http://doi.org/10.35847/JCarpentieri.JAmorim.DMallows.PFreire.2.1.6

Research Article

Credibility, Relevance, and Policy Impact in the Evaluation of Adult Basic Skills Programs: The Case of the New Opportunities Initiative in Portugal

J.D. Carpentieri, University College London, Institute of Education

David Mallows, University College London, Institute of Education

José Pedro Amorim, University of Porto, Faculty of Psychology and Education Sciences, Centre for Research and Intervention in Education, and Paulo Freire Institute of Portugal

Abstract

Adult basic education (ABE) policies aim to help adults improve their literacy, numeracy and information and communications technology skills, as well as their qualifications, often in pursuit of economic gains such as better employment and earnings. The large-scale improvement of skills and qualifications has been referred to as a wicked policy problem, suggesting that it is extremely difficult and perhaps even impossible to achieve success in this policy domain. Evaluations have highlighted these challenges, with many programs showing little or no impact. Between 2006 and 2012, the Portuguese government ran a large-scale adult education program, the New Opportunities Initiative (NOI), which focused on the recognition and validation of adults' existing skills and the development of literacy and numeracy. The NOI was evaluated twice, in 2009 and in 2012. These two evaluations produced very different findings and outcomes: the first evaluation found the NOI to be a success, and led to continued investment, but the second evaluation reached more negative conclusions and was used as a rationale for de-funding the program. In this article we analyze these two sets of evaluations, investigating the reasons for their starkly different conclusions. We find that, while both evaluations had strengths, they also suffered from serious methodological and/or theoretical weaknesses. These weaknesses are *part of a broader pattern of evaluation errors* that characterize the field of ABE more generally and which make it more likely that ABE policies will continue to fail. Using the conflicting NOI evaluations as case studies, we offer potential solutions to ABE's evaluation problem, emphasizing the need to collect long-term longitudinal evidence on the causal mechanisms through which policy goals may be achieved.

Correspondence: J.D. Carpentieri, j.carpentieri@ucl.ac.uk; David Mallows, d.mallows@ucl.ac.uk; José Pedro Amorim, jpamorim@fpce.up.pt;

Author Note: The contribution of Dr Amorim to the development of the paper was supported by the Portuguese Foundation for Science and Technology (FCT) and by the European Social Fund – Human Capital Operational Programme (POCH) from Portugal 2020 Programme –, in the framework of the contract established under the transitional rule of Decree Law 57/2016, amended by Law 57/2017; and by the Portuguese Government, through the FCT, under the strategic funding awarded to CIIE – Centre for Research and Intervention in Education [grant no. UID/CED/00167/2013; UID/CED/00167/2019]. There are no relevant stipulations to this funding.

In modern economies, qualifications and skills are increasingly important. Studies such as the Programme for the International Assessment of Adult Competencies' (PIAAC) Survey of Adult Skills (OECD, 2013) highlight strong correlations between low qualification levels, low levels of literacy and numeracy, and negative outcomes such as low wages, unemployment, poor health, and reduced social and political engagement. Comparisons of British cohorts born in 1958 and 1970 indicate that the negative impacts of poor basic skills and low qualifications grow over time as economies evolve (Bynner, 2002), and have lifelong impacts (Parsons & Bynner, 2007). Such evidence has had an impact on policy, moving skills and qualifications from the margins to the mainstream of policy (Hamilton & Hillier, 2006), and encouraging governments to invest in adult basic skills, e.g., programs such as England's Skills for Life (Department for Innovation, Universities and Skills, 2007) as well as more general adult education interventions such as Sweden's Knowledge Lift (Albrecht, Van den Berg, & Vroman, 2005). However, with very limited exceptions (Gyarmati et al., 2014), evaluations of such interventions have shown little or no impact on participants' basic skills (Carpentieri, 2015; Reder, 2016), nor on their earnings or employment outcomes (Albrecht et al., 2005; Metcalf & Meadows, 2009). These null findings have proven problematic for advocates of such programs.

Schwandt (2009), a leading theoretician of evaluation science, emphasizes the need for evaluations to be credible and relevant, at both methodological and theoretical levels. Methodological credibility refers to the trustworthiness of the evidence used in the evaluation: can we believe the information presented to us? Methodological relevance focuses on whether that evidence is appropriate for addressing the evaluation's research questions. Methodological credibility and relevance play a central role in evaluation's legitimization function (Legorreta, 2015), through which governments demonstrate that: (a) they are acting on evidence and reason rather than instinct and ideology, and (b) their policies are effective and resources are being used wisely. This legitimization function is essential within the modern welfare state, which is characterized by a demanding public and competing claims for investment (Le Grand, 2003; Pierson, 2001).

In addition to generating methodologically credible and relevant evidence, evaluations need to be theoretically credible and relevant. Theoretical credibility refers not to the quality of an evaluation's evidence but to the appropriateness of its design (Schwandt, 2009). An evaluation may produce methodologically robust evidence, but be based on an inaccurate understanding or "program theory" (Chen, 1990; Pawson & Tilley, 2004; Weiss, 1995) of how change may be achieved, and thus provide an inaccurate assessment of an intervention's outcomes, impacts or value. Program theory describes the processes through which programs are presumed to

produce outcomes (Donaldson & Gooler, 2003); the direct and indirect causal pathways through which programs are hypothesized to achieve their aims (Chen, 1990; Weiss, 1995). Program theory focuses on mechanisms, by which we refer not to program activities but to the changes within the participants that those activities facilitate. These changes, in turn, may lead to the desired outcomes. Programs 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 & Tilley, 2004). For example, a program theory may be simple and linear, e.g., a program's "dose" of literacy instruction will directly increase adults' literacy skills, or more complex, e.g., a program will increase adults' literacy practices, and these increases in practices will in turn serve as mechanisms that contribute, over a sufficient amount of time, to improvements in literacy skills (Reder, 1994, 2009a, 2009b, 2012, 2014a). If a literacy program is implicitly or explicitly based on the more complex of these two theories but the evaluation of that program is based on the simpler theory, there will be a mismatch between program theory and evaluation design, thus weakening the evaluation's theoretical credibility.

Loss of credibility through theoretical misspecification occurs even if the evidence used by an evaluation is methodologically credible and relevant. For example, if an adult literacy program focuses primarily on improving participants' literacy practices (perhaps as a means towards long-term improvement of literacy skills), but an evaluation of that program focuses only on short-term impacts on literacy skills, the evaluation is not a credible assessment of the intervention's impacts, no matter how robust the evidence it has collected: the theory that the evaluation is testing is not the same as the program theory underpinning the intervention itself.

In addition to being theoretically credible, evaluations should be theoretically relevant. Theoretical relevance refers to the contribution of an evaluation to knowledge cumulation (Pawson & Tilley, 2004). Knowledge cumulation may refer to the assessment of an individual program via an evaluation, or an evaluation's broader contribution to program theory within the field, i.e., its contribution to increased understanding of the causal pathways through which programs may achieve their aims (Pawson, 2013). Table 1 provides a summary overview of methodological and theoretical credibility and relevance.

Wicked Policy Problems

The centrality of evaluation-based decision-making may present particular challenges when

Table 1: Methodological and theoretical credibility and relevance

	METHODOLOGICAL	THEORETICAL
Credibility	Trustworthiness or believability of the evidence	Appropriateness of the evaluation design for assessing intervention impact
Relevance	Appropriateness of the evidence for addressing the evaluation's research questions	Contribution of the evaluation to knowledge cumulation

governments seek to address so-called "wicked" policy problems such as adult skills and education (Payne, 2009). Wicked policy problems have a number of characteristics (Alford & Head, 2017; APSC, 2007; Rittel & Webber, 1973) that make it difficult to develop successful interventions, or 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 an individual's family, community, and other social networks. This social complexity is mirrored at the intervention level, with service provision likely to require the cooperation of multiple agencies across multiple government departments and/or policy domains. Perhaps most importantly from an evaluative standpoint, the mechanisms of causal change to address wicked problems may be complex or difficult to identify and are likely to require long-term behavior change. Unsurprisingly, wicked policy problems are likely to be associated with a history of chronic policy failure, with efforts to address such problems having failed repeatedly and across a range of contexts: while the policy problem may be clear, the "solution" is likely to be difficult to identify and operationalize. This has certainly been the case in adult skills (see e.g., Albrecht et al., 2005; Carpentieri, 2015; Metcalf & Meadows, 2009; Reder, 2016).

In this paper we will argue that, when evaluating interventions targeted at wicked policy problems such as adult skills, methodological credibility and relevance are necessary but insufficient evaluation conditions. Evaluations of adult skills programs have too frequently settled for methodological credibility and relevance while under-emphasizing the importance of theoretical credibility and relevance. As such, they have potentially reached inaccurate conclusions about program impact and have certainly made insufficient contributions to knowledge cumulation. Wicked policy problems demand that evaluations seek not just to evaluate individual initiatives but to move the field forward through cumulation of knowledge about how programs might work, why, for whom and in what contexts (Pawson & Tilley, 2004).

One of the most ambitious policies aimed at addressing the wicked problem of adult skills and qualifications was Portugal's New Opportunities Initiative (NOI), which ran from 2005 to 2013. NOI was a large-scale adult education and training program with a focus on the recognition and validation of adults' existing skills and the development of literacy and numeracy. The Portuguese adult population has one of the lowest levels of high school completion in Europe (Eurostat, 2019).[1] The NOI was an attempt to address this under-qualification (MTSS/ME, 2006) by providing routes through which adults could achieve school-level qualifications through adult education. As such, the NOI represented a "paradigm change in policy" (Carneiro, 2011, p. 29) that would systematically and sustainably address the chronic policy failure characterizing adult education and skills in Portugal.

The NOI was subject to two evaluations, in 2010 and in 2012. The first evaluation concluded that NOI was achieving its aims. The second drew the opposite conclusion and was used as justification for the cancellation of the policy. In this article we analyze these two sets of evaluations, investigating the reasons for and impacts of their different

1 In 2005, when NOI was launched, only 26% of the adult population had at least upper secondary, far from the 68% OECD and EU average (OECD, 2007). Nowadays, this figure has increased to 49% in Portugal and 78% in EU (Eurostat, 2019).

conclusions. In doing so, we draw comparisons between the NOI evaluations on one hand and evaluation approaches in adult basic skills on the other. The paper is structured as follows. After first describing the Portuguese policy context and the evaluation's goals, methods and findings, we then assess the credibility and relevance of the two sets of NOI evaluations, at both the methodological and theoretical levels. After discussing the policy uses of these evaluations, we conclude by providing recommendations for an evaluation strategy suitable to a broad range of wicked policy problems, including adult basic skills.

Telling the Story: The New Opportunities Initiative, the Political Context and the External Evaluations

The New Opportunities Initiative

The NOI was an unprecedented, large-scale national program of adult education that ran from December 2005 to March 2013. The NOI's main ambition was to "achieve mass schooling at the level of [upper] secondary" (MTSS/ME, 2006, p. 10). Within the initiative, secondary education was seen as "the minimum level" necessary for individuals to function in the modern "knowledge-based economy," and to be able to acquire and retain, throughout life, new skills (MTSS/ME, 2006, p. 3). The NOI set out to "accelerate the qualification levels of the Portuguese people" (MTSS/ME, 2006, p. 10) through processes of recognition, validation and certification of competences (RVCC) and participation in adult education and training (AET) courses. Both routes, RVCC and AET courses, gave participants the possibility of gaining certificates of equivalence at primary, lower, and upper secondary levels.

RVCC focused mainly on the collection of evidence of adults' lifewide and lifelong learning. That is, what they had learned throughout their lives, in formal, non-formal and informal contexts. However, not all knowledge was equally valued – the recognition and validation were limited to a set of competences defined by the frameworks for primary and secondary education. The AET courses, on the other hand, were designed mainly for the acquisition of new learning, although they did incorporate recognition of what participants already knew.

The First Evaluation, Coordinated by Roberto Carneiro

In 2007, the Ministers of Education and Labor invited Roberto Carneiro, ex-Minister of Education (1987-1991), to coordinate an external evaluation of the NOI. This started in April 2008 with a first set of evaluation results published in 2009 (Carneiro et al., 2009; Carneiro, Centro de Sondagens e Estudos de Opinião, Lopes, Cerol, & Magalhães, 2009a, 2009b; Carneiro, Liz, Machado, & Burnay, 2009; Carneiro, Mendonça, & Carneiro, 2009; Carneiro, Valente, Carvalho, & Carvalho, 2009) and a second set of results published the following year (Carneiro et al., 2010). The evaluation focused mainly on the perceptions of NOI of those involved as participants or professionals. Carneiro and colleagues took a primarily *emic* approach (Morris, Leung, Ames, & Lickel, 1999) to the collection of data, using focus groups, face to face and telephone interviews, case studies of NOI Centers, and an online survey to focus on stakeholder experiences of and perspectives on NOI. The evaluation engaged with

a broad range of stakeholders: adults enrolled in NOI,[2] adults who met conditions for access but did not apply, NOI professionals, employers, local opinion makers, civic associations, and academics.

One of the main foci of the evaluation was what Carneiro et al. (2010, p. 9) termed "the emergence of a brand." Policymakers were keen to understand stakeholder perceptions of the NOI as a public policy, and as a brand signaling a shift in attitudes to ABE. The evaluation also focused on the quality of service of the NOI Centers and stakeholders' satisfaction with this; the quality of the qualification processes and the assessment of key competences; and the impact of the initiative on participants.

The stated intention of the government in introducing NOI was to create massive brand awareness in order to affect a "paradigm change in policy" (Carneiro, 2011, p. 29), raising both awareness and credibility of adult education as a public good. Carneiro found that NOI was perceived, by target audiences and those who worked within the initiative, as a public (service) brand with clear values. It was seen as accessible, flexible and inclusive and as providing valorization of each individual and their life wide and lifelong experience of learning. However, the NOI "brand" was also perceived by stakeholders as being too closely linked to a specific political party and thus potentially time limited.[3]

NOI's professionals recognized (and celebrated) NOI's success indicators. However, the evaluation highlighted some indicators of inefficiency, such as adults remaining on waiting lists for long periods of time, as well as doubts about the comparability of the learning systems employed at the centers. Of equal concern was the certification of the learning processes, with questions about the validity, rigor, and comparability of the processes used. Some small business owners were concerned about increasing training costs without evidence of short-term impact on business results. Local opinion makers (e.g., academics, journalists, commentators) were the most critical of NOI. There were also doubts about the relative ease and the short duration of the learning processes, on the one hand, and the school-like nature of much of the provision, on the other.

Participants also reported strong reinforcement of self-esteem and an increase in motivation to continue learning, as well as a general improvement in soft skills such as self-management and initiative, adaptability, interaction, and communication. Parents said that they felt better able to support their children in school.

The 2011 Election Campaign: A Shift of Government and Policies

The NOI was a flagship policy of the XVII and XVIII Constitutional Governments. Following victory in the 2005 election, the Socialist Party had introduced policies of modernization with the stated aim of closing the educational gap between Portugal and its more developed neighbors in Europe, which was deemed to have a negative impact on the economy, social cohesion and personal development (MTSS/ME, 2006).

The NOI was an important topic in the 2011 election campaign. The opposition candidate Pedro Passos Coelho of the Social Democratics, the main center-right party in Portuguese politics,

2 The adults were at three different stages of the learning process: on a waiting list, in training (RVCC and AET courses), already certified.

3 NOI was a flagship policy of the Socialist Government (Carneiro, 2010) and had been the subject of heated cross-party debate. For example, during an election campaign, a representative of the Social Democratic party said that "the Engineer Sócrates [leader of the Socialist Party)] is convinced that he can exchange diplomas for votes" (RTP, 2011).

argued that NOI was a "scandal" (JN, 2011), an expensive "mega-production[4] giving credit and certifying ignorance." He promised "an external audit" and the end of the NOI (RTP, 2011).

In the aftermath of these statements, Joaquim Azevedo, who contributed to Carneiro's evaluation, said that a direct assessment of the quality of the training provided under NOI had not been carried out, as the evaluation focused on measuring the perceptions of those involved in the Initiative, and supporting the self-assessment of the New Opportunities Centers (Viana, 2011). Carneiro himself had noted that his evaluation had focused not on the quality and rigor of the certification process, but the perception of that quality and rigor among the people involved (Viana, 2011).

Shortly after the 2011 election, which was won by the Social Democratic party, the new Minister of Education and Science of the XIX Government, a coalition of the two right-wing parties in Portugal, criticized the NOI on the same grounds of inefficiency – NOI "ran poorly overall," he argued (Crato, 2011) – and for the lack of rigor and consistency in the certification process, suggesting that "handing out diplomas is not the solution." Following the election, a second evaluation of the NOI was commissioned by the new government.

The Second Evaluation, Coordinated by Lima

The second evaluation, coordinated by Francisco Lima, opted for an *etic* or outsider approach to program evaluation, explicitly taking a "diametrically opposed path to the previous evaluation" (Lima, Silva, & Fonseca, 2012b, p. 28). Rather than seeking to understand the perceived impacts of the NOI on stakeholders' lives, and the success or otherwise of the NOI in affecting a paradigm shift in popular understanding of adult education in Portugal, Lima et al. (2012a, 2012b) sought to measure participants' performance in the labor market in just two dimensions: earnings and employment status.

Lima et al. (2012a, 2012b) did not collect primary data. Instead, they drew on secondary analysis of two large data sets: an NOI database which recorded the learning outcomes of participants[5] and the national social security register of individuals' unemployment and other social benefits. These two datasets were linked on an individual level, allowing for quasi-experimental comparison of earnings and employment status among matched NOI participants and non-participants.

Lima et al. (2012b) found that participation in processes of RVCC did not increase the probability of transition into employment, nor did RVCC typically have an impact on earnings.[6] However, participation in AET courses was associated with a small but statistically significant increase in the probability of transition into employment, and there was also a positive relationship between AET course completion and an increase in earnings for participants who were already employed (Lima et al., 2012a).

Following the publication of the Lima evaluation the Social Democratic government moved to end the NOI. Silva et al. (2018) shows the magnitude of this de-investment. Between 2007 and 2011 the number of enrolments in the NOI ranged from 243,971 to 283,399. In 2012, enrolments decreased

4 It could also be conceptualized as mega-choreography, a stage production.

5 The System of Information and Management of the Educational and Training Provision (SIGO)

6 With the exception of participants with a higher level of education (secondary level) at the start of the process and in combination with modular training.

very significantly and, by 2013, had shrunk to just 28. NOI, which had been launched with the aim of affecting a "paradigm change" (Carneiro, 2011, p. 29) in adult education in Portugal, had effectively been closed down.

Carneiro's Methodological Weaknesses

Perhaps the most obvious difference between the two evaluations is their methodological approach. Whereas Carneiro's evaluation was primarily *emic*, i.e., focused on qualitative "insider stories" of stakeholders' experiences and perceptions of NOI, coupled with self-report quantitative data collected from stakeholders, Lima's evaluation was an *etic*, large-*N*, quantitative, quasi-experimental analysis of matched treatment and control groups. In discussing their methodology, Lima et al. (2012b) criticized Carneiro's methods, suggesting that Carneiro had the relationship between perceptions and impacts backwards: rather than basing assessment of program impacts on stakeholders' subjective perceptions (as Carneiro had done), Lima and colleagues argued that evaluations should be based on more objective measures of program impacts, and that these measures should then form the basis for the evaluator's perceptions about the program.

In advancing this opinion, Lima et al. (2012b) did not criticize the credibility of Carneiro's evidence (i.e., its believability or trustworthiness) but rather its methodological relevance. In Schwandt's (2009) framework, methodological relevance refers to the validity of the evidence, i.e., the appropriateness of the evidence for the evaluative claims made on its behalf. In drawing on qualitative self-report evidence to assess program outcomes such as gains in literacy and "learning to learn" skills (see e.g., Valente, Carvalho, & Carvalho, 2011), the Carneiro evaluation produced evidence that, while highly relevant for understanding learner experiences and perspectives, was markedly less relevant for measuring change over time due to program processes and activities. In doing so, the Carneiro evaluation opened itself to methodological criticisms of the sort advanced by Lima and colleagues.

Lima's Theoretical Weaknesses

The OECD (2002) defines evaluation as "the systematic and objective assessment of an ongoing or completed project, program, or policy," and suggests that evaluations "should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process" (p. 21).

Despite this characterization of evaluation as "objective assessment," a great deal of subjective decision-making goes into evaluation design. Political actors, whether funders or evaluators themselves, may exercise a high degree of discretion in establishing the criteria for program assessment, and this discretion can play a central role in determining evaluation results (Pollitt, 2013). In Portugal, we see evidence of this discretion in action, via a shift in how the key policy problem underlying NOI was characterized and evaluated.

NOI sought to address the wicked problem of adult skills and qualifications, a problem that had arisen at least in part through generations of underinvestment in Portuguese education. Wicked problems such as adult skills and qualifications compel governments to rethink traditional approaches. Accordingly, NOI was highly ambitious in scope: the policy sought to radically reshape Portugal's adult education system, and Portuguese adults' attitude to that system (Carneiro, 2011). Such an ambitious set of objectives creates opportunities for evaluators, but also challenges.

NOI's program theory was predicated on the notion that increasing both the supply of and demand for adult skills and qualifications would have positive impacts on attitudes to and uptake of adult learning opportunities, which would in turn have positive long-term impacts on employment and earnings, amongst other outcomes. Carneiro's evaluation therefore focused primarily on issues of supply and demand, assessing public acceptance of the NOI brand and stakeholder perceptions of program quality. Despite its methodological weaknesses, the Carneiro evaluation did achieve a high level of theoretical credibility, in that the evaluation design closely matched (and sought to assess the effectiveness of) the program theory underpinning NOI. In contrast, the Lima evaluation had a much narrower focus, measuring only short-term program impacts on earnings and employment.

The discretionary, subjective decisions of evaluators and/or their funders shape evaluation processes and results, making evaluations less objective than they might otherwise appear. However, appearances play a central role in the relationship between politics and evaluation. The conceptualization of evaluation as an objective, strictly rational and technical tool allows evaluations to be used as "mechanism[s] to disguise the politics involved" in decision-making (Legorreta, 2015, p. 62). Evaluations serve a legitimizing function, allowing governments to symbolically demonstrate that their actions are driven by evidence rather than ideology (Legorreta, 2015), even when this is not the case. Thus in addition to playing an instrumental role in policy-making by providing credible and relevant evidence of program effectiveness, evaluations may play a symbolic role, allowing policymakers to wave "the flag of evaluation to claim a rational basis for action (or inaction), or to justify pre-existing positions" (Henry & Mark, 2003, p. 264). Evaluations provide a "cloak [or mask] of rationality" that decision-makers can use to cover or disguise ideological decisions (Legorreta, 2015, p. 62).

We suggest that Lima's focus only on earnings and employment outcomes – as important as these outcomes are – is an example of this symbolic function of evaluation. By conducting a methodologically rigorous evaluation, Lima provided decision-makers with a seemingly objective assessment of NOI, and this assessment provided the Social Democratic government with a mask of rationality that was used to justify ending the NOI, which was so closely associated with the previous Socialist Party government. Lima et al.'s high degree of methodological credibility and relevance (particularly in comparison to Carneiro's lower methodological relevance) masked the subjective, discretionary decision-making underpinning their evaluation design. Despite appearing methodologically "objective", the Lima evaluation was theoretically mis-specified, in that it was based not on NOI's underpinning program theory but on a more reductive theory focused solely on short-term earning and employment outcomes. By focusing only on these outcomes Lima evaluated a complex, broad-ranging, long-term program using a somewhat simplistic, linear evaluation design.

Wicked Problems Require Knowledge Cumulation

Such theoretically mis-specified evaluations are unfortunately common in adult basic skills: the field is littered with methodologically credible and relevant evaluations that, because they were theoretically mis-specified, were likely to produce null findings. In England, for example, two successive evaluations of the national adult literacy and numeracy program (Cook, Morris, Cara,

Carpentieri, & Creese, 2013; Panayiotou, Hingley, & Boulden, 2018) were predicated on the notion that the program's dose of literacy instruction would directly increase adults' literacy skills, and that this increase would be sufficiently large and rapid to be measurable when comparing pre- and post-tests. In both evaluations, this proved untrue. In the United States, several randomized controlled trials (e.g., Miller, Esposito, & McCardle, 2011) have been predicated on the same dose-response design and have reached similarly negative conclusions. Through his Practice Engagement Theory, Reder (1994, 2009b) has provided a more realistic hypothesis, suggesting that whereas adult basic skills programs are unlikely to produce measurable short-term impacts on literacy and numeracy skills, they do lead to measurable increases in literacy and numeracy practices; these practice gains, in turn, serve as mechanisms that contribute, over a sufficient amount of time, to improvements in literacy and numeracy skills.

In focusing on the role of practices as a mechanism for skills gain, Reder implicitly addresses one of the key weaknesses of many program evaluations in wicked fields: their over-emphasis on a small range of politically high profile short-term outcomes, and their lack of attention to how, why, in what context, for whom, and over what time period those outcomes may be achieved and sustained (Pawson, 2013; Pawson & Tilley, 2004). Though they may be methodologically credible and relevant, such evaluations are theoretically limited because they do not delve into the program's "black box" – i.e., they do not provide sufficient evidence of the causal mechanisms through which programs achieve impact (Stame, 2004). Nor do they provide sufficient information for program designers seeking to improve the theories on which future programs can be based. Policymakers, rightly and urgently "moved by the need to tackle serious social problems" such as adult skills, focus only on program outcomes and impacts, and "gloss over what is expected to happen [in the program], the how and why" (Stame, 2004, p. 58). In such cases, evaluations lack theoretical relevance, i.e., they do not help us understand how desired outcomes are most likely to be achieved. This theoretical relevance is essential to policy development in wicked fields.

In Portugal, neither set of NOI evaluations generated sufficient evidence of how NOI might achieve its aims, through what mechanisms, in what contexts, and over what length of time. The Lima evaluation, for example, investigated economic and employment outcomes, but was much less interested in the mechanisms through which they might be achieved. This is in contrast to a quasi-experimental study of the economic impacts of England's Skills for Life Adult Literacy and Numeracy Strategy (Metcalf & Meadows, 2009) which, in addition to collecting evidence on employment and earning outcomes, collected evidence on mechanisms supporting employability such as self-esteem and motivation to participate in training and education. Metcalf and Meadows (2009) argued that these mechanisms may, over time, facilitate the economic outcomes of interest. Lima appears to have been un-interested in such processes.

This lack of contribution to broader program theory is in some ways more notable in the Carneiro evaluation – precisely because this was a more theoretically ambitious evaluation than Lima's. Carneiro considered a broad range of outcomes, including changes in literacy practices, but did not engage in sufficient consideration of how these outcomes may interact in causal chains over time to produce NOI's desired goals.

Even while seeking to evaluate a "paradigm shift in policy," Carneiro adopted a traditional evaluation approach focused on program outcomes and impacts, with insufficient attention to the conceptualization and operationalization of program mechanisms. This evaluation was meant to be developmental, not just summative – as such, it should have made meaningful contributions to program theory. It failed to do this, in large part because of a lack of focus on mechanisms. As with Lima's evaluation (2012a, 2012b), the black box of NOI was not opened and explored.

The relevance of the two evaluations thus goes only as far as the program (NOI) being assessed and does not extend to the field as a whole. Such an approach may be both efficient and sufficient in policy fields where program theory is well developed, i.e., areas in which stakeholders can turn to well-evidenced theories of how to achieve their policy aims. Adult skills are not such a field.

Conclusion

In this article, we have used the NOI evaluations as a case study of methodological and theoretical credibility and relevance in evaluations of interventions in wicked policy areas. Our analysis illustrates strengths and weaknesses in both sets of evaluations, both at the level of evidence use and evaluation design. With regard to the credibility and relevance of the evidence used in the two sets of evaluations, Carneiro's largely *emic* evidence was relevant for claims about stakeholder perceptions but was insufficient for assessment of program impacts on earnings and employment. In these areas, Lima's evidence was more relevant. However, with regard to the theoretical credibility of the two sets of evaluations, we suggest that Lima's methodological rigor masks a reductive, theoretically mis-specified evaluation approach which was inappropriate to NOI's program theory. This aspect of our analysis highlights the central role that the "hidden politics" of evaluation design may play in shaping evaluation design (Legoretta, 2015).

In this analysis, we have highlighted the parallels with evaluations of adult basic skills interventions. Lima's methodologically rigorous but theoretically mis-specified evaluation is reminiscent of a number of major adult literacy and numeracy evaluations, in terms of the evaluation design's misalignment with program theory. Analogous to the notion of the "mask of rationality" through which evaluations legitimize ideological decision-making, there is a "mask of credibility" through which evaluators and evaluation funders convince themselves that methodological credibility and relevance is sufficient. It is not. Methodological rigor is necessary but is not by itself sufficient as an evaluation design based on an unrealistic or unsupported program theory is an exercise in futility and does not contribute sufficiently to knowledge cumulation. As we have argued, evaluations in wicked policy fields need to go beyond merely assessing the intervention at hand; they need to actively contribute to program theory in the field as a whole (Pawson & Tilley, 2001). Collective commitment to knowledge cumulation is essential for overcoming wicked policy problems: intervention studies in wicked policy areas need to keep some focus on the forest, not just their individual tree.

In basic skills, one of the few studies to attempt to do this is the Longitudinal Study of Adult Learning (LSAL) (Reder, 2009a). Using longitudinally repeated measures of literacy and numeracy skills and practices over a seven-year period (Strawn, Lopez, & Setzler, 2007), LSAL was able to test and support Practice Engagement Theory's hypothesis that program-driven increases

in literacy and numeracy practices would lead, over time, to improved literacy and numeracy skills. One of the keys to LSAL's positive impacts is the long-term nature of the study: participants were tracked over seven years, allowing researchers time to focus on mechanisms, not just outcomes. Thus, LSAL was able to test and contribute to program theory in a way that neither NOI evaluation, nor evaluations such as those conducted by Cook et al. (2013) and Metcalf and Meadows (2009) did. Metcalf and Meadows (2009) have suggested that their own 3-year evaluation was unlikely to have covered a long enough period of time for employment and earnings effects to become evident. Notably, Reder (2014b) found that whereas adults with more than 100 hours of basic skills program participation did not show earnings gains (compared to non-participants) in the first 5 years of LSAL, after 9-10 years, participants showed large comparative gains.

> Pawson and Tilley (2001) have argued that evaluation is: cursed with short-termism. Programs are dispatched to meet pressing dilemmas, evaluations are let on a piecemeal basis, methods are chosen to pragmatic ends, and findings lean towards parochial concerns. Our hope, possibly against hope, is for a future evaluation culture that is more painstaking and for an evidence base that is more cumulative. (p. 322)

We share this hope and suggest that LSAL shows a possible way forward. To avoid repetitive and non-productive short-termism in adult skills evaluations, there is a need for long-term evaluations and a long-term approach to knowledge cumulation. Longer term longitudinal evaluations would give researchers an improved chance of developing a clearer understanding of the intermediary causal mechanisms that lead to policy relevant outcomes such as skills gains, better employment and increased earnings. Greater understanding of causal mechanisms (including the time required for such mechanisms to take effect) would allow for the development of more nuanced and robust program theories. This would in turn lay the groundwork for more sensible evaluation indicators and program targets. If improved adult skills are an investment worth making – and they certainly are – then so too is improved program evaluation. Without the latter, our progress towards the former will be far slower.

References

Albrecht, J., Van den Berg, G. J., & Vroman, S. (2005). *The knowledge lift: The Swedish adult education program that aimed to eliminate low worker skill levels. IZA Discussion Paper No. 1503*. Bonn: Institute for the Study of Labor (IZA).

Alford, J., & Head, B. W. (2017). Wicked and less wicked problems: A typology and a contingency framework. *Policy and Society, 36*(3), 397-413. doi:10.1080/14494035.2017.1361634

Amorim, J. P. (2016). *Literacy in Portugal: Country report. Adults*. Köln: ELINET – European Literacy Policy Network. Retrieved from http://www.eli-net.eu/fileadmin/ELINET/Redaktion/user_upload/Portugal_Adults_Report1.pdf.

APSC (2007). *Tackling wicked problems: A public policy perspective*. Canberra, Australia: Australian Public Service Commission.

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

Carneiro, R. (2011). 'New Opportunities' and new government: A paradigm change in policy. In R. Carneiro (Ed.), *Accreditation of prior learning as a lever for lifelong learning: Lessons learnt from the New Opportunities Initiative, Portugal* (pp. 29-79). UNESCO, MENON and CEPCEP.

Carneiro, R., Valente, A. C., Liz, C., Lopes, H., Cerol, J., Mendonça, M. A., & Melo, R. Q. (2010). *Iniciativa Novas Oportunidades: Resultados da avaliação externa (2009-2010)*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Mendonça, M. A., & Carneiro, M. A. (2009). *Análise da Iniciativa Novas Oportunidades como acção de política pública educativa. Caderno temático 1*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Liz, C., Machado, M. R., & Burnay, E. (2009). *Percepções sobre a Iniciativa Novas Oportunidades. Caderno temático 2*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Valente, A. C., Carvalho, L. X., & Carvalho, A. X. (2009). *Estudos de caso de Centros Novas Oportunidades. Caderno temático 3*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Centro de Sondagens e Estudos de Opinião, Lopes, H., Cerol, J., & Magalhães, P. (2009a). *Painel de avaliação de diferenciação entre inscritos e não inscritos na Iniciativa Novas Oportunidades. Caderno temático 4*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Centro de Sondagens e Estudos de Opinião, Lopes, H., Cerol, J., & Magalhães, P. (2009b). *Estudo de percepção da qualidade de serviço e de satisfação. Caderno temático 5*. Lisboa: Agência Nacional para a Qualificação.

Carneiro, R., Melo, R. Q., Jacinto, F., Caldeira, H., Salvado, I., Marmelo, M., Reis, S., . . . Rondão, C. (2009). *Auto-avaliação de Centros Novas Oportunidades: Adequação do SIGO às necessidades de avaliação. Caderno temático 6*. Lisboa: Agência Nacional para a Qualificação.

Carpentieri, J. (2013). Evidence, evaluation and the tyranny of effect size: A proposal for more accurately measuring programme impacts in adult and family literacy. *European Journal of Education, 48*(4), 543-556. doi:10.1111/ejed.12046.

Carpentieri, J. (2015). Adding new numbers to the policy narrative: Using PIAAC data to focus on literacy practices. In M. Hamilton, B. Maddox, & C. Addey (Eds.), *Literacy as numbers: Researching the politics and practices of international literacy assessment* (pp. 93-110). Cambridge: Cambridge University Press.

Chen, H. T. (1990). *Theory-driven evaluations*. London: Sage.

Cook, J., Morris, M., Cara, O., Carpentieri, J., & Creese, B. (2013). *Investigating the benefits of English and maths provision for adult learners. BIS research papers 129a & 129b*. London: Department for Business, Innovation and Skills.

Crato, N. (2011, September 13). *TVI news*. Retrieved from https://www.youtube.com/watch?v=BYXaK0cdI6I.

Department for Innovation, Universities & Skills (UK) (2007). *World-class skills: Implementing the Leitch review of skills in England*. Norwich: The Stationery Office.

Donaldson, S. I., & Gooler, L. E. (2003). Theory-driven evaluation in action: Lessons from a $20 million statewide work and health initiative. *Evaluation and Program Planning, 26*(4), 355-366. doi:10.1016/S0149-7189(03)00052-1.

Eurostat (2019). *At least upper secondary educational attainment, age group 25-64 by sex*. Retrieved from https://ec.europa.eu/eurostat/tgm/table.?tab=table&init=1&language=en&pcode=tps00065&plugin=1.

García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2017). *Quantifying the life-cycle benefits of a prototypical early childhood program (No. w23479)*. National Bureau of Economic Research.

Gyarmati, D., Leckie, N., Dowie, M., Palameta, B., Hui, T. S. W., Dunn, E., & Hébert, S. (2014). *UPSKILL: A credible test of workplace literacy and essential skills training*. Ottawa: Social Research and Demonstration Corporation.

Hamilton, M., & Hillier, Y. (2006). *Changing faces of adult literacy, language and numeracy: A critical history*. Stoke-on-Trent: Trentham.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314. doi:10.1177/109821400302400302

JN (2011, May 16). *Passos Coelho promete reformular 'escândalo' das Novas Oportunidades*. Retrieved from http://www.jn.pt/micro-sites/eleicoes-legislativas-2011/noticias/psd/interior/passos-coelho-promete-reformular-escandalo-das-novas-oportunidades-1853545.html.

Le Grand, J. (2003). *Motivation, agency, and public policy: Of knights and knaves, pawns and queens*. Oxford: Oxford University Press.

Legorreta, P. C. G. (2015). *The hidden politics of evaluation: Towards a smarter state?* (Doctoral dissertation). Retrieved from http://etheses.whiterose.ac.uk/11124/.

Lima, F., Silva, H., & Fonseca, T. (2012a). *Avaliação dos cursos de educação e formação de adultos e formações modulares certificadas: Empregabilidade e remunerações*. Lisboa: Instituto Superior Técnico, Universidade Técnica de Lisboa, and CEG-IST, Centro de Estudos de Gestão do IST.

Lima, F., Silva, H., & Fonseca, T. (2012b). *Os processos de reconhecimento, validação e certificação de competências e o desempenho no mercado de trabalho*. Lisboa: Instituto Superior Técnico, Universidade Técnica de Lisboa e CEG-IST, Centro de Estudos de Gestão do IST.

Metcalf, H., & Meadows, P. (2009). Outcomes for basic skills learners: A four-year-longitudinal study. In S. Reder & J. Bynner (Eds.), *Tracking adult literacy and numeracy skills: Findings from longitudinal research* (pp. 225-241). New York, NY: Routledge.

Miller, B., Esposito, L., & McCardle, P. (2011). A public health approach to improving the lives of adult learners: Introduction to the special issue on adult literacy interventions. *Journal of Research on Educational Effectiveness, 4*(2), 87-100.

Morris, M. W., Leung, K., Ames, D., & Lickel, B. (1999). Views from inside and outside: Integrating emic and etic insights about culture and justice judgment. *Academy of Management Review, 24*(4), 781-796.

MTSS/ME (2006). *Novas Oportunidades: Iniciativa no âmbito do Plano Nacional de Emprego e do Plano Tecnológico*. Lisboa: MTSS/ME.

OECD (2002). *Glossary of key terms in evaluation and results-based management*. Paris: OECD/DAC. Retrieved from https://www.oecd.org/dac/evaluation/2754804.pdf.

OECD (2007). *Education at a glance 2007*. Paris: OECD.

OECD (2013). *OECD skills outlook 2013: First results from the Survey of Adult Skills*. Paris: OECD. doi:10.1787/9789264204256-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: NRDC.

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

Pawson, R., & Tilley, N. (2001). Realistic evaluation bloodlines. *The American Journal of Evaluation, 22*(3), 317-324. doi:10.1016/S1098-2140(01)00141-2

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

Payne, J. (2009). Divergent skills policy trajectories in England and Scotland after Leitch. *Policy Studies, 30*(5), 473-494.

Perry, A., Amadeo, C., Fletcher, M., & Walker, E. (2010). *Instinct or reason: How education policy is made and how we might make it better*. Reading: CfBT Education Trust.

Pierson, P. (Ed.). (2001). *The new politics of the welfare state*. New York, NY: Oxford University Press.

Pollitt, C. (2013). The logics of performance management. *Evaluation, 19*(4), 346-63.

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, NY: State University of New York Press.

Reder, S. (2009a). The development of literacy and numeracy in adult life. In S. Reder & J. Bynner (Eds.), *Tracking adult literacy and numeracy skills: Findings from longitudinal research* (pp. 59-84). New York, NY: Routledge.

Reder, S. (2009b). Scaling up and moving in: Connecting social practices views to policies and programs in adult education. *Literacy and Numeracy Studies, 16*(2) & *17*(1), 35-50.

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

Reder, S. (2014a). *The impact of ABS program participation on long-term literacy growth*. Washington, D.C.: U.S. Department of Education.

Reder, S. (2014b). *The impact of ABS program participation on long-term economic outcomes*. Washington, D.C.: U.S. Department of Education.

Reder, S. (2016). 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* (pp. 37-59). Paris: OECD. doi: 10.1787/5jm0v44bnmnx-en

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

Roche, C., & Kelly, L. (2012). *The evaluation of politics and the politics of evaluation. Background paper 11*. Developmental Leadership Program.

RTP (2011, May 17). *Sócrates e Passos Coelho polemizam sobre 'Novas Oportunidades'*. Retrieved from http://www.rtp.pt/noticias/politica/socrates-e-passos-coelho-polemizam-sobre-novas-oportunidades_v442794.

Schwandt, T. A. (2009). Toward a practical theory of evidence for evaluation. In S. Donaldson, C. Christie & M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice* (pp. 197-212). Thousand Oaks, CA: Sage.

Silva, G. X., Valente, A. C., Simões, F., Santos, D., Freire, M., Alves, M. J., Lameira, S., & Duarte, T. (2018). *Sistema nacional de qualificações – 10 anos*. Lisboa: Agência Nacional para a Qualificação e o Ensino Profissional.

Stame, N. (2004). Theory-based evaluation and types of complexity. *Evaluation, 10*(1), 58-76. doi:10.1177/1356389004043135

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.

Valente, A. C., Carvalho, L. X., & Carvalho, A. X. (2011). Bringing lifelong learning to low-skilled adults: The New Opportunities Initiative. In R. Carneiro (Ed.), *Accreditation of prior learning as a lever for lifelong learning: lessons learnt from the New Opportunities Initiative, Portugal* (pp. 145-182). UNESCO, MENON and CEPCEP.

Viana, C. (2011, May 18). *Avaliação das Novas Oportunidades não incide sobre a qualidade da formação*. Retrieved from https://www.publico.pt/2011/05/18/portugal/noticia/avaliacao-das-novas-oportunidades-nao-incide-sobre-a-qualidade-da-formacao-1494711.

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