subject of the evaluation using more qualitative methods such as thick description and case study (Parlett and Hamilton 1976, Stake 1986) or approaches that draw on the traditions of connoisseurship and criticism (Eisner 1985). Finally, in terms of 'who' should participate in evaluation and determine its outcomes, the history of evaluation exhibits a wide range of perspectives from those which give the key role to the evaluators themselves (Scriven 1976), through those who focus on the importance of commissioners and managers (Stufflebeam 1983) to those who seek to engage a wider range of stakeholders (Patton 1997; Guba and Lincoln 1989), including some who place a particular emphasis on the participative processes (Cousins and Earl 1995, Torres and Preskill 2001) or on the engagement of the disempowered (House 1991, Fetterman 1996). We will return to the 'how' and 'who' questions later. However, the primary focus of this paper is a particular approach to the 'what' question: that of what we call 'level models'.

'Level' models for evaluating CPD

We have used the term 'level models' to describe a family of evaluation approaches that share the characteristic of tracing the effects of training and development

interventions through a series of 'levels' each of which more closely approaches the 'ultimate' intentions or outcomes of the intervention. Although rarely made explicit, these models draw on an evaluation tradition which posits that programme design and implementation involve a series of inter-related components and the role of evaluation is to assess one or more of these components and the inter-relationships between them. Such ideas are in embodied in Stake's (1967) antecedent-transaction-outcome approach and Stufflebeam's (1983) aforementioned CIPP, among others. In fact, most writers trace these models back to the influential work of Kirkpatrick (1998), which was originally conceived in a series of journal articles in 1959. This model identifies four levels of outcome for interventions: (i) participants' reactions, (ii) participants' learning, (iii) changes in participants' behaviour, and (iv) desired results. More recently, in relation to the more specific topic of teachers' professional development, Guskey (2000) has presented a similar model. In this model he replaces changes in participants' behaviour with 'the use of new knowledge and skills' and replaces organisational results with 'student outcomes'. He also adds an additional level – 'organisational support and change' - between levels (ii) and (iii). Models such as these have influenced much official advice on the evaluation of CPD. For example, in its advice to schools, England's Training and Development Agency for Schools (TDA)[1] suggests that 'impact evaluation should focus on what participants learn, how they use what they have learned and the effect on the learning of children and young people' (TDA 2007, p. 2).

Such models, while enormously influential, have not gone unchallenged. For example, Alliger and Janak (1994) suggest that Kirkpatrick's model is based on three assumptions that may not hold in practice: that each successive level is more informative to the evaluator than the previous one; that each level is caused by the

previous level; and that each succeeding level is correlated with the previous level. Similar criticisms might be made of Guskey's model, For example, he argues that 'each higher level builds on the one that comes before. In other words success at one level is necessary for success at the levels that follow' (Guskey 2000, p. 78).

However, the various factors that Guskey identifies under 'Level 3, Organizational Support and Change', are not a consequence of the previous stage as the other levels are, but a set of conditions for the previous stages to lead to the next ones. This point is effectively picked up by Holton (1996) in his critique of Kirkpatrick's approach. He argues, not only that the levels are not necessarily sequential (for example, positive reactions may not be a necessary pre-condition for effective learning), but also that the model is inadequate in a more general sense for explaining evaluation findings:

> For example, if only the four levels of outcome are measured and a weak correlation is measured between levels two and three, all we really know is that learning from training was not associated with behaviour change. In the absence of a fully specified model, we don't know if the correlation is weak because some aspect of the training effort was not effective or because the underlying evaluation model is not valid.' (Holton 1996, p. 6)

In other words, we don't know whether poor outcomes are the result of a poorly designed programme or of factors which lie outside the programme itself. Holton goes on to develop a more complex model that identifies influences beyond the intervention that are likely to determine, first, whether the intervention will result in learning, second, whether any learning will be transferred into improved participant performance, and third, whether such increased performance will influence organisational results. In doing so, he considers variables relating to the individual participant (e.g., motivation), the programme (e.g., whether it enables the individual to try out new ideas in practice) and the organisation (e.g., whether effective transfer is rewarded).

Using similar ideas, Leithwood and Levin (2005) explore a range of models for evaluating the impact of both leadership and leadership development that embody various combinations of variables. In particular, they distinguish between what they call 'mediating' and 'moderating' variables. Mediating factors are analogous to the intermediate levels described above, in that they lie on an assumed causative path from the 'independent variable' (for example, the leadership development intervention) to the 'dependent variable' (i.e. the final outcome). Moderating variables, in contrast, are described as 'features of the organizational or wider context of the leader's work that interact with the dependent or mediating variables…[and] potentially change the strength or nature of the relationships between them' (Leithwood and Levin 2005, p. 12). These authors give examples of variables relating to the characteristics of students, teachers, leaders and the organisation, making the important point that, depending on how the theory or framework is used to guide the study, the same variable might be defined as a moderator, a mediator or a dependent variable. Thus, for example, 'employee trust' might be a dependent variable (the purpose of training programme), a mediator (a step on the assumed causative path from leadership development to improved employee motivation or performance) or a moderator (a factor in the work context that influences whether employees respond positively to leadership development activities).

The various level models described above have been used by their authors in a variety of ways. Kirkpatrick's original model, for example, was developed (as was Guskey's modification of it) for the pragmatic purpose of enabling training evaluators to carry out their task more systematically. Indeed, in his rejoinder to Holton's critique, Kirkpatrick claims the widespread use of his approach (he doesn't use the term 'model') for that purpose as the main evidence for its validity (Kirkpatrick

1996). Alliger and Janak's (1994) and Holton's (1996) critiques of Kirkpatrick, and Leithwood and Levin's work, in contrast, are based on a more traditional research-oriented approach. They seek to model empirically the factors that influence training and development outcomes through identifying key variables, specifying the relationships between these, and measuring them. Such approaches lead to more complex models of relationships than either Kirkpatrick's or Guskey's and also to the likelihood that different patterns of variables may be identified in different situations.

Developing a new model

The discussion above suggests that level models raise two key questions. First, what causative relationships are assumed to hold between a training or development experience and various kinds of potential outcomes? Secondly, how does the experience interact with situational factors associated with individuals and organisational arrangements and how do these interactions affect outcomes? When, in our earlier studies, we used a modified Kirkpatrick/Guskey model, we found that it fell short in enabling us to deal with these key questions, as the critiques in the previous section suggested. This led us to look to build on these authors' - and Leithwood and Levin's - work to develop a model of the effects of CPD programmes using a broader set of types of variables. Our model emerged through a series of multiple method studies undertaken mainly for what was then called the National College for School Leadership (NCSL), in England[2], including:

- Evaluations of individual leadership development programmes, both those designed to develop individual leaders, such as Leading from the Middle (LftM) and Leadership Pathways (Simkins 2009) and those designed to develop teams such as the Multi-Agency Team Development (MATD) Programme.
- Comparisons between programmes, such as our examination of the impact of in-school elements of three leadership development programmes, Leading from the Middle (LftM), the National Professional Qualification for Headship (NPQH) and the Leadership Programme for Serving Heads (LPSH) (Simkins, Coldwell and Close 2009).

- Studies of complex programmes with multiple aims, recipients and forms, such as the 14-19 Leadership and Management Development programme to support implementation of the new Diplomas (Coldwell and Maxwell 2008, Maxwell, Simkins and Coldwell 2009).

The frame for the model - shown in outline form in Figure 1 - is constructed around the following sets of key variables, and their interactions:

- Interventions: The CPD activities themselves.
- Antecedents: Those factors associated with individual participants that affect their ability to benefit from the opportunities offered to them.
- Moderating factors: Variables in the school and wider environment that influence whether, and how, the interventions lead, via the achievement of intermediate outcomes to produce final outcomes. These factors help to explain why apparently similar activities have different consequences for different individuals, teams and schools.
- Intermediate outcomes: Those outcomes of the CPD activities that are conceived to be pre-conditions for the achievement of the final outcomes, particularly learning, changes in participant behaviour and engagement in particular tasks or activities.
- Final outcomes: The intended effects of the CPD activities, primarily relating to effects on organisations, teachers and students.

These variables interact in often complex ways which are sensitive to the details of design and implementation of particular CPD activities.

Insert Fig 1 about here

As we noted above, our early studies of school leadership programmes were strongly influenced by the Kirkpatrick model. In responding to evaluation briefs, we focused primarily on various outcome levels, from participant reactions to impact on pupil learning (Intermediate Outcomes 1-3 and Final Outcomes 1-2 in Figure 1), although we recognised the difficulties in applying such models in practice. These include the complexity of outcomes (both intended and unintended) and the time taken for final outcomes to be achieved (especially at the level of student learning). However, as we carried out more studies we identified a number of themes that led us to extend our model in a number of ways and to identify limitations to the situations where such approaches can be applied.

First, we came to appreciate the importance of participants' motivations in influencing how they approached programmes and the impact of this on their programme experience. For example, in relation to Leadership Pathways (a programme for middle and senior leaders), some participants saw the programme as a step on the road to promotion while others wanted to use it to take stock and decide whether, for example, they eventually wanted to become a head. This affected whether they treated the programme quite instrumentally – as a necessary entry hurdle to mandatory preparation for headship – or as an opportunity for personal learning and growth. These factors, in turn, affected the extent and quality of participant engagement with various aspects of the programme. We found similar motivational differences in other programmes: for example between headteachers more recently in post who wanted to use LPSH to improve their performance and some very experienced heads who wanted to use it to take personal stock at an advanced stage in their career (Simkins, Coldwell and Close 2009).

Secondly, where the school was a partner in programme delivery, the ways in which the school engaged was critical for programme success. This was most obviously exemplified by the different ways in which programme coaching roles were interpreted and the effectiveness with which they were carried out, differences that typically reflected deeper issues around school priorities and school culture (Simkins et al 2006, Simkins 2009).

These two factors – participant motivations (and the factors that influence these) and organisational context – broadly correspond with Leithwood and Levin's 'moderating factors', but we found it useful to distinguish between them. Thus we used the term, 'Antecedents' for factors associated with participants' engagement with the programme, and 'Moderating Factors' for those associated with the

organisational and wider context in which the programme operates. Each of these can help explain why outcomes may differ for different participants and in different contexts. Furthermore, the consideration of Moderating Factors led us to recognise the importance of feedback loops through the role of leadership development programmes in developing aspects of individual, group and organisational capacity. For example, the experience of being coached led programme participants to develop skills that they could use with others. Consequently, we added Final Outcome 3 to our model.

Finally, new challenges emerged when we moved from the evaluation of leadership development programmes targeted at individuals to other kinds of programmes. Two programmes in particular illustrate this theme. Our evaluation of the Multi Agency Team Development Programme (MATD) not only required us to import into the model outcome variables relating to team learning and team effectiveness, it also led us to revisit antecedents and moderating factors from a team perspective. For example, one key issue was whether groups of participants recruited for the programme actually were teams, which raised further questions concerning the necessary characteristics of a 'team'. Another concerned how the ways in which a group was located within organisational structures helped or hindered the achievement of both learning and effectiveness in both the short and long runs.

Another case was the evaluation of the LSIS/NCSL 14-19 Leadership and Management Development Programme. This programme enabled participating organisations to access a range of development interventions for groups and individuals including national open seminars, bespoke workshops in consortia of schools and colleges, group and individual coaching between and within organisations, action learning sets and organisational development activities. Here, as we gathered data on each element of the programme, we realised it was not possible

to create a model for the programme as a whole. The programme's 'menu driven' nature meant that different choices were made by individual and organisational participants, so that different kinds of outcomes might be located in different organisations each with their own moderating and/or mediating factors. It proved impossible to encompass this complexity in a single level model of the type we had used previously. So, for example, it was possible and useful to use the model for examining individual coaching interventions, but not for exploring impacts of combinations of coaching and other interventions.

Drawing our learning together, a number of issues emerge clearly. First, we used the level model essentially heuristically. The model evolved - and increased in complexity - in response to both the differing designs of particular leadership development programmes and our emergent findings. Secondly, while we found our key categories of variables – antecedents, interventions, intermediate and final outcomes and moderating factors – quite robust, we had to recast these in relation to differences in detail between the various programmes we evaluated. Where programmes comprised different kinds of interventions as sub-components these needed to be modelled separately; and where the programme was overly complex in terms of the relationship between interventions and participants it had to be abandoned.

There was a third issue, however. When we used our model to gather data from participants and other stakeholders, we became increasingly aware of the ways in which participants and other actors constructed their own mental models of what these programmes were about, the outcomes that they were pursuing and the ways in which aspects of programme delivery were expected to influence these. Sometimes these differed from our own construction of the programme designers' intentions.

Thus, as we have seen, participants might use the programmes to take personal stock in relation to where they wanted their future careers to go, or see them as a set of hurdles necessary for promotion which needed to be jumped as economically as possible. Such motivations could lead to engagement in ways that were inconsistent with the programme's presumed primary objective of enhancing leadership competence. Similarly, schools could frame the in-school tasks which many programmes involved as traditional 'management projects' with which they were familiar from other contexts, emphasising 'getting things done' rather than seeing them as vehicles whose potential for learning needed to be carefully thought through and nurtured. Consequently, we were increasingly faced with the need to make a distinction between the 'design model' and the 'model in reality' as it was perceived by key actors. Whereas we attempted to formulate the former from programme documentation and evaluation briefs, it became clear that participants, their schools and other key actors often constructed their own versions of desired programme processes and outcomes which were not necessarily consistent with 'official' expectations.

These inconsistencies or contradictions were often important findings from our studies, which led us to look afresh at models such as ours not simply from a retrospective practical perspective but going back to their underpinning principles. It is to these ontological and associated epistemological issues that we now turn our attention.

Evaluation ontologies and level models

We have already noted both that the purposes of the authors of earlier models varied, from the intention to provide practical help for those engaged in evaluation to the more research-focused and empiricist, while our own approach developed

heuristically within the essentially pragmatic context of commissioned programme evaluations. In this section we consider more carefully the theoretical underpinnings of these approaches and their location within the broader literature on evaluation and social research. By so doing, we hope to provide a more secure theoretical basis for understanding CPD evaluation, thereby elucidating both the limitations of level models, and the possibilities for developing such models that better reflect the complexity of the social world. We present a threefold categorisation of approaches to evaluation based on different underpinning ontological positions familiar from social theory: the positivist or naïve realist position, the realist position and the constructivist position. These represent a modification of the distinction between quantitative and qualitative methods described earlier. As we will go on to outline, the second and third of these can be seen as critiques of the first.

The first category of evaluation approaches takes a broadly positivist view of the nature of social reality, drawing on a tradition dating back via Durkheim to Compte. These approaches assume there is a close relationship between the observable, which is captured via careful data gathering, and the objective reality of the social world. Such approaches often utilise experimental or quasi-experimental evaluation designs which attempt to measure impacts by controlling for factors that might confound such impacts. Typically, these types of studies can tell us something about effects of CPD in very limited but highly valid ways. A useful example here is Wayne et al's (2008) discussion of professional development impacts on pupil outcomes in the US. These authors discuss Carpenter et al's (1989) study which

> randomly assigned 40 first-grade teachers to two groups. One group received a brief 4-hour PD [Professional Development] programme. The other received an extensive 80-hour program known as cognitively guided instruction (CGI)… The students of the teachers who received CGI outperformed the [others] on three of the six student achievement measures. (Wayne et al 2008, p. 469)

Such findings provide some evidence of the effects of CPD in specific areas of pupil performance but the more general learning is less clear. Blamey and Mackenzie (2007, pp. 440-441) argue that such approaches flatten out 'variations in context' by treating interventions as 'unified entities through which recipients are processed and where contextual factors are conceptualised as confounding variables' rather than essential ingredients in understanding causal processes at work. In this case, we know that the intervention worked in some ways to improve pupil learning, but, as Wayne et al [ibid] note, such studies 'have not yet provided the kind of guidance needed to steer investments in PD'. Whilst evaluation designs of this kind are rare in the UK CPD evaluation literature, their underlying ontology and successionist view of causation (x causes y because, having attempted to rule out confounding factors, x is associated with and is temporally prior to y) is consistent with level models as used in the UK and elsewhere. Our own model, and others we discuss above such as Leithwood and Levin's, draw on this tradition in that they tend to use models highly reliant on, and derived from and modified by, empirical data. Just as social research in this tradition has been critiqued by more recent philosophical traditions, the next two positions discussed below can be seen, therefore, as different types of responses to this first position.

The second set of approaches sets out to be explicitly driven by theory rather than data, and includes the group of post-positivist approaches ‘realist(ic) evaluation’ (Pawson and Tilley 1997), ‘theory of change’ (Connell and Kubisch 1998) and ‘programme theory’ (Rogers et al 2000) approaches. These evaluation approaches draw on what is now usually called the "critical realist" social theory of Roy Bhaskar (1998), developed by others, notably - particularly in relation to the education field - Margaret Archer (1995). They share the ontological position that there are real,

underlying causal mechanisms that produce regularities observable in the social world. The level model tradition can be seen to fit in to this group, since the application of level models to programme and other evaluations can be thought of as using a theory-based approach. However, as we will go on to argue, level models including ours tend to underplay the complexity of the social world discussed by the theorists working within the critical realist paradigm in social science and evaluation research.

For realist evaluators and social researchers, the mechanisms that produce regularities are derived through what can be thought of as 'middle-range' theories: those 'that lie between the minor but necessary working hypotheses…. and the all-inclusive systematic efforts to develop a unified theory' (Merton 1968 p. 39). This is the sense in which such approaches are described as theory-based. These mechanisms operate in specific contexts to produce particular sets of outcomes. Hence these approaches have a generative view of causation, in contrast with the data-driven successionist view shared by positivist/naïve realist positions (Pawson and Tilley 1997). Viewed from this perspective, the role of the evaluator is to uncover such combinations of context, mechanisms and outcomes. These approaches have a strong focus on learning from evaluation about why and how programmes work, not just 'what works'. However, they can be criticised for failing to provide highly valid findings in the way that is claimed for experimental studies. From this perspective, the processes underlying the workings of CPD programmes are complex in a number of ways. In particular, they are embedded both within wider social structures and in specific contexts; they tend to lead, in context, to 'regularities' (in programmes, these are usually described as outcomes), they are unstable over time, and, since they underlie what is observable, observable data is necessarily incomplete.

Turning again to level models, two key issues emerge from this discussion. First, from these perspectives, level models tend not to provide enough detail of the theory or mechanisms underlying the levels of the model, and therefore are inadequate in explaining why particular outcomes occur in particular contexts. The processes indicated by the arrows that link the boxes in such models remain largely opaque. Secondly, for evaluators working with this post-positivist tradition, any single framework such as a level model cannot deal with all the possible combinations of context, mechanism and outcomes that may create change in a programme (Blamey and Mackenzie 2007). The discussion of our own approach in the previous section, indicating the difficulties we faced in dealing with programmes that support groups as well as individuals, or that comprise multiple interventions, illustrates this well. From a realist viewpoint, the evaluator should look at a number of possible mechanisms and compare their explanatory power in any given context in order to learn from them (Pawson and Tilley 1997). There is no inherent reason why level-type models cannot at least partly address this point, if they are underpinned by a theory-based understanding of the nature of learning and development, and are flexible and adaptable to the specifics of the programme or experience being examined. This is true of our model, which as we have shown is essentially a highly adaptable frame for constructing a variety of specific models to gather and interpret data. It is in fact a 'meta-model' to be redefined in each project. Nevertheless, one can still persuasively argue that any single model or even meta-model is inherently limited and limiting in its approach to understanding social processes and the complexity of the social world.

Finally, we need to consider a third category of ontological approaches to evaluation, which again can be seen as being in opposition to the first position above. This is based on an underlying ontological position that the social world is

constructed by the actors engaged within it. Associated with this is the epistemological position that knowledge of the social world can only be obtained through the perspectives of individuals and these perspectives may legitimately differ (Berger and Luckmann 1966, Denzin 2001). Evaluators from this tradition - which we label a constructivist position - concentrate on the perspectives and constructed meanings of programmes, their workings and outcomes from the viewpoints of all of those involved. Some of these positions - particularly Guba and Lincoln's 'fourth generation evaluation' (Guba and Lincoln 1989) - seem to us to be extreme, seeing no possibility in generating knowledge about a programme beyond that which is subjective, specific to particular instances and negotiated among a wide range of stakeholders. This underplays a more general constructivist position, namely that programme purposes may be contested, that individuals may experience interventions in different ways, and understanding these contestations and experiences may provide important information that can contribute to our understanding of how interventions work (Sullivan and Stewart 2006). This is the essence of the final point in the previous section about the ways in participants in the programmes that we have evaluated impute different personal and organisational purposes to programmes.

Level models can address this in part by treating their components as subject to interpretation rather than simply in terms of a priori specification and we have done this in many of our evaluations. Nevertheless, many theorists of professional development would be unhappy with this, tending to be deeply suspicious of any training and development model that they feel to be underpinned by reductionist ideas associated with performativity agendas (Fraser et al 2007), and level models are easily characterised in this way. The emphasis of such critics would be on the capacity of professional development to facilitate professional transformation and teacher

autonomy and agency (Kennedy 2005, Cochran-Smith and Lytle 1999). It could be argued that the enhancement of professional autonomy and the encouragement of genuine critique are just particular outcomes that can easily be incorporated into a level model. However, often implicit in these models are instrumentalist assumptions about the role of training and development programmes in promoting specific outcomes, which are typically pre-determined and measured in particular ways rather than emergent and constructed by the participants themselves. The models are concerned with promoting 'what works' rather than enabling practitioners to engage with 'what makes sense' (Simkins 2005). This leads to a deeper concern about the relationship between level models and the nature of professional learning itself. Webster-Wright argues, for example, that:

> Evaluative research often compares methods of delivery of PD [professional development] through evaluating learning outcomes, focusing on evaluating solutions to the problem of learning rather than questioning assumptions about learning… In addition, the majority of this research focuses on special factors affecting PD (program, learner or context) rather than studying the holistic, situated experience of learning.' (2009, p. 711).

She argues for a distinction to be made between professional development (PD) and professional learning (PL) and for studies to focus on the latter. This would involve an approach that 'views learner, context, and learning as inextricably inter-related, and investigates the experience of PL as constructed and embedded within authentic professional practice' (p. 713). This is a very different approach from that embodied in level models.

Conclusion

It was proposed at the beginning of this paper that evaluators need to address three key questions: what should be the focus of evaluation; how should these aspects be investigated and whose views should count in the evaluation. It was further suggested that level models focus on the first of these questions – the 'what'. However,

consideration both of our experience of using level models and of the theoretical perspectives discussed above makes it clear that things are not so simple.

Firstly, the analysis in this paper suggests that, while level models can be used in the positivist tradition to structure evaluations of well defined development programmes with clearly identifiable target groups and intended outcomes, perhaps more significant is their potential for exploring heuristically the workings of such programmes through identifying key variables, the possible relations between them and the ways in which these variables and relationships can be constructed: an 'inquiry' rather and 'audit review' approach to evaluation (Edlenbos and van Buuren 2005). However, the models also have limitations. From a realist perspective they do not typically give enough attention to the real mechanisms through which outcomes are achieved, either in their specificity or complexity; and from some constructivist perspectives they are based on reductionist instrumental assumptions that pervert the complex reality of genuine professional learning.

Secondly, level models need to be implemented and, in doing this, evaluators make choices about the kinds of data to gather, who to collect it from and what weight to give to it. Alkin and Ellett (1985) suggest three dimensions against which models or theories of evaluation should be judged: their methodological approach (from quantitative to qualitative), the manner in which the data are to be judged or valued (from unitary – by the commissioner or the evaluator - to plural), and the user focus of the evaluation effort (from instrumental to enlightenment). Alkin (2004) uses these broad dimensions to develop an ‘evaluation theory tree’, attempting to place each key writer on evaluation into one of these areas based on a judgement about their primary concern while recognising that this inevitably over-simplifies many writers’ views.

Level models are not easily placed on any of these dimensions. For those evaluators in the first of the traditions we identified above, the aim may be to specify intended outcomes, measure these and determine whether or not they have been achieved: a typically quantitative, unitary and instrumental approach. For others who reject such a position, such models may nevertheless be of value. For realists they provide one starting point for seeking to understand the complex reality of professional development and the mechanisms through which learning and other outcomes occur in a variety of contexts. And for some constructivists, the idea of multiple models which reflect the differing perspectives of various stakeholders may be of value. In each of these cases evaluations are likely to draw on more qualitative, plural and/or enlightenment-oriented approaches than positivist approaches do.

These complications emphasise the need to consider always, when and how level models are used. In making these decisions attention needs to be given to the purposes of evaluation and to the nature of the programme, activity or process being evaluated. In their comparison of two 'theory-driven' approaches to evaluation, Blamey and Mackenzie (2007) argue that 'theory of change' approaches are most apt for complex, large-scale programme evaluations and examining links between their different strands, whereas 'realist evaluation' approaches suit examinations of learning from particular aspects of programmes or from less complex programmes. From our experience of attempting to apply level models to a range of programme evaluations, it appears that the strengths of level models are similar to those of 'realist evaluation' models in that they can be particularly useful in uncovering the workings of well defined development programmes with clearly identifiable participant groups. Nevertheless, the emphasis on learning programmes is significant here: continuing professional development is, or should, comprise much more than programmes. Two

final consequences arise from this. First, there will be many areas of CPD activity for which level models are inappropriate and other evaluation approaches must be sought. These might include approaches such as biographical studies or rich case studies, which seek to see professional learning as an emergent personal and social process rather than one simply embodied in inputs and outputs. They might also include approaches that engage the learners much more explicitly as partners in the evaluation process than many commissioned evaluations typically do. Second, the necessary incompleteness of any one model (including level models as a family) requires us to aim explicitly to develop our theoretical understanding of the social world and in this way to 'make evaluations cumulate' (Pawson and Tilley 1997).

This leads to a final point. There is an added complexity for evaluators, such as ourselves, working in the arena of publicly funded evaluation research. On the one hand, as evaluators commissioned to evaluate government programmes we normally work under the expectation that we will generate results that are essentially instrumental: in Easterby-Smith's (1994) terms, results that 'prove' (or not) programme outcomes and perhaps also contribute to 'improving' programme design. However, the ways in which evaluation purposes are constructed raise important ethical issues (Elliott and Kushner 2007), and beyond this as academics our stance has a strong enlightenment focus, with a major concern for 'learning' about the programmes we study, placing them in context and, insofar as this is possible, generating understanding that can be extended beyond the case at hand (Torres and Preskill 2001; Coote et al, 2004). The analysis in this paper, by exploring the ways in which level models have been used to evaluate CPD programmes while explicitly linking them to underlying ontological positions, helps to explore this tension. It is all

too easy - and sometimes unavoidable - to succumb to the desire of contractors, whether explicit or not, to take an essentially positivist stance to evaluation. However, by doing so the real potential for learning may not be fully capitalised upon. In most of the work referred to here, we have been able to avoid this temptation, but the relationship between 'ownership', methodology and integrity is one that requires constant attention.

Footnote:

1. The TDA (Training and Development Agency for Schools) is an agency of the UK government, responsible for the training and development of the school workforce in England, administering funding, developing policy and monitoring initial teacher education and continuing professional development of teachers and other school staff.
2. England's National College for School Leadership, now renamed the National College for Leadership of Schools and Children's Services is one of the largest national leadership development enterprises in the world. Largely funded by government and with a total budget about £121 million in 2008/09, it runs or commissions a very wide range of leadership development programmes targeted at leaders at all career stages and now covering all children's services, not just schools. The titles of the programmes referred to in the text are largely self-explanatory, except for Leadership Pathways which is programme targeted at middle and senior leaders not yet eligible for the National Professional Qualification for Headship. For further details see www.nationalcollege.org.uk.

Acknowledgements: Thanks to John Coldron, Bronwen Maxwell and anonymous reviewers for comments on earlier drafts.

Mike Coldwell is Director of the Centre of Education and Inclusion Research (CEIR), Sheffield Hallam University.
Tim Simkins is Professor of Educational Leadership and Management in CEIR.

References

Alkin, M., 2004. Evaluation Roots: testing theorists' roots and influences. London: Sage.

Alkin, M. and Ellett, F., 1985. Evaluation models and their development. In: T. Husen and N. Postlethwaite, eds. International Encyclopaedia of Education: research and studies. Oxford: Pergamon, 1760-1766

Alliger, G. and Janak, E., 1994. Kirkpatrick's levels of training criteria: thirty years later. In: C. Scheier, C. Russell, R. Beatty and C. Baird, eds. The Training and Development Sourcebook. Amherst, MA: HRD Press**,** 229-228.

Archer, M., 1995. Realist Social Theory: The Morphogenetic Approach. Cambridge: Cambridge University Press.

Bennett, J., 2003. Evaluation Methods in Research. London: Continuum.

Berger, P. L. and Luckmann, T., 1966. The Social Construction of Reality: A Treatise in the Sociology of Knowledge, Garden City, NY: Anchor Books
Bhaskar, R.A., 1998. The Possibility of Naturalism (3rd edition). London: Routledge.
Blamey, A. and Mackenzie, M., 2007. Theories of change and realistic evaluation: peas in a pod or apples and oranges'. Evaluation, 13 (4), 439-455.
Blamey, A. and Mackenzie, M., 2007. Theories of change and realistic evaluation: peas in a pod or apples and oranges'. Evaluation, 13 (4), 439-455.
Campbell, 1975. Assessing the impact of planned social change. In G.M. Lyons (Ed.) Social Research and Public Policies. Hanover, NH: Dartmouth College Public Affairs Centre.
Cochran-Smith, M. and Lytle, S., 1999. Relationships of knowledge and practice: teacher learning in communities. Review of Research in Education, 24 (1), 249-306.
Coldwell, M. and Maxwell, B., 2008. Evaluation of the CEL/NCSL 14-19 Leadership and Management Development Programme Final Report. Internal report.
Connell, J. and Kubisch, A., 1998. Applying a theory of change approach to the evaluation of comprehensive community initiatives: progress, prospects and problems. In: K. Fulbright-Anderson, A. Kubisch, and J. Connell, eds. New Approaches to Evaluating Community Initiatives: Vol 2 Theory, measurement and analysis. Washington DC: Aspen Institute, 15-44
Cook, T.D., Scriven, M., Coryn, C.L.S. and Evergreen, S.D.H., 2010. Contemporary thinking about causation in evaluation: a dialogue with Tom Cook and Michael Scriven. American Journal of Evaluation, 31 (1), 105-117.
Coote, A., Allen, J. and Woodhead, D. (2004). Finding out what works: Building knowledge about complex community-based initiatives. London: King's Fund.
Cousins. J. and Earl, L., (Eds.) 1995. Participatory evaluation in education: Studies in evaluation use and organizational learning. London: Falmer.
Cronbach, L., 1982. Designing evaluations of educational and social programs. San Francisco: Jossey-Bass
Denzin, N.K., 2001. Interpretive Interactionism (2nd edition). London: Sage Publications.
Easterby-Smith, M., 1994. Evaluation of Management Development, Education, and Training, Gower.
Edlenbos, J. and van Buuren, A., 2005. The learning evaluation: a theoretical and empirical exploration. Evaluation Review, 29 (6), 591-612.
Eisner, E., 1985. The art of educational evaluation: a personal view. London: The Falmer Press
Elliott, J. and Kushner, S., 2007. The need for a manifesto for educational programme evaluation. Cambridge Journal of Education, 37 (3), 321-36.
Fetterman, D., 1996. Empowerment evaluation: An introduction to theory and practice. In D.M. Fetterman, S.J. Kaftarian and A. Wandersman (Eds.) Empowerment evaluation: Knowledge and tools for self-assessment and accountability. Thousand Oaks, CA: Sage.
Fraser, C., Kennedy, A., Reid, L. and McKinney, S., 2007. Teachers' continuing professional development: contested concepts, understandings and models. Journal of In-Service Education, 33 (2), 153-169.
Guba, Y. and Lincoln, E., 1989. Fourth Generation Evaluation. London: Sage.
Guskey, T., 2000. Evaluating Professional Development. Thousand Oaks, CA: Corwin Press.

Hansen, H.F., 2005. Choosing evaluation models: a discussion on evaluation design. Evaluation, 11 (4), 447-462.
Holton, E., 1996. The flawed four-level evaluation model. Human Resource Development Quarterly, 7(1), 5-22.
House, E.R., 1978. Assumptions underlying evaluation models. Educational Researcher, 7 (3), 4-12.
House, E.R., 1991. Evaluation and social justice: Where are we? In M.W. McLaughlin and D.C. Phillips (Eds.) Evaluation and education: At quarter century. Chicago: University of Chicago Press, 233-47.
Kennedy, A., 2005. Models of continuing professional development: frameworks for analysis. Journal of In-Service Education, 31(2), 235-250.
Kirkpatrick, D., 1996. Invited reaction: Reaction to Holton article. Human Resource Development Quarterly, 7(1), 23-25.
Kirkpatrick, D., 1998. Evaluating Training Programmes: the four levels. 2nd ed. San Francisco: Berrett-Koehler.
Leithwood, K. and Levin, B., 2005. Assessing School Leadership and Leadership Programme Effects on Pupil Learning. Nottingham: Department for Education and Skills.
Maxwell, B., Coldwell, M. and Simkins, T., 2009. Possibilities of partnerships as sites for learning: leadership development in English 14-19 Diploma consortia. Paper presented at the annual conference of the British Educational Research Association, Manchester, September.
Merton, R., 1968. Social Theory and Social Structure. New York: Free Press.
Parlett, M., and Hamilton, D., 1976. Evaluation as illumination: a new approach to the study of innovative programmes. Occasional Paper No. 9. Edinburgh: Centre for Research in the Educational Sciences.
Patton, M.Q., 1997. Utilization-Focused Evaluation (2nd edition). Beverly Hills: Sage.
Pawson, R. and Tilley, N., 1997. Realistic Evaluation. London: Sage.
Rogers, P., Hacsi, T., Petrosino, A. and Huebner, T.A. Eds., 2000. Program Theory in Evaluation: challenges and opportunities. San Francisco: Jossey-Bass.
Scriven, M., 1976. Evaluation bias and its control. In G.V. Glass (Ed.), Evaluation Studies Review Annual, Vol 1. Beverly Hills: Sage.
Simkins, T., 2005. Leadership in education: 'What works' or 'what makes sense'? Educational Management Administration and Leadership, 33 (1), 9-26.
Simkins, T., 2009. Integrating work-based learning into large-scale national leadership development programmes in the UK. Educational Review, 61 (4), 391-405.
Simkins, T., Coldwell, M., Caillau, I., Finlayson, H. and Morgan, A., 2006. Coaching as an in-school leadership development strategy: experiences from Leading from the Middle. Journal of In-Service Education, 32 (3), 321-340.
Simkins, T., Coldwell, M. and Close, P., 2009. Outcomes of in-school leadership development work: a study of three NCSL programmes. Educational Management Administration and Leadership, 37(1), 29-50.
Stake, R., 1967. The countenance of educational evaluation. Teachers College Record, 68, 523-540.
Stake, R., 1986. Evaluating educational programmes. In D. Hopkins, (Ed.). Inservice Training and Educational Development, London: Croom Helm.
Stufflebeam, D.L., 1983. The CIPP model for programme evaluation. In: G. Madaus, M.S. Scriven and D.L. Stufflebeam. Eds. Evaluation Models: viewpoints on educational and human services evaluation. Boston: Kluwer-Nijhoff, 117-141.

Stufflebeam, D.L. and Shinkfield, A., 2007. Evaluation: theory, models and applications. San Francisco: Jossey-Bass.
Stufflebeam, D.L. and Webster, W.J., 1980. An analysis of alternative approaches to evaluation. Educational Evaluation and Policy Analysis, 2 (3), 5-19.
Sullivan, H. and Stewart, M., 2006 Who owns the theory of change? Evaluation, 12 (2), 179-199.
Torres, R.T. and Preskill, H., 2001. Evaluation and organizational learning: past, present and future. American Journal of Evaluation, 22 (3), 387-95
Training and Development Agency for Schools, 2007. Impact evaluation of CPD. London: TDA.
Tyler, R.W., 1942. General statement on evaluation. Journal of Educational Research, 35, 492-451
Wayne, A.J., Yoon, K.S., Zhu, P., Cronen, S. and Garet, M., 2008. Experimenting with teacher professional development. Educational Researcher, 37 (8), 469-479.
Webster-Wright, A., 2009. Reframing professional development through understanding authentic professional learning. Review of Educational Research, 79 (2), 702-739.