The evolution of evidence problems in Adult Literacy and Numeracy

The field of Adult Literacy and Numeracy (ALN) has an evidence problem. This paper analyses that problem through the lens of Berriet-Soliec *et al.*'s (2014) tripartite evidence typology, which suggests that there are three overarching types of evidence available to researchers and policymakers: Type 1 evidence, which provides descriptions of the scope of social problems; Type 2 evidence measuring the impacts of programmes addressing those problems; and Type 3 evidence of the mechanisms through which programme impacts are realised. Whereas ALN was once characterised by a lack of Type 1 evidence of ALN problems and Type 2 evaluative evidence of programme impacts, there is now an abundance of these evidence types. However, ALN finds itself in an 'evidence impasse': it repeatedly appears that programmes aimed at improving ALN skills do not do so. This impasse may be the product of evaluation failure, not programme failure. I argue that evaluations should place greater emphasis on the production and analysis of Type 3 evidence of the mechanisms through which adult basic skills may be improved over time. This paper outlines a strategy for doing so, with a particular emphasis on long-term impact of literacy and numeracy practices on basic skills.

Key words: literacy practices; PIAAC; literacy skills; basic skills; evaluation

Introduction and background

The field of Adult Literacy and Numeracy (ALN) has an evidence problem. On one hand, there is an abundance of quantitative evidence documenting the scope and scale of basic skills[1] difficulties in countries around the world. Currently, the Programme for the International Assessment of Adult Competencies' (PIAAC's) Survey of Adult Skills (SAS) (OECD, 2013a) allows for international comparisons of basic skills – a process originally begun by the International Adult Literacy Survey (IALS) (OECD, 1995, 1997, 2000) and continued by the Adult Literacy and Lifeskills Survey (ALL) (2005, 2011). The field of ALN now has more evidence than ever before of the scale of adult literacy and numeracy problems, and the negative outcomes associated with those problems. For example, PIAAC's SAS highlights strong cross-sectional correlations between poor basic skills and negative outcomes such as low wages, unemployment, poor health, and reduced social and political engagement (OECD, 2013). Longitudinal research in Britain (Bynner and Parsons, 2006; Parsons and Bynner, 2007) has found similarly strong relationships between low basic skills in adulthood and negative outcomes later in adulthood across a range of domains. Comparisons of British cohorts born in 1958 and 1970 has further suggested that the negative impacts of poor basic skills have grown over time – that is, the associations between poor basic skills and negative life outcomes are stronger for individuals born in 1970 than for those born in 1958 (Bynner, 2002).

Findings such as these have encouraged governments to make significant investments in ALN programmes. With this increased policy interest has come an increased focus on programme accountability: when large amounts of money and policy time are invested in

---

[1] In the current paper, the phrase 'Adult Literacy and Numeracy (ALN)' is used interchangeably with 'basic skills'.

ALN programmes, policy makers feel a heightened need to rigorously evaluate those programmes, in order to measure impact and justify expenditure (AUTHOR 1, publication details withheld for anonymity).

However, most of the evidence about the impacts of these programmes has not been positive. On average, ALN programmes do not seem to make a difference to participants' skills, at least when those skills are measured immediately after programme completion (Sheehan-Holt and Smith, 2000; Reder, 2016). Some studies (e.g. Brooks, 2007; Rhys Warner *et al.,* 2008) have found that programme participants improved their literacy and/or numeracy skills, when comparing results on pre- and post-tests. However, these studies did not involve comparison groups. Whereas one high-quality comparison group study has reported quantifiably measurable skills gains for participants (Gyarmati *et al.,* 2014), most comparison group studies have found little evidence of participants leaving ALN programmes with quantifiably better basic skills than when they entered (AUTHOR 1, publication details withheld for anonymity; Reder, 2017). The evidence on changes in national skills levels is also largely negative. In England, for example, a comparison of national basic skills surveys published in 2003 (DfES) and 2011 (BIS) shows no improvement in basic skills levels during that period, despite globally unprecedented ALN investment during this period. Comparison of results on IALS and PIAAC's SAS is less straightforward: due to implementation differences across surveys, perceived changes in results over time should be interpreted with caution with regard to literacy (Pacagnella, 2016). And due to different conceptualisations of numeracy in IALS and PIAAC, numeracy results on the two surveys are not directly comparable OECD (2013b). However, the literacy results indicate moderate, statistically significant gains in England between IALS and PIAAC (six points or 0.1 Standard Deviations) (BIS, 2013; Pacagnella, 2016).

Mixed findings such as these give rise to questions of whether adult literacy and numeracy policies represent worthwhile investments of taxpayer money (Sheehan-Holt and Smith, 2000). These findings also illustrate the 'evidence impasse' that ALN finds itself in: policy makers see strong evidence for the need to invest in the improvement of ALN skills, but do not see clear evidence that those investments are paying off. This impasse is not unique to a small number of countries, but is characteristic of the field of ALN as a whole.

In this paper, I analyse the factors contributing to this impasse and thus to policy failure in ALN. Drawing on theoretical literature in the field of Evaluation Science (Donaldson, 2012), I focus on tensions and inter-relations amongst three different types of evidence that inform ALN policy: 1) evidence of basic skills problems; 2) evidence of the impacts of ALN programmes on those problems; and 3) evidence of the mechanisms through which those impacts may be achieved. I use this tripartite typology of evidence (Berriet-Solliec *et al.,* 2014) to analyse the evolving history of evidence problems in ALN. I then describe a 'policy window' (Kingdon, 2010) through which the current impasse may be addressed, leading to a more efficient and productive approach to programme evaluation in the field. In particular, this approach would lead to the greater emphasis on the collection and analysis of evidence on literacy and numeracy practices, and the impact of those practices on basic skills gain.

Tensions underpinning ALN research and policy

ALN evidence is characterised by a number of historical tensions. Perhaps the key tension prior to the mid-1990s and the appearance of IALS was that between advocates' calls for more government interest in ALN on the one hand, and governments' complacency and uninterest on the other (Hamilton and Hillier, 2006; Bailey, 2006). Nowadays, a central tension is the question of what outcomes should be focused on by policy makers, programme

staff and programme evaluators.

International assessments such as PIAAC have focused primarily on measures of individuals' literacy and numeracy skills. In doing so, they have privileged and advanced a conceptualisation of literacy and numeracy as 'portable, decontextualised' (Reder and Davila, 2005, p. 172) cognitive and technical skills that individuals can apply relatively consistently across a range of contextual settings, including home and work (Street, 1984; Green and Howard, 2007). An implicit assumption of this 'autonomous' (Street, 1984) approach is that literacy and numeracy skills can and indeed should be measured via standardised assessments such as PIAAC.

Critics of this focus on autonomous skills (e.g. Barton and Hamilton, 1998, 2000) have argued that skills are neither so easily conceptualised nor measured, and that the policy focus on literacy and numeracy skills is overly reductive (Belzer and St. Clair, 2005; Hamilton *et al.,* 2015). Such critics have argued that literacy and numeracy should be understood not primarily as decontextualized, quantifiably measurable skills, i.e. what people are capable of doing, but in terms of what they actually do, i.e. the practices in which individuals engage in their daily lives. However, this practice-focused approach to literacy and numeracy is not without its own tensions. At its most straightforward level, a practice-focused approach to ALN focuses on the literacy and numeracy activities that individuals undertake – for example, reading books, writing text messages and calculating the cost of one's groceries. In this approach, practices are conceptualised as 'common or typical activities or tasks' that individuals engage in (OECD, 1995). This 'individual activities' approach to practices has been at the heart of a number of studies (e.g. Sheehan-Holt and Smith, 2000; Reder, 2009, 2017; Nienkemper and Grotlüschen, 2019). It is also at the heart of PIAAC's approach to participants' use of their literacy and numeracy skills – their practices.

An alternative conceptualisation of practices focuses on literacy and numeracy not primarily as individual activities but as *social practices* (Street, 1984; Barton and Hamilton, 1998). Whereas the social practices approach does take account of the literacy and numeracy activities that individuals engage in (e.g. reading books and sending text messages), it places greater emphasis on the contextual nature of those activities and the social roles and meanings ascribed to literacy and numeracy practices by the individuals participating in them. In this social practices framework, 'practices' is thus a more abstract concept – for example, literacy practices are not just 'what people do with literacy', nor are they 'observable units of behaviour' (Barton and Hamilton, 2000, p. 7-8); they are 'cultural ways of utilising literacy' which involve 'values, attitudes, feelings and social relationships'. The social practices approach thus goes beyond ALN practices as individual-level activities, instead emphasising literacy and numeracy as social acts shaped by and embedded in context and community.

The social practices approach typically takes a strong epistemological stance (Esposito, *et al.,* 2014), privileging qualitative approaches to the study of practices, with a particular emphasis on ethnographic studies focused on adults' own perspectives on literacy and numeracy, and the meanings that individuals and groups ascribe to such practices (e.g. Barton and Hamilton, 1998). This methodological stance contrasts with the quantitatively orientated, measurement-focused approach to literacy and numeracy as individual activities, e.g. in PIAAC. Researchers focused on skills gain have typically adopted a quantitative approach, whereas those focused on literacy and numeracy practices have typically been qualitative. These methodological and epistemological tensions are redolent of the 'paradigm wars' (Tashakkori and Teddlie, 1998) that have characterised much social science research in recent decades. Whereas some ALN scholars (e.g. Reder 2009; Esposito *et al.,* 2014) have sought to bridge these divides, policymakers typically place greater value on quantitative

evidence, and have argued that such evidence is required in order to justify investment (AUTHOR 1, publication details withheld for anonymity). A key reason that social practice theories have gained little policy traction in the field of adult basic skills is the difficulty inherent in operationalising the qualitatively oriented social practices approach at great enough scale to be relevant to many policymakers (Reder, 2017). These difficulties are inherent in tensions regarding the unit of analysis in ALN studies: whereas quantitative and qualitative methods may both focus on individuals as the unit of analysis, qualitative approaches are more likely to focus on groups, or to seek to understand individuals in context, e.g. engagement in practices as part of social interaction (see for example Duncan, 2015). Despite its theoretical importance, the social practices approach has remained on the margins of ALN policy.

My analysis in this paper focuses not on tensions between skills and practices, or on tensions between quantitative and qualitative approaches, nor even on tensions between policymakers and qualitative researchers, but on less well-noted tensions amongst the different types of evidence that shape ALN policy. Berriet-Solliec *et al.* (2014) provide a typology of three types[2] of evidence available to researchers and policymakers:

- Type 1: Evidence of *presence*, i.e. the presence of social, economic or educational problems

---

[2] In addition to describing different *types* of evidence, Berriet-Solliec and colleagues highlight the different *levels* of evidence that researchers may collect. There are a range of approaches to evidence levels in the research literature (see Cartwright and Hardie, 2012 for a summary), with most such approaches producing hierarchies running from high quality systematic reviews at the top to single case observations and/or expert opinions at the bottom. The focus of the current paper is not on levels of evidence in the field of ALN, but on types of evidence, the relationships between those types, and the importance of those relationships for research and policy.

- Type 2: Evidence of *impact*, e.g. the impacts of programmes or policies which seek to alleviate the problems illustrated by Type 1 evidence

- Type 3: Evidence of *mechanisms*, e.g. the causal processes through which a programme has positive impacts on a problem.

Type 1 evidence provides policymakers and researchers with evidence 'on the state of the world' (Berriet-Solliec *et al.*, 2014, p. 198), i.e. the presence of problems and/or the impacts of these problems. The key aims of Type 1 evidence are to learn about and describe the world. By enabling policymakers to identify and quantify the scope and impacts of social problems, Type 1 evidence may provide impetus for policies and programmes which seek to make a positive impact on the world.

Type 2 evidence focuses on policy or programme outcomes and impacts. Type 2 evidence is typically generated through programme evaluations, which seek to produce evidence of programmes' effectiveness at 'difference making' (Berriet-Solliec *et al.,* 2014, p. 198) as they address the problems identified by Type 1 evidence. In ALN, the primary outcome that policy makers have focused on is programme impact on basic skills. To justify this focus, policy makers have appealed to Type 1 evidence showing the strong correlations between basic skills problems and a range of negative outcomes in life.

Type 3 evidence focuses not on programme outcomes, but on the mechanisms through which those outcomes are achieved. The key aim of Type 3 evidence is to improve understanding of how interventions facilitate outcomes. Mechanisms mediate the relationship between programme resources and activities on the one hand and programme outcomes on the other (Pawson and Tilley, 1997, 2004; Pawson, 2013). Type 3 evidence of mechanisms thus provides information on the causal pathways through which programme impacts may be achieved. As such, Type 3 evidence helps policy and programme stakeholders to develop and improve programmes, and to refine the programme theories underlying policy interventions.

Programme theory is the process or processes through which programmes are presumed to produce outcomes (Donaldson, 2001): programme theory describes the direct and indirect causal pathways through which programmes are hypothesised to achieve their aims (Chen, 1990; Weiss, 1995). Programme theory focuses on mechanisms, by which we refer not to programme activities but to the changes within the participants that those activities facilitate. These changes, in turn, may lead to the desired outcomes.

Programme theory is at the heart of all interventions, whether stakeholders are aware of that theory or not. Interventions are by their nature theories made manifest, and all 'social programs are based on explicit or implicit theories about how and why the program will work' (Weiss, 1995, p. 66). Programmes are not simply assumed to create change by their very existence, they are instead grounded on theoretical assumptions about the processes through which outcomes will be achieved (Pawson and Tilley, 2004). In ALN, for example, one such assumption is what I refer to as the 'skills dose' hypothesis: the assumption that a 'dose' of literacy or numeracy instruction will lead to quantifiably measurable skills gains. This assumption is apparent in the typical pre/post-skills assessment evaluation design.

The evolution of evidence problems in ALN

The field of ALN has long suffered from an evidence problem, but that problem has evolved over time. Prior to the appearance of IALS in the mid-1990s, the primary evidence problem in the field was the lack of rigorous quantitative evidence about adult basic skills. This lack of Type 1 evidence allowed the governments of wealthy countries to underestimate and/or ignore the scope and scale of adult literacy and numeracy problems: ALN was viewed as a marginal field, one which did not merit significant policy interest or financial investment. Since the mid-1990s, however, IALS, then ALL and now PIAAC's SAS have provided a large and convincing body of Type 1 evidence showing a high prevalence of poor basic skills

in wealthy countries, and strong correlations between basic skills on the one hand and outcomes such as employment status, earnings, social and political engagement, health, trust and well-being on the other (OECD, 2013a; AUTHOR 1, publication details withheld for anonymity).

However, the cross-sectional nature of these surveys has meant that they have only been useful for establishing correlations. In this regard, the Type 1 evidence problem characterising ALN has not yet been fully addressed: whereas we have an abundance of rigorous international evidence of basic skills problems and their correlates, we suffer from an almost complete lack of long-term longitudinal evidence illustrating the life course impacts of those problems. One country where such evidence does exist is the UK, where the British birth cohort studies have produced longitudinal Type 1 evidence of the life course impacts of literacy and numeracy problems (Bynner, 2002; Bynner and Parsons, 2006; Parsons and Bynner, 2007). The longitudinal findings from these birth cohort studies about the negative impacts of poor basic skills have been consistent with the cross-sectional findings of international assessments, suggesting that correlations between low basic skills and negative life outcomes may indeed be causal.

The global proliferation of Type 1 evidence of basic skills problems and their correlates has had significant policy impact (AUTHOR 1 *et al.,* publication details withheld for anonymity). In addition to national governments' heightened focus on basic skills, trans-national organisations such as the European Commission (EC) and the Organisation for Economic Cooperation and Development (OECD) have emphasised the need to improve basic skills levels in wealthy countries (EU-HLGL, 2012; OECD, 2013a). This heightened emphasis has contributed to greater investment in ALN programmes. In England, for example, the government responded to IALS with unprecedentedly high investment in basic skills provision (NAO, 2008).

Even those who have been critical of the nature and focus of Type 1 international assessment evidence have recognised the positive impact that this evidence has played in moving adult basic skills from the margins to the mainstream (Hamilton and Hillier, 2006) of modern policy making, by encouraging government interest and investment in ALN (AUTHOR 1, publication details withheld for anonymity).

This development has transpired concurrently with a growth in programme accountability requirements: in the modern state, programmes and policies must justify their investment by providing evidence of their effectiveness and impact. This has led to a proliferation of Type 2 evidence assessing programmes' 'difference making' (Berriet-Solliec et al., 2014, p. 198) with regard to basic skills. Typically, these evaluations have focused on measurable, decontextualized skills gains, rather than other outcomes such as ALN practices and non-cognitive impacts such as improved self-confidence and attitudes to learning (Tett and Maclachlan, 2007). Evidence about skills gain is typically generated through pre/post-tests of literacy and numeracy skills, i.e. comparisons of programme participants' basic skills at the start of the programme with their skills at programme completion or within a relatively short period after programme completion, e.g. 6-12 months.

Policy makers have been repeatedly criticised for their demand for quantitative evidence of programme impacts on literacy and numeracy skills (Hamilton *et al.,* 2015). While expressing sympathy for these critiques, AUTHOR 1 (publication details withheld for anonymity) notes that ALN programmes exist in a heavily contested policy environment: the modern welfare state is characterised by competing claims for investment in social programmes: interventions in one policy area (e.g. education) must ultimately compete with those in other areas (e.g. health) for government funding. Within the field of education itself, some influential researchers argue that money invested in adult education would be better spent on children (Heckman, 2006). The European Commission (2006) has argued that adult

learning's relative lack of programme impact data (in comparison to other policy fields) hampers efforts at securing government funding for ALN. Policy fields – and programmes within them – that can show good return on investment, using quantitative evaluations of impact, have a comparative advantage.

As noted above, however, the Type 2 evidence that has been generated has been largely negative: few ALN programmes appear to have generated the impacts that policy funders seek. While there is more Type 1 evidence than ever of the need to improve basic skills, the Type 2 evidence that it is possible to do so through classroom or workplace programmes is at best mixed, and at worst negative: policy-makers have extensive Type 1 data on proficiency, but these data do not point the way to improving proficiency. This is the evidence impasse.

Faced with this impasse, one might argue that the improvement of adult basic skills is a 'wicked' policy problem (Rittel and Webber, 1973) that may simply be too difficult to solve (or, more accurately, lessen). Wicked policy problems have a number of characteristics (Briggs, 2007; Alford and Head, 2017) that make it difficult to develop successful interventions, and difficult to develop appropriate evaluation designs for assessing success. A wicked policy problem is likely to have multiple, overlapping causes or antecedents, and multiple, overlapping consequences. There is social complexity at the user level: 'individual' problems are influenced by that individual's family, community, and other social networks. Perhaps most importantly from an intervention and evaluation standpoint, wicked problems cannot be solved by generic principles or linear heuristics (Blackman *et al.,* 2006). The mechanisms of causal change to address wicked problems may be complex and/or difficult to identify, and are likely to require long-term behaviour change.

Wicked problems can be contrasted with 'tame problems', the solutions to which are easier to identify and implement (*ibid.*). Tame problems are complicated but solvable: they

may be highly complex and require high-level skills and cross-organisational coordination to address (think of successfully launching a rocket), but the path to success can be mapped out and followed, given sufficient skills and resources. A key problem in ALN (and other policy fields) is that wicked problems involving education and skills are erroneously believed by policy makers to be tame. As a result, the programme theories underpinning interventions are mis-specified: they do not work because they are based on an incorrect theory of how change may be achievedUnsurprisingly, wicked policy problems are likely to be associated with a history of chronic policy failure, with efforts to address such problems having faltered repeatedly and across a range of contexts. This has certainly been the case in ALN: programmes have repeatedly failed to improve skills, and national skills levels have remained stagnant. However, I argue that the key failure is not a failure of programmes, but a failure of evaluation focus and design. In particular, the failure lies in the repeated focus on collecting Type 2 evidence on basic skills gain, before collecting sufficient evidence on the mechanisms through which basic skills gains might be achieved. This failure is the result of the flawed assumption that ALN programmes will work the way policymakers want or expect them to – that is, the assumption that the prevailing 'skills dose' hypothesis is correct. The evidence is consistent and compelling that it is not.

Berriet-Solliec *et al.* (2014, p. 199) argue that in any field in which causal pathways are not clearly understood, i.e. any field in which programmes keep failing to deliver the expected or hoped for results, evaluations science should refocus its attention: it should complement its focus on the collection of Type 2 evidence of programme impacts with an equal or even greater level of attention to collecting Type 3 evidence of programme mechanisms. A focus on Type 3 evidence is particularly important if 'the causal structure is [more] complex' than expected. Other evaluation scientists (e.g. Chen, 1990; Weiss, 1995; White, 2009; Pawson, 2013) agree with the need to develop a richer understanding of the

mechanisms of change, rather than repeatedly seeking Type 2 evidence of impact in the absence of Type 3 evidence of how impact may be achieved. Evaluations in ALN need to delve deeper into programmes' 'black boxes' (Stame, 2004) – i.e. they need to focus more on how and why skills gain happens, not just ask if it does. Focusing too much on Type 2 evidence of impact has negative impacts on the quality of evaluations and the design of future programmes, and leads to a glut of evaluations that *appear* relevant to policy makers, in terms of assessing the short or medium term programme impacts, but a paucity of evaluations that truly are relevant, through adding to programme theory in this challenging policy area.

Within ALN, a notable example of mechanism-focused theory is Practice Engagement Theory (PET) (Reder, 1994), which advances the hypothesis that increased ALN practices are the key mechanism leading to skills gain. This hypothesis is supported by findings from the Longitudinal Study of Adult Learning (LSAL) (Reder, 2009). Using longitudinally repeated measures of literacy and numeracy skills and practices over a seven-year period (Strawn *et al.,* 2007), LSAL was able to test PET's hypothesis that positive changes in practices would lead to improved skills in the long term. Whereas this study found no short-term impact of programmes on participants' skills, it did find short-term impacts on practices, and found that the presence of practice changes was correlated with long-term skills gains. That is, the causal pathway for skills gain was:

- Participation in an ALN programme did not generally lead to skills gains, as measured soon after programme completion.
- For many programme participants, however, participation did lead to measurable increases in literacy and numeracy practices.
- Over the course of several years, programme participants who experienced programme-related gains in literacy and numeracy practices went on to achieve measurable gains in literacy and numeracy skills.

14

In mapping this causal pathway, LSAL provides both Type 2 evidence of impact and Type 3 evidence of the mechanism through which that impact was achieved. LSAL did not evaluate a specific programme; rather it followed a cohort of American high school dropouts over time, and assessed the long-term impacts on these individuals of any ALN programme participation. As such, LSAL may offer greater insights into programme impacts in general, as opposed to the impacts of one programme in particular. LSAL also suggests that while conducting longitudinal research on marginalised groups offers a complex range of challenges, methodological aims can be met in the presence of sufficient resources and appropriate research designs (Strawn *et al.,* 2007). Turnbull (2002) has argued in favour of evaluation approaches that focus on programme mechanisms rather than specific programme models. LSAL effectively does this by focusing not on what one or more programmes do, but on a particular mechanism of change (practice gain) that is may be achieved across a range of programme models or types. It is the mechanism that matters, not the specific programme model.

Evidence of practices as a mechanism for basic skills gain in adulthood has also come from the 1970 British Cohort Study. Following a representative sample of more than 9,000 Britons since birth into their fifth decade, this study has found that individuals who read for pleasure frequently in adulthood experienced larger vocabulary gains between adolescence and midlife than those who did not read for pleasure (Sullivan and Brown, 2014). Whereas this finding is observational rather than the product of an experimental or quasi-experimental study, it is robust to rigorous controls for factors including socio-economic background and childhood and adolescent vocabulary test scores.

Implications: A policy window for practices?

Findings such as these suggest a need to rethink the programme theory underpinning

evaluations (Reder, 2012), moving away from the 'skills dose' hypothesis implicit in the typical pre/post-test evaluation design, and placing more emphasis on the measurement of practices. Building on this argument, Coben and Alkema (2017) explore a number of possible approaches for developing numeracy-specific practice measures, and argue that the historical tensions in ALN between practices and skills can at least in part be overcome by including measures of both in programme evaluation.

However, a greater evaluative focus on practices requires more than evaluator interest and desire. The 'hidden politics' of evaluation (Legorreta, 2015) play a central role in shaping evaluation design. Policymakers fund evaluations that fit their assumptions or hypotheses about the outcomes that matter, and the causal mechanisms leading to those outcomes. Policymakers thus need to be convinced of the importance of practices in the development of basic skills, to a greater degree than advocates of the social practices approach have yet achieved. The presence of Type 1 quantitative evidence about literacy and numeracy practices may play a positive role in this process, by helping to open a 'policy window' (Kingdon, 2010). The opening of such a window occurs when newly available data converges with policy trends to create new opportunities or impetuses for addressing social problems. The launch of England and Wales' Skills for Life adult basic skills strategy in 2001 provides an example (Fowler, 2005). In that instance, a national government fearful of diminished economic competitiveness in the global market responded IALS' Type 1 evidence of basic skills problems to embark on policy change. Type 1 evidence of low literacy and numeracy created a sense of crisis, while also illuminating possible gains if that crisis could successfully be responded to. A similar phenomenon was witnessed in Ireland, where funding for adult literacy programmes had been stagnant for years prior to IALS, but then increased 18-fold in the six years after the publication of that country's unexpectedly poor results (Bailey, 2006). In these and other countries, Type 1 evidence helped to move the issue of

adult basic skills from the margins to the mainstream (Hamilton and Hillier, 2006) in a way that qualitative evidence of the problem could not. It did so by providing policy makers with robust quantitative evidence describing the scope and scale of the issue and the social problems associated with it. Importantly, this Type 1 evidence converged with other policy factors to create a policy window; I do not mean to suggest in this paper that information is enough (on its own) to propel policy action.

In a smaller but still meaningful way, PIAAC data on practices could also help to open a policy window, with positive impacts on evaluation science and the ALN evidence impasse. IALS and ALLS produced some practice data, but PIAAC's data on literacy and numeracy practices is much richer, covering a range of literacy and numeracy activities at work and in everyday (non-work) life. (Reder *et al.* (2016) provide a valuable discussion of the scope and utility of PIAAC practice data.) PIAAC's relatively rich data on practices and their correlates represents an important step forward in the range and quality of Type 1 evidence of practices: there has been a significant expansion in the evidence 'on the state of the world' (Berriet-Solliec *et al.,* 2014, p. 198) that can be incorporated into ALN policymaking. Along with the production of this evidence, the OECD has argued that policymakers should take ALN practices seriously if they want to improve national skills levels: while the cross-sectional design of the SAS does not allow for robust causal claims, PIAAC does recommend greater engagement in literacy and numeracy practices as a mechanism for improved basic skills (OECD, 2013a).

I am not so naive as to believe that Type 1 evidence of literacy and numeracy practices will lead governments to devote more attention to practices than to skills. Governments worry about skills crises, not crises of practice. However, there is some indication of a positive chain of events that has led national and transnational policy actors (and evaluation funders) to place a greater focus on practices. LSAL's findings contributed to

a greater focus on practices by PIAAC (Reder *et al.,* 2016) and the European Commission (EU-HLG, 2012). PIAAC's interest in practices, in turn, has influenced evaluation approaches in England: a 2013 evaluation of national basic skills provision (Cook *et al.*) included quantitative measures of literacy and numeracy practice, as did a 2018 evaluation (Panayiotou, *et al.*). Nienkemper and Grotlüschen (2019) suggest that PIAAC practice data on the distribution of literacy practices within Germany can play a positive role in shaping basic skills education. Practices may, at least in a small way, be moving towards the mainstream of policy thinking.

While this slightly greater policy attention to practices is a positive development, it is only a stepping stone in the necessary evolution of ALN evidence, and may mask another key tension underlying the evidence impasse in ABE: that between a focus on practices as outcomes only or as outcomes and mechanisms of skills gain. Both the English evaluations cited above operationalised practices only as a programme outcome, rather than as a mechanism of long-term skills gain. This, I would argue, is precipitate. LSAL and Sullivan and Brown's cohort study analysis provide valuable evidence that increased practices lead, in the long term, to improved skills – but these are only two studies. The hypothesis requires more testing. This testing should be conducted in other national and local contexts and with different types of programme participants. Other hypotheses, e.g. suggestions that improved confidence and/or enhanced learner identity (Tett and MacClachan, 2007) contribute to long-term skills gains, should also be tested. The field of ALN needs much more Type 3 evidence of potential programme mechanisms: the field needs to go through a hypothesis-testing phase in order to break out of the current evidence impasse.

Knowledge cumulation about the mechanisms of change will be key to overcoming this impasse. Evaluations should focus not only on individual programmes, as the England evaluations did, but should also actively seek to add to theoretical knowledge in the field. A

collective commitment to knowledge cumulation (Pawson and Tilley, 2001) is essential for overcoming wicked policy problems: evaluations should have one eye on the programme in question, but the other eye on their contribution to programme theory. We need to understand the forest if we are to grow stronger trees. Most importantly, evaluators should develop and test hypotheses about mechanisms leading to skills gain. This, I would suggest, should be the key current aim for evaluation science in ALN.

If this aim is to be pursued, and the tension between practices and skills is to be overcome, another tension will need to be addressed. This is the tension between short-termism and the development of robust programme theory. Spurred by quantitative evidence on the negative impacts of poor skills, policymakers tend towards self-defeating short-termism (HLGL, 2012). Evaluation science in ALN and other fields all too frequently finds itself in a situation in which policymakers, rightly and urgently 'moved by the need to tackle serious social problems' highlighted by Type 1 evidence, expect these problems to be 'solved' in straightforward ways and in relatively short periods of time (Stame, 2004, p. 58). In such cases, evaluation funders may focus only on programme outcomes, and 'gloss over what is expected to happen [in the programme], the how and why' (*ibid.*) and when.

We need a different way. A longer-term longitudinal approach to evaluation is essential to developing programme theory in ABE: temporally mis-specified evaluations (i.e. evaluations that are too short to capture causal change) are by definition theoretically mis-specified, and theoretical mis-specification is at the heart of ABE's evidence impasse. Only by overcoming short-termism and focusing more on knowledge cumulation through long-term evaluations can we develop a genuine understanding of the impacts of ALN programmes, and the causal pathways leading to those impacts. Rather than repeated two-to-three-year evaluations focused on outcomes, ALN needs a broad set of evaluations focused on mechanisms and outcomes, and covering a broad range of national and participant

contexts. The goal would be knowledge cumulation, in the form of a robust body of Type 3 evidence on the mechanisms of change.

Once these mechanisms and their temporal requirements are more fully understood, ALN evaluations could then shift their focus back to short-term Type 2 evidence collection. However, they would make this switch from a position of knowledge about mechanisms and outcomes, rather than mere hope about programmes' impact on skills. If it were confirmed across a range of contexts and programme types that improvements in literacy and numeracy practices lead to long-term improvements in basic skills, we could then begin to evaluate literacy and numeracy programmes on their effectiveness at improving practices. We could do so with evidence-based confidence that these practice gains would in time lead to skills gain, as in LSAL. Evaluation design could then be based on robust, evidence-informed programme theory, rather than the hopeful but un-evidenced 'skills dose' hypothesis that currently predominates.

The likelihood of this happening is of course small: long-term evaluations are seen as too expensive for typical budgets and too long for normal policy cycles (AUTHOR 1, publication details withheld for anonymity). However, both problems can be addressed, given sufficient will. For example, long-term evaluation studies of 7-10 years could be funded by national research councils and supranational bodies such as the European Commission, with a focus on mechanisms of skills gain rather than on particular programmes *per se*. Such studies would allow researchers to generate and accumulate the Type 3 evidence of mechanisms that is needed to move out of the current evidence impasse.

Conclusion

This paper has charted the evolution of the evidence problem in ALN, and how this evolution has led to the current evidence impasse. This evolution has involved a number of historical

tensions in the field. The first such tension was that between advocates' calls for more government interest in ALN on the one hand, and governments' complacency and uninterest on the other. This tension was largely resolved through the proliferation of large-scale international assessment evidence in the form (initially) of IALS. In illustrating the scope and seriousness of the basic skills problem, IALS's Type 1 evidence moved adult literacy and numeracy from the margins to the mainstream of policy.

However, this forward step gave rise to another tension, that between practices and skills. Responding to IALS data, governments typically adapted a 'reductive' focus on skills (Belzer and St Clair, 2005), to the neglect of practices. At the same time, tensions over accountability grew. Across all policy fields (not just ALN), a heightened emphasis on programme accountability meant that programmes were increasingly required to justify their existence through quantitative evidence of impact/return on investment, while qualitative evidence of literacy and numeracy practices was largely ignored by policymakers.

These tensions within ALN between interest and uninterest (Hamilton and Hillier, 2006), practices and skills (Reder, 2009), accountability and autonomy (Pinsent-Johnson, 2015), and quantitative and qualitative evidence (Esposito *et al.,* 2014) have been well documented. In the current article I have focused on a more abstract but no less important tension influencing ALN evidence: the tension between evaluatory focuses on Type 2 evidence of programme outcomes and Type 3 evidence of programme mechanisms. In doing so, I have argued that ALN evaluations need to place greater emphasis on the production and analysis of robust quantitative evidence of mechanisms such as ALN practices that may lead to skills gain, as suggested by Practice Engagement Theory (Reder, 1994) and the Longitudinal Study of Adult Skills (Reder, 2009).

I have suggested that PIAAC's production of Type 1 evidence on practices may facilitate such a focus. This Type 1 evidence, I suggest, may support greater evaluative focus

on Type 3 evidence of mechanisms, which could in turn improve the relevance and utility of Type 2 evidence in the field. However, this will only happen if the field of ALN begins to take a longer-term approach to programme evaluation, so that researchers can develop a sufficiently nuanced understanding of how, and over what time period, mechanisms such as practice gains might lead to improvements in literacy and numeracy skills. Without significant strides forward in the long-term longitudinal collection and analysis of Type 3 evidence of the mechanisms of skills gain, ALN will remain in its current evidence impasse, perhaps for generations to come.

References

Alford, J., & Head, B. W. (2017). Wicked and less wicked problems: A typology and a contingency framework. Policy and Society, 36(3), 397–413.

Bailey, I. (2006). Overview of the adult literacy system in Ireland and current issues in its implementation. In J. Comings, B. Garner, & C. Smith (Eds.), Review of adult learning and literacy: Connecting research, policy and practice (Vol. 6, pp. 197–240). New Jersey: Lawrence Erlbaum Associates.

Barton, D., & Hamilton, M. (2000). Literacy practices. In D. Barton, M. Hamilton, & R. Ivanič (Eds.), Situated literacies: Reading and writing in context (pp. 7–15). London: Routledge.

Barton, D., & Hamilton, M. (1998). Local literacies: Reading and writing in one community. London: Routledge.

Belzer, S., & St. Clair, R. (2005). Back to the future: Implications of the neo-positivist research agenda for adult basic education. Teachers College Record, 107(6), 393–1411.

Berriet-Solliec, M., Labarthe, P., & Laurent, C. (2014). Goals of evaluation and types of evidence. Evaluation, 20(2), 195–213.

BIS (2013). The International Survey of Adult Skills 2012: Adult literacy, numeracy and problem solving skills in England. BIS research paper number 139. London: Department for Business, Innovation and Skills.

BIS (Department for Business, Innovation and Skills) (2012). 2011 skills for life survey: A survey of literacy, numeracy and ICT levels in England. BIS research paper number 81. London: Department for Business, Innovation and Skills.

Blackman, T., Elliott, E., Greene, A., Harrington, B., Hunter, D. J., Marks, L., . . . Williams, G. (2006). Performance assessment and wicked problems: The case of health inequalities. Public Policy and Administration, 21(2), 66–80.

Briggs. (2007). Tackling wicked problems: A public policy perspective. Canberra: Australian Public Service Commission.

Brooks, G. (2007). Effective teaching and learning: Reading. London: NRDC.

Bynner, J. (2002) Literacy, numeracy and employability. Nathan: Adult Literacy and Numeracy Australian Research Consortium. BIS Retrieved from http://eric.ed.gov/?id=ED473579

Bynner, J., & Parsons, S. (2006). New light on literacy and numeracy: Results of the literacy and numeracy assessment in the age-34 follow-up of the 1970 cohort study. London: National Research and Development Centre for Adult Literacy and Numeracy (NRDC.

OECD. (2013b). Relationship of the Survey of Adult Skills (PIAAC) to other international skills surveys. In The survey of adult skills: Reader's companion.Paris:Organisation for Economic Cooperation and Development. doi:10.1787/9789264204027-8-en.

Paccagnella, M. (2016). Literacy and Numeracy Proficiency in IALS, ALL and PIAAC. OECD Education Working Papers, No. 142. Paris: Organisation for Economic Cooperation and Development. doi:10.1787/5jlpq7qglx5g-en

Panayiotou, S., Hingley, S., & Boulden, K. (2018). Quantitative program of research for adult English and maths: Longitudinal survey of adult learners. Final research report. London: Department for Education.

Parsons, S., & Bynner, J. (2007). Illuminating disadvantage: Profiling the experiences of adults with entry level literacy or numeracy over the lifecourse. London: National Research and Development Centre for Adult Literacy and Numeracy (NRDC.

Pawson, R. (2013). The science of evaluation: A realist manifesto. London: Sage.

Pawson, R., & Tilley, N. (1997). Realistic evaluation. London: SAGE.

Pawson, R, & Tilley, N. (2001). Realistic evaluation bloodlines. The American Journal of Evaluation, 22(3), 317–324.

Pawson, R., & Tilley, N. (2004). Realist evaluation. London: Cabinet Office.

Pinsent-Johnson, C. (2015). From an International Adult Literacy Assessment to the Classroom: How test development methods are transposed into curriculum. In M. Hamilton, B. Maddox, & C. Addey (Eds.), Literacy as

numbers: Researching the politics and practices of international literacy assessment (pp. 187–206). Cambridge: Cambridge University Press.

Reder, S. (1994). Practice engagement theory: A sociocultural approach to literacy across languages and cultures. In B. Ferdman, R. M. Weber, & A. Ramirez (Eds.), Literacy across languages and cultures (pp. 33–74). Albany: State University of New York Press.

Reder, S. (2009). 'Scaling up and moving in: Connecting social practices views to policies and programs in adult education'. Literacy and Numeracy Studies, 16(2), Vol 17, no. 1, 35–50.

Reder, S. (2012). The longitudinal study of adult learning: Challenging assumptions. Montreal: Centre for Literacy.

Reder, S. (2016). Chapter 3. Skill use: Engagement in reading, writing and numeracy practices. In A. Grotlüschen, D. Mallows, S. Reder, & J. Sabatini (Eds.), Adults with low proficiency in literacy or numeracy OECD Education Working Papers, No. 131 Paris: OECD. doi:10.1787/5jm0v44bnmnx-en

Reder, S. (2017). Adults' engagement in reading, writing and numeracy practices. Applied Linguistics Faculty Publications and Presentations. 22. Retrieved from https://pdxscholar.library.pdx.edu/ling_fac/22

Reder, S., & Davila, E. (2005). Context and literacy practices. Annual Review of Applied Linguistics, 25, 170–187.

Rhys Warner, J., Vorhaus, J., Appleby, Y., Bathmaker, A., Brooks, G., Cole, P., . . . Pearce, L. (2008). The learner study: The impact of the skills for life strategy on adult literacy, language and numeracy learners. London: NRDC.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169.

Sheehan-Holt, J., & Smith, C. (2000). Does basic skills education affect adults' literacy proficiencies and reading practices? Reading Research Quarterlyl, 35(2), 226–243.

Stame, N. (2004). Theory-based evaluation and types of complexity. Evaluation, 10(1), 58–76.

Strawn, C., Lopez, C., & Setzler, K. (2007). It can be done: Sample retention methods used by the longitudinal study of adult learning. Portland, OR: Portland State University.

Street, B. V. (1984). Literacy in theory and practice (Vol. 9). Cambridge: Cambridge University Press.

Sullivan, A., & Brown, M. (2015). Reading for pleasure and progress in vocabulary and mathematics. British Educational Research Journal, 41(6), 971–991.

Tashakkori, A., & Teddlie, C. (1998). Mixed methodology: Combining qualitative and quantitative approaches. In Applied Social Research Methods (Vol. 46). Thousand Oaks: Sage.

Tett, L., & Maclachlan, K. (2007). Adult literacy and numeracy, social capital, learner identities and self-confidence. Studies in the Education of Adults, 39(2), 150–167.

Turnbull, B. (2002). Program theory building: A strategy for deriving cumulative evaluation knowledge. American Journal of Evaluation, 23(3), 275–290.

Weiss, C. (1995). Nothing as practical as good theory. In J. Connell, A. Kubisch, L. Schorr, & C. Weiss (Eds.), New approaches to evaluating community initiatives: Concepts, methods and contexts (pp. 65–92). New York, NY: The Aspen Institute.

White, H. (2009). Theory-based impact evaluation: Principles and practice. Journal of Development Effectiveness, 1(3), 271–284.