A Selection of the Best Papers from the 5th Biennial South African Monitoring and Evaluation Association Conference 2015

Abstract

This abstract describes that the Best Papers that were selected from the 5th Biennial South African Monitoring and Evaluation Association Conference 2015.

In the article "Evaluation opportunities and challenges: Perspective of an Emerging Evaluator by Paul Kariuki and Purshottama S. Reddy, they argue that despite the increased demand for monitoring and evaluation opportunities, emerging evaluators are challenged to access these opportunities due to lack of requisite experience , qualification and support. Based on the author's personal perspectives, this article calls for a change in mind-set on the part of all key stakeholders who should by create an enabling environment for emerging evaluators to develop their capacity to their full potential.

In the second article "Harmonising M&E Systems: Positioning the data atom at the centre" by Johan Olivier and Mr Georg Bernhoeft, argue that M&E systems that have traditionally been developed and designed from the reporting requirement perspective are doomed to fail. They have designed a technology solution to harmonise the M&E systems that consists of a number of inter-related work-package requirements where the "data" work-package is the point of departure and the "reporting" work-package is the end-product.
In the third article "Theory of Change Development for South African Education Programmes: a means to increase ownership and utilisation of evaluation findings" by Ms Eleanor Hazell, Ms Hazel Mugo, Ms Carmel Marock, reflects on TOC development for three evaluations of education programmes, to discuss the advantages and pitfalls of different methods, how toc development can stimulate dialogue and consensus building and may increase ownership and utilisation of evaluation findings.

In the fourth article "Professionalising Monitoring and Evaluation for Improved Performance and Integrity: Opportunities and Unintended Consequences "by Richard M Levin, argues that professionalisation of M&E practices will contribute to the establishment of a recognised "Evaluation Profession", through a professional body and affiliated professionals. The consequence of this will potentially be differentiation between those who hold qualifications and are 'legitimately' working in the field of evaluation and those who are 'unqualified 'but who have gained experience. It may also create hierarchical divisions within the occupational classes of evaluators and a new division between 'professionals' and 'amateurs'. To avoid pitfalls and unintended consequences, it is critical to look at how the profession or occupation will be specified and classified and the consequences this would have on existing practitioners and emerging communities of practice. Professionalisation must be a transformative project sensitive to the unintended consequences, which avoids elitism and exclusion.

The fifth article "The impact of a remedial reading programme on second language grade 4 students in KZN: Evidence from a randomised impact evaluation" by Stephen Taylor, Brahm Fleisch, Volker Schöer, the authors report on a randomised controlled trial of a remedial programme designed to boost the English reading and literacy skills of grade 4 students, for whom English is a First Additional Language. The paper describes some of the challenges involved in implementing a randomised controlled trial in the context of the South African school system. The paper also reflects on how this sort of impact evaluation presents a challenge to the con-

ventional research in education policy, but also creates valuable opportunities for inter-disciplinary collaboration to take knowledge further.

In the sixth article "Turning a Mess into a Message: Evaluating Complex Programmes Using the Keystone" by Gordon Freer and Sebastian Lemir, in developing an evaluation framework for a five year M4P (Making Markets Work for the Poor) DFID funded programme in Ghana, the authors developed the keystone node approach, that centres on the examination of keystone nodes within the theory of change (ToC). Like the keystone in a stone arch, if a theory of change keystone node fails, the theory of change crumbles at that point. As such, each node contains critical assumptions or steps of logic that are vital to the intervention theory progressing from one stage to the next. By focusing the evaluation on 'keystone nodes', evaluation efforts target the most learning-rich steps within the theory of change, allowing for a better understanding of complex, multi-year programmes.

This paper has demonstrated that improving the acquisition of reading in the early grades is central to the education quality challenge in developing countries, and especially in South In the final article "Using Impact evaluation for education policy innovations: the case of early grade literacy in South Africa"by Nompumelelo MohohlwaneStephen Taylor have described of two recent RCTs focusing on early grade reading in South Africa. The RCT of the Reading Catch-up Programme, implemented in Pinetown, Kwa-Zulu Natal has provided important lessons. The findings have highlighted the need for a valid counter-factual in measuring impact, which is a strength of the RCT methodology. The second RCT discussed, the Early Grade Reading Study (EGRS), which is being implemented in 230 schools in the North West province provides an exciting opportunity for further learning. It is anticipated that the findings will address some of the questions emerging from the RCUP study as well as provide substantive information on the binding constraints in the teaching of language in South African schools.

Contents

The impact of a remedial reading programme on second language grade 4 students in KZN: Evidence from a randomised impact evaluation.

Theory of Change Development for South African Education Programmes: A means to increase ownership and utilisation of evaluation findings.

Turning a Mess into a Message: Evaluating Complex Programmes Using the Keystone Node Approach

Using Impact evaluation for education policy innovations: the case of early grade literacy in South Africa

Evaluation Opportunities and Challenges: Perspectives of an Emerging Evaluator

First Author:
Paul Kariuki
Co-Authors: Purshottama S. Reddy

Abstract

The field of evaluation is growing rapidly on all fronts, fuelled mainly by rapid institutionalization of monitoring and evaluation across all levels of government. This growth has resulted in an increased demand for competent monitoring and evaluation practitioners. This trend is further propelled by the need to analyse processes and the results of development endeavours by means of evaluation in a bid to provide a nuanced perspective of the effectiveness of a development policy or programme. While there are thus ample opportunities for emerging evaluators to immerse themselves in the evaluation field, a lack of experience excludes many from active economically gainful evaluation opportunities. Most evaluation opportunities are extremely competitive and demand experience. This is exclusionary and challenging for an emerging evaluator who does not have the requisite experience to gain entry into the field; those who indeed gain entry cannot do so without external support. Furthermore, the majority of emerging evaluators are also challenged in terms of professional qualifications for a range of reasons such as a lack of financial resources to access professional education in monitoring and evaluation.

Based on the author's personal perspectives, this article calls for a change in mind-set on the part of all key stakeholders in the evaluation field. The evaluation community is tasked with the daunting responsibility of creating an environment that enables emerging evaluators to grow. In terms of capacity enhancement, organizations should provide emerging evaluators with professional development opportunities to enhance their knowledge, ideas and skills in order for them to be productive in executing their responsibilities and to offer high quality services.

The articles concludes by suggesting that, given the necessary support, emerging evaluators have the capacity to be innovative and offer fresh perspectives as well as new ideas that could improve an organization's work significantly. They thus deserve to be supported to develop their capacity to their full potential.

Keywords: Emerging evaluator, monitoring, evaluation, competencies, opportunities, evaluation capacity development

Introduction

There is growing pressure from donors locally and internationally on development organizations to demonstrate accountability and values for money through various evaluation activities (Lennie et al, 2015). As a result, many of these organizations have resulted to relying on professional evaluators externally contracted to conduct program implementation and impacts evaluations (Naccarella et al, 2007). However, with the changes being experienced by civil society organizations worldwide as a result of dwindling financial resources, contracting external evaluators is costly for them to manage such services. The better option is thus grooming staff internally who can undertake evaluation responsibilities and provide informed recommendations as to which interventions work, why they work as well as those that do not and why they fail. In this article, emerging evaluators are defined as these evaluation professionals that provide evaluation services within the organization whilst
developing their specialized monitoring and evaluation capacities under the guidance of seasoned evaluation practitioners.

It is appreciable that the sudden surge in demand for evaluation expertise has opened significant opportunities for emerging evaluators to grow their professional capacities. This phenomenon is largely supported by the increased need to generate reliable information within the prescribed standards so that decision makers can utilize it to improve interventions. It is noted that organizations that invest in systematic enquiry of their processes and interventions perform significantly better than average organizations (Segone et al, 2006:31-32). They are proactively responding to their contexts and have a better nuanced understanding of the role of monitoring and evaluation in their organizations. The staff team understand that monitoring and evaluation is a necessary practice not only for enhancing organizing learning but also for setting new priorities, strategies and reconsideration of existing standards (ibid). Emerging evaluators act as conduits for facilitating this shared understanding in an organization among colleagues, challenging assumptions so that all can think critically about their work and explore alternatives that have potential to shift their work at a high level of operation.

This article explores in detail the role of emerging evaluators in organizations. It begins with a discussion on their usefulness in the organizations in second section. It then follows in the third section with a definition of key terminology of key concepts followed by a discussion about the essential prerequisites of an emerging evaluator in the fourth section. In the fifth section, the paper discusses the barriers that emerging evaluators encounter in the course of their career journey. The sixth section discusses the role of the evaluation community in supporting them in their professional growth and career development. It concludes with a summation of key issues in the seventh section.

The Role and Usefulness of Emerging Evaluations in

Emerging evaluators have a significant role they play in an organization. These are described briefly in this section, namely:

Promotion of shared understanding about the organization and its interventions

Evaluation as a profession is growing rapidly globally, attracting scores of people interested in it across a range of disciplines (Donaldson and Christie, 2006:243). Carden (2007:7) claims that evaluation as a science is "alive and flourishing, constituting the most exciting career a young social scientist may embark upon". Given the increase in interest to enter the profession, there is significant increase of emerging evaluators in many organizations providing

technical services by offering assistance in examining programme effectiveness on beneficiaries, showing which programmes are working optimally and why among other aspects (Forss et al, 2006: 31). They enhance shared understanding, about results, increase participatory engagement among colleagues, thereby increasing chances of ownership as well as enhance programme and organizational development (Dabelstein, 2003:365). As such, emerging evaluators are a critical human resource as they provide evaluation results that enable enhanced programming by organizations in order that they can respond to beneficiaries needs appropriately.

Facilitate the development of useful professional networks and organizational alliances
Moreover, emerging evaluators facilitate the development of professional networks within and out of the organization as they engage with other professionals about their work (Forss et al, 2006: 34). They share their views and other knowledge products such as reports with other colleagues in a professional network. In doing so, they enhance their own understanding of their work, gain useful

insights and perspectives from colleagues from other organizations represented in an evaluation as well as other professional associations. This exposure broadens an emerging evaluator's professional worldviews, and ultimately benefits the organization in the end as the quality of work is enhanced overtime due to repeat exposure to professional networking opportunities.

Enhanced communication of organization's programmes to external audiences
Connected to the above two points, due to the opportunity to share programme results and participating in professional networking opportunities, emerging evaluators become effective communicators of their work over time. This aspect of the evaluation profession is essential because well-communicated evaluation results enhance not only a shared understanding about organization's work internally but also becomes a useful talking point for resource mobilization. Globally, donors are interested in knowing the impact of their investment on beneficiaries through the various programmes they support. To satisfy this quest, often donors require that the organizations they support financially provide evaluation reports regularly to "demonstrate the effectiveness" of their programmes (Rossi, Lipsey and Freeman, 2004: 250). An emerging evaluator has to develop this critical skill in order to bring real value in an organization in terms of attracting partnerships that enhance its work in various ways such as project funding, project collaborations with other organizations as well as formation of organizational networks and alliances that enhance the services offered to the society in general (Donaldson and Christie, 2006:249).

Building of strong inter-personal professional relationships in an organization
Relationships matter in evaluation, it is a mark of a professional evaluator (Wilcox and King, 2014:3). Emerging evaluators have a critical role to play in building strong inter-personal professional relationships among colleagues in an organization. As such, building relationships is vital for ensuring ownership of programmes and harnessing commitment as well as accountability from programmes team for effective implementation and follow-up activities to the end (Rossi, Lipsey and Freeman, 2004). In those instances where an emerging evaluator is weak in relationship building, there are difficulties experienced in terms of programme implementation, data collection flow, reporting as well as harnessing accountability in terms

of individual responsibility in case of non-performance. In the converse, where an emerging evaluator has excellent inter-personal relationship with colleagues, the difficulties mentioned above do not exist in most cases, as the team feels appreciated, understood and have a real stake in the overall effectiveness of the programme. They are also committed to support organization's efforts as a collective. This skill set is a must for an emerging evaluator and brings enormous value into an organization.

Strengthening of the organizational programmes

Emerging evaluators are at the coal face of every programme as they assist in conceptualization design and evaluation of the interventions in the organization. As such, by providing technical guidance and leadership in all these phases, they "strengthen the value of services" offered by an organization through its programmes (Forss et al, 2006: 35). In all these phases, programme goals, outputs and outcomes as well as individual roles and responsibilities are clarified and the team have a shared understanding of the intervention.

Promotes use of evaluation results in the organization

As primary producers of evidence in an organization through project evaluation activities, emerging evaluators encourage and promote the use of evaluation findings to improve organizational

programming. Segone et al (2006:35) argues that increased utilization of evaluation results increase not only the value of an organization's programming but also enhances mutual accountability among staff in an organization. Evaluation results are then used for decision making and learning. As such, emerging evaluators play a strategic role in an organization in that through their services, the entity is aligned to its vision. They identify what works, what doesn't and why, thereby ensuring that the organization provides services that are responsive to the needs of their beneficiaries.

Boosting of morale among colleagues in an organization

In most instances, emerging evaluators due to their work in an organization as monitoring and evaluation officers, they encourage colleagues to work together in implementing project activities, collection of routine data and final project evaluation. Forss et al (2006:37) argues that emerging evaluators understand that evaluation being an intellectual activity, can be daunting and therefore encouraging staff members to be involved in such an activity sends a message to the team that they are part of the organization and their input matters. This inclusion is known to inspire staff and encourages them to commit to the organization.

In summary, emerging evaluators are a vital part of an organization's human resource base. They bring in technical expertise and other intangible benefits such as their ideas that add value to the organization in many ways. From this perspective, it is plausible to argue that emerging evaluators in an organizational context serve to strengthen its programmes and other interventions offered thereby enhancing the overall quality of its services. The next section discusses the role of the evaluation community in supporting emerging evaluators to realize their potential as evaluation practitioners.

Definition of Terminology

Mentorship

The conventional understanding of mentorship is described as the relationship in which ex-

perienced, adult individuals help younger ones learn to navigate the world of work (Gomez, Ali and Casillas, 2014: 51). It is a relationship premised on the understanding that one person with knowledge and expertise passes onto to another to enhance their career development and growth (Warnberg et al, 2003). Mentoring relationships are primarily focused on a proteges' professional development and undergo various changes overtime, as they are not static relationships. Notably, mentors choose the level of involvement they desire with their proteges as the roles are defined as the relationship grows and interpersonal bonds are strengthened (Guroian, 2008:79). At this point, the relationship has ceased to be a one-way to one of interaction as peers in the same profession, resulting into mutually beneficial exchanges. As such, mentoring in the context of this article, is understood to mean a developmental relationship that is embedded in a career context (Ragins and Kram, 2007: 5). This is the distinguishing characteristic of mentorship from all other personal relationships.

Opportunities

In the context of this article, the term opportunity is defined broadly as the type of professional experiences and privileges available to an emerging evaluator to participate in an evaluation activity (O'Brien, 1999:3). It also means access to evaluation-related educational training for professional development as an evaluator (ibid). However, to further enhance our understanding of opportunity, there are two key factors involved, first, choice and second, affordability. Opportunity impacts these two factors differently. As far as choice is concerned, opportunity is determined by equality of access to professional evaluation education opportunities (Podems et al, 2014:86). Access to these opportunities is controlled by mechanisms most often beyond the control of an emerging evaluator

such affordability, which makes access to opportunities unequal. Affording the costs of professional growth can be inhibitive for most emerging evaluators as they are often out of reach for most of them. Therefore, affordability is strongly proportionate to the issue of equal opportunity. Stated differently, only those emerging evaluators that have the means to participate will have the opportunity to advance professionally. Furthermore, if the means are not available, affording a professional development as an emerging evaluator is an impossibility without financial aid and institutional support from an employer.

Evaluation Capacity Development

According to Labin et al (2012:328), evaluation capacity development is a "complex phenomenon involving issues of individual learning, organizational change, sustained change, program processes and outcomes". However, in the context of this article, the phrase is taken to mean providing staff in an organization the skills to conduct rigourous evaluations and doing this in a way that both acknowledges the local context and ensures that such evaluations become part of the routine practice of the entity (Naccarella et al, 2007: 232). This further means that organizations must have sufficient level of resources to ensure such practices are rooted as part of the organizational culture. This assertion is further underscored by Preskill (2010:224) who argues that evaluation capacity development is about "increasing the sustainability of professional evaluation practices" in an organization. On his part, Ba Tall (2009:123), suggests that evaluation capacity development is a continuing long-term process of learning and change management. On their part, Lennie and Tacchi (2013:94) contend that evaluation capacity development is part of institutionalizing evaluation and creation of an evaluation culture in an organization. They see process as vital for emerging evaluators as part of enhancing their professional development goals.

Competencies

Generally, competencies are defined as those performance capabilities needed by an individual to demonstrate knowledge, skill and ability (Dooley et al, 2004: 317). Deist et al (2005:30) further claims that the term encompasses possession of the necessary attributes to perform competently. However, this definition is narrow and reductionist as it suggests that context does not matter. I argue that context does matter in defining an individual competencies, even though competencies are centred on individuals, they cannot be viewed independent of the context in which individuals work. Podems et al (2014:83) asserts that people do not have competencies independent of their contexts, meaning the environment in which they work. The constructivist and interpretative paradigms, define competencies from a phenomenological point of view, suggesting that competency is a function of the environment in which it is practised, where the "worker and work form one entity through lived experience of work" (Sandberg, 2000:50). This broad view of the term concurs with Hoffman's definition which describes competency as the positive combination of knowledge, ability and willingness of an "individual to cope successfully and responsibly in changing situations" (Hoffmann, 2013:3). Stated differently, competence is active handling of knowledge in order to execute professional responsibilities or specific tasks in a given role or position (Ennis, 2008:4). In the context of this article, competencies for emerging evaluators are a combination of specialist knowledge, personal abilities, individual attributes and willingness to carry out specific tasks in the context of their work.

Prerequisites of an Emerging Evaluatior

Wisdom emerges when theory meets practice through deep reflection and applied knowledge. It is not enough to gain specialist knowledge without determination to apply it. Ramirez et al (2013:5) posits that emerging evaluators require mentoring support, one in which experienced evaluators mentor upcoming evaluators, creating an environment where everyone learns together. However, Conley-Tyler (2005:5) argues that for emerging evaluator to develop professionally the following personal attributes are critically vital to possess, namely:

a) Willingness to learn

Every emerging evaluator must be willing to learn and explore possibilities for growth. This attitude is important for accessing growth opportunities available within an organization as well as those outside of organization such as conferences, workshops and seminars to mention a few. It is imperative that an emerging evaluator possess a teachable disposition in order to gain insights and acquire knowledge previously unavailable before a task was undertaken (Patton, 2010:46).

b) Knowledge of organization's program and operations

An emerging evaluator must be well versed with the organization, its vision, programs, philosophies, policies, procedures, personnel and management. Inadequate understanding of organization's programs limits an emerging evaluator capacity to conduct realistic evaluations with plausible recommendations for actions (Braskamp et al, 1987:66).

c) Knowledge of context

Connected to the above point, understanding the context in which the organization operates.

An emerging evaluator ought to appreciate that context is not static, it is constantly in motion being influenced by different factors the political, financial and cultural contexts. Importantly, an emerging evaluator ought to be adaptable to changing contexts and understanding the organization's work in those environments. Faase and Pujdak (1987:81) claims that an emerging evaluator must be attuned to the uniqueness of changing contexts, possessing a heightened sensitivity in order to interpret an organization's work correctly. Moreover. It is this through this understanding of the context that emerging evaluators see the connectedness and wholeness of the organizations' work, thereby able to appreciate that all environments are interrelated and connected to each other.

d) Willingness to receive feedback

Another vital attribute of an emerging evaluator is willingness to appreciate feedback. Whilst feedback is important for one's learning, it can be positive or negative depending on the way it is given. In most instances, feedback is a consequence of performance, as it is given to (Hattie and Timperley, 2007:81):

assess one's work;
improve future assignments;
understand subject matter better;
inspire confidence;

In this case, an emerging evaluator has an opportunity to confirm, add to, overwrite or restructure information as one deems it fit to make sense. Since feedback does not have any effect in a vacuum, it must be received as a gift by an emerging evaluator for one to apply it and see its influence on one's work (Race, 2001:3). Kluger and DeNisi (1996) argue that great results are observed when there is high commitment from an emerging evaluator to actively use any feedback received from a skilled and experienced evaluator. This understanding is essential as it aids an emerging evaluator to comprehend succinctly the discrepancy between the current and desired understanding about the tasks performed.

e) Accountable

An emerging evaluator needs to be accountable to oneself and colleagues. It is an intrinsic attribute that makes a person responsible for their own actions and commit to deliver results on any endeavours wholeheartedly. In the context of this article, to be accountable means commitment to see tasks accomplished as required timeously. It also implies that one is able to provide a recovery plan if a task is not accomplished on time or goes off track. As such, an accountable emerging evaluator is always ready to take ownership if a task is not completed on time as planned or the resultant effects are deviant from the expected outcomes.

f) Flexibility

Flexibility in this paper means the extent an emerging evaluator is responsive to the context in which one is operating. It also means that one is able to adapt to changing work environments in order to achieve the desired results. Flexibility allows an emerging evaluator to see things from different perspectives, which enriches the quality of work (Patton, 2010:45). However, since nothing in the evaluation context is static, inflexibility results to stuckness and can lead to non-performance on the part of an emerging evaluator. Moreover, inflexibility limits interaction between an emerging evaluator and colleagues further minimising opportu-

nities for learning and improvement of one's work.

g) Availability

An emerging evaluator must be readily available to undertake any evaluation assignment upon request (Conley-Tyler, 2005:6). This attribute allows an emerging evaluator to take advantage of every opportunity to learn and enhance their own professional capacities. Conversely, unavailability limits an emerging evaluator exposure to growth opportunities and strains relationships with colleagues, especially in a team context.

BARRIERS AND CHALLENGES EXPERIENCED BY EMERGING EVALUATORS

Emerging evaluators like many professionals experience hurdles on their way up to become recognized professionals in the field of their expertise. Whilst this is not a new phenomenon in career development, the concerns are the following aspects, namely:

Organizational Context, Culture and Capacity

The demands of an evaluation responsibility in an organization can be overwhelming and misunderstood. Often, colleagues who do not understand the role of evaluation and evaluators in an organization, perceive evaluation as "policing", a profession aiming to find out any ills that other staff members may be involved in. The view of the profession is rooted in ignorance, more often on the part of staff and sometimes perpetuated by organizational leadership. The resultant effect is that it limits growth possibilities of an emerging evaluator, who has to face attrition tendencies from fellow colleagues. Moreover, if an organization does not promote utilization of evaluation results, it limits its own learning capacity, a feature that stifles an emerging evaluator ability to entrench evaluation results to influence organization's programming (Ortenbald, 2013:25). Therefore, organizational culture and context can impede an emerging evaluator's professional growth and development.

Lack of exposure to professional development opportunities

Generally, a sustained lack of exposure to career and professional development opportunities reduces the professional edge of any professional over a time period. It is essential that a professional accesses opportunities to sharpen his/her skills and knowledge. However, if an emerging evaluator does not have opportunities to interact with colleagues in the field through educational activities such as conferences, workshops and seminars to ensure they stay on the professional edge. The exposure increases an emerging evaluator's confidence as one interacts with colleagues in different settings and share thoughts as well as ideas about professional issues in their field.

Inadequate professional experience

Any profession is driven by expertise. In instances where a professional does not have adequate expertise, it becomes challenging to execute one's responsibilities effectively with confidence. An emerging evaluator faces such limitations especially if one is not adequately acquainted with technical experience such as data collection, its analysis, its management as well as development of monitoring systems. This limitation can severely impede professional development and can be exacerbated if an emerging evaluator is not given the requisite training to develop this technical expertise.

Lack of incentives to grow professionally

Incentives to grow professionally are critical to an emerging evaluator professional development and career growth. Organizations that do not encourage their staff members to grow

professionally, risk to be stagnant as there is no motivation to be creative and innovative. Like any professional, emerging evaluators need incentives such as sponsorships to attend professional monitoring and evaluation events such as conferences, workshops and seminars. The effect of such investments incentivises them to aspire higher heights in their professional development, especially if they are self-motivated intrinsically as individuals and are goal-driven. Organizational environments that are supportive of personal and professional development, are highly likely to create cultures in which high levels of motivation among staff, including emerging evaluators can flourish.

Shifting organizational policies regarding programmes

The complexity of an organization's interventions can necessitate changes aimed at redirecting its efforts. However, proper organizational structures must be established in order to support the changes. Emerging evaluators are often not part of organizational decision-making processes and may find themselves in a precarious condition when planned policies affect programming. In those instances where policies are changed frequently may cause staff instability and distraction from priorities. Emerging evaluators are prone to such distraction. Importantly, whilst introducing organizational policies regularly may be useful in the short, such structural changes may pose serious threats on capacity of emerging evaluators to adapt and still deliver quality services. In order to mitigate for such policy shifts, it is imperative to involve staff in organizational development processes to increase their legitimacy within the organization as well as enhance ownership among staff.

Inadequate financial resources in the organization

Globally, donor funding to organizations is diminishing rapidly. The resultant effect of this trend is strained organizational capacity to support emerging evaluators' professional growth needs. In most cases, organizations prioritize other aspects of their core work that they deem important and staff development may not be one of them. Dwindling financial flows into organizations may pose professional growth challenges for emerging evaluators. In this case, emerging evaluators must be willing to invest own resources and not rely solely on their organizations for their professional development.

The Role of the Evaluation Community

As discussed thus far, emerging evaluators are a vital human resource in an ever-growing and dynamic development world. As such, the evaluation community have a role to play to ensure that these human resource capital is well-nurtured to ensure they stay the course and provide high quality services. Therefore the evaluation community is thus obligated professionally to identify common stumbling blocks that impede their growth. They must also explore ways in which such stumbling blocks can be addressed to facilitate their progress as professional evaluators.

Moreover, the evaluation community has a professional obligation to assist emerging evaluators in establishing professional networks amongst themselves so that they can interact, exchange ideas, good practices and discuss collaboration amongst themselves (Coggshall et al, 2012:20). In this way, emerging evaluators experience and practice sharing insights that unlocks their professional potential through insights that they gain through such networks. Connected to the above point, evaluation community should create platforms through which

emerging evaluators can share their own work, insights and experiences with a wider evaluation, scholarly and development community (Zukoski and Luluquisen, 2002:12). It is in these engagement platforms such as professional evaluation conferences, seminars and workshops to mention a few, where emerging evaluators gain professional insights while they bring in their new perspectives and ideas.

Additionally, the evaluation community needs to offer mentoring opportunities to emerging evaluators to guide them as they navigate their way in the profession. Mentoring facilitates understanding of the ethos of the profession, especially when seasoned and experienced professional evaluators genuinely engage themselves to pour themselves into their protégés (Guroian, 2008:14). This professional relationship, when defined clearly and expectations spelt out upfront and agreed upon, is a rewarding experience for both mentors and protégés. Technical aspects such as methodological challenges, data analysis, data presentation and report writing are addressed in such relationships and overtime an emerging evaluator increases his or her professional capacities in these vital evaluation areas.

Conclusion

This article has argued that emerging evaluators are a critical professional human resource in any organization. They provide essential technical services that are beneficial to an organization in many ways. In a rapidly changing and competitive professional context, emerging evaluators need to be guided during their normative years of their careers so that they can realize their potential. Against this perspective, this article calls for a change in mind-set on the part of all key stakeholders in the evaluation field in responding to the daunting responsibility by the evaluation community in creating an environment that enables emerging evaluators to grow and thrive. Similarly, organizations should provide emerging evaluators with professional development opportunities to enhance their knowledge, ideas and skills confident execution of their responsibilities. As such, given the necessary professional support, emerging evaluators have the capacity to be innovative and offer fresh perspectives that could improve an organization's work significantly. They thus deserve to be supported to develop their capacity to their full potential.

Bibliography

Ba Tall, O. 2009. The Role of national, regional and international evaluation organizations in strengthening country-led monitoring and evaluation systems. In Segone, M. (ed) Country-led Monitoring and Evaluation Systems. Better Evidence, Better Policies, Better Development Results. Geneva: UNICEF. 119-34.

Braskamp, L.A., Brandenburg, D.C. and Ory, J.C. 1987. "Lessons about clients expectations. In Nowakowski, J. (ed). The Client Perspective on Evaluation: New Directions for Program Evaluation, No. 36. Jossey-Bass, San Franscisco. Page 66.

Carden, F. 2007. The Real Evaluation Gap. Alliance Vol. 12(4): 53-4.

Coggshall, J., Rasmussen, C, Colton, A, Milton, J and Jacques, C. 2012. Generating Teaching Effectiveness: The Role of Job-Embedded Professional Learning in Teaching Evaluation. http://www.tqsource.org/Publications/JEPD%20Issue%20Brief.pdf accessed on 10/09/2015.

Conley-Tyler, M. 2005. A Fundamental Choice: Internal or External Evaluation? Evaluation Journal of Australasia. Vol. 4 (new series). Nos. 1&2. Pages 3-11.

Dabelstein, N. 2003. Evaluation Capacity Development: Lessons Learnt. Evaluation Vol. 9(3):365-9. Deist, F. and Winterton, J. 2005. What is Competence? Human Resource Development International

Vol.8 No.1. Pages 27-46.

Donaldson, S. and Christie, C. 2006. Emerging Career Opportunities in the Transdiscipline of Evaluation Science. In Donaldson, S.I. Berger, D.E. and Pezdek, K. (Eds). Applied Psychology: New Frontiers and Rewarding Careers. Mahwah, New Jersey. Pages 243-259.

Ennis, M. 2008. Competency Models. A review of literature and role of employment and training Administration. U.S. Department of Labour.

Forss, K, Kruse, S, Taut, S. 2006. Chasing a ghost? An essay on participatory evaluation and capacity development. Evaluation Vol. 12(1): 128-44.

Gomez, R. L., Ali, A., & Casillas,W. (2014). Mentorship and the professional development of culturally responsive evaluators in the American Evaluation Association's Graduate Education Diversity Internship (GEDI) program. In P. M. Collins & R. Hopson (Eds.), Building a new generation of culturally responsive evaluators through AEA's Graduate Education Diversity Internship program. New Directions for Evaluation, 143, 49–66.

Guroian, V. 2008. Literature and real meaning of mentorship. Center for Christian Ethics. Baylor University. Pages 1-8.

Hattie, J. and Timperley, H. 2007. The power of feedback. Review of Educational Research. Vol. 77. Pages 1-31.

Hoffman, T. and Adombent, M. 2013. The Concept of competencies in the context of education for sustainable development. http://www.esd-expert.net accessed on 03/09/15.

Labin, S. Duffy, J, Meyers, D. 2012. A research synthesis of the evaluation capacity building literature. American Journal of Evaluation. Vol. 33 (3):307-38.

Lennie, J, and Tacchi, J. 2011. Researching, monitoring and evaluating communication for development: Trends, challenges and approaches. Report on a literature review and consultations

with Expert Reference Group and UN Focal Points on C4D. Prepared for the United Nations Inter- Agency Group on Communication for Development, New York: UNICEF. http://www.unicef.org/cbsc/files/RME-RP-Evaluating_C4D_Trends_Challenges_Approaches_Final-2011.pdf accessed on the 10/09/2015.

Lennie, J, Tacchi, J, Wilmore, M and Koirala, B. 2015. A holistic, learning-centred approach to building evaluation capacity in development organizations. Evaluation. Vol. 21(3):325-343.
Naccarella, L. Pirkis, J, Kohn, F. 2007. Building Evaluation Capacity: Definitional and practical implications from an Australian case study. Evaluation and Program Planning. Vol. 30(3):231-236.
O'Brien, B. 1999. What is opportunity? Defining, Operationalizing and Measuring the goal of postsecondary Educational Opportunity. Institute of Higher Education Policy. Boston, MA. Pages 3-28.
Ortenbald, A. 2013. What do we mean by "learning organization"? In Ortenbald, A. (ed) Handbook of Research on the Learning Organization: Adaptation and Context. Cheltenham: Edward Elgar. Pages 22- 34.
Patton, M.Q. 2010. Future Trends in evaluation. In Segone, M. (ed). From Policy to Results. Developing Capacities for Country Monitoring and Evaluation Systems. New York: UNICEF. Pages 44-57.
Podems, D, Goldman, I and Jacob, C. 2014. Evaluator Competencies: The South African Government Experience. The Canadian Journal of Program Evaluation. Vol.28 No.3. Pages 71-85.
Preskill, H. 2010. Exploring effective strategies for facilitating evaluation capacity building. In Segone, M. (ed) From Policy to Results. Developing Capacities for Country Monitoring and Evaluation Systems. New York: UNICEF. Pages Pages 224-38.
Ragins, B. and Kram, K. 2007. The roots and meaning of mentoring. Journal of Vocational Behaviour. Vol.89:1-13.
Ramirez, R. and Broadhead, D. 2013. Utilization Focused Evaluation. A Primer for Evaluators. Southbound.Sdn.Bhd. Penang, Malaysia. Page 5.
Rossi, P. H., Lipsey, M. W., and Freeman, H. E. 2004. Evaluation: A systematic approach (7th ed.) Thousand Oaks, CA: Sage.
Wilcox, Y. and King, J. 2014. A professional grounding and history of the development and formal use of evaluator competencies. The Canadian Journal of Program Evaluation. Vol.28 No.3. Pages 1-28.
Zukoski, A. and Luluquisen, M. 2002. Participatory Evaluation. What is it? Why do it? What are the challenges? Policy and Practice. Issue. No.5. Pages:1-6.

Harmonising M&E Systems: Positioning the data atom at the centre.

Prof. Johan Olivier and Mr Georg Bernhoeft
Particip - GIZ

Prof. Johan Olivier
johan.olivier@particip.com
083-452-5539

Abstract

A recent assessment of government M&E systems in South Africa "…highlighted the varied and contrasting conceptualisations of M&E systems across sectors and departments. These, at times amorphous, understandings of M&E systems may make it too difficult to develop a standardised M&E system diagnostic that applies equally appropriately across the different systems"(p. vi). Furthermore, the assessment found that "While there have been advances in M&E systems in government over the past two decades, data governance remains weak … data needs are not well planned for and co-ordinated and there are huge inefficiencies and duplication in fragmented forms of data collection…."(p. iii) [*1] .

This paper is based on our current work designed to harmonise M&E systems across government. This assignment, funded by GIZ, is a joint initiative between the DCoG, DPSA, National Treasury, and the DPME in the Presidency. Our current assessment of M&E systems has shown that M&E systems have traditionally been developed and designed from the reporting requirement perspective. Our central argument is that this is the wrong approach and that such initiatives are doomed to fail. In order for M&E systems to be successful, the design starting point should be the data requirements – up to the level of the data atom, the smallest piece of information. Today's technological advances have made data alive, self-aware, and with its own rules.

Our approach in harmonising M&E systems has been designed around this realisation and consists of a number of inter-related work-package requirements where the "data" work-package is the point of departure and the "reporting" work-package is the end-product. The intermediate work packages are preconditions for a successful M&E system and include, among others, defining roles and responsibilities, data governance issues, and project/programme objectives and implementation strategies, as well as project and performance tracking. It is only when all these have been correctly executed that M&E and its associated systems, and ultimately reporting, will be successful.

1. A Study into the State and Use of Monitoring and Evaluation in Government: Synthesis Report and Strategic Proposals for Continuous Improvement, DPME, November 2014.

Introduction

Monitoring and Evaluation (M&E) systems enable governments and development agencies to measure and assess their performance as well as their capacity to deliver the services as set out in their mandates. In addition, M&E systems enable them to measure the outputs (goods and services) provided and give an indication of the extent to which these outputs contribute to the desired results, outcomes and impacts. Furthermore, an M&E system also aims at enabling the development partners to better understand the causes of good and poor performance. Consequently, effective M&E systems are an important element to improve service delivery - also at decentralized level.

The decision to launch an intervention to harmonise M&E systems across the South African Government was made after it was agreed that significant challenges confront Government in this area. The intervention, which is a joint effort between the Department of Planning, Monitoring, and Evaluation (DPME) in the Presidency, National Treasury (NT), the Department of Cooperative Government (DCoG), and the Department of Public Service and Administration (DPSA), was launched in April 2015 and funded by the 'Governance Support Programme (GSP) [*2] which is implemented by GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit) on behalf of the German Government. The formal brief for this intervention is to: Harmonise national and provincial M&E systems for local government. In other words, the harmonisation efforts should have real implications and benefits for local government.

This paper is based on our current work designed to harmonise M&E systems across government. The paper documents the results of our assessment of M&E systems in the South African government and our approach to develop a solution.

The paper begins with an overview of the motivation behind this initiative. This is followed by a brief review of the results of our review of current M&E systems after which we present our approach to the harmonisation of M&E systems. The paper is concluded with a few summative remarks and implications for future system development.

Background

The Diagnostic Overview of the National Planning Commission conducted in 2011 identified nine areas of weakness, which hamper development in today's South Africa. Many of them are linked (as root or effect) to local service delivery, i.e. high rates of unemployment, substandard quality of basic education, poorly located and inadequate public infrastructure, spatial divide of communities, an ailing public health system, an uneven performance of the public sector and wide-spread incidences of corruption.

Significant strides have been made since 1994 in the establishment of M&E systems for project and programme management across all spheres of government. The publication of the "Policy Framework for the Government-Wide Monitoring and Evaluation (GWM&E) System" by the Presidency in 2007 was a significant development in public sector management.

---

2. The Governance Support Programme (GSP) is a partnership programme agreed to between the Governments of South Africa and Germany. The programme provides technical, policy and process advice to support the South African government in remedying systemic shortcomings. The objective of GSP is that public institutions have improved their service delivery, in cooperation with the private sector and civil society.
3. The Presidency: DPME, National Evaluation Policy Framework, 23 November 2011, p. iii.

While effective monitoring and evaluation (M&E) is recognized as being an essential part of sound project and programme management for many years, it has seen an increased formalisation across government as a whole and been transformed into a "Government-wide Monitoring & Evaluation System" (GWM&E). The National Evaluation Policy Framework (NEPF) was approved on 23 November 2011. This set out the approach in establishing a National Evaluation System for South Africa. The purpose of the NEPF is " … to ensure that credible and objective evidence from evaluation is used in planning, budgeting, organisational improvement, policy review, as well as on-going programme and project management, to improve performance." *[3]

The underlying purpose is:

- Improving policy or programme performance - providing feedback to managers;
- Improving accountability for where public spending is going and the difference it is making;
- Improving decision-making e.g. on what is working or not-working;
- Increasing knowledge about what works and what does not with regards to a public policy, plan, programme, or project.

The NEPF envisages a National Evaluation Plan (NEP) which is updated annually including the key interventions across government which are seen as a national priority. The NEPF together with the National Evaluation Plan 2015-2018 provide further context within which the current intervention takes place.

The establishment of a GWM&E System is an important step in measuring the performance of the government and – on the basis of the assessed performance strengths and weaknesses – improve the delivery of goods and services. The nature and quality of the data obtained however depend on the effectiveness, efficiency and quality of the different M&E systems. The M&E systems are meant to be connected and harmonized across government institutions, which is however not always the case.

Despite the progress that has been made to date in the area of GWM&E, the results of recent assessments show that much remains to be done to optimise M&E in the South African context. A recent assessment of government M&E systems in South Africa "…highlighted the varied and contrasting conceptualisations of M&E systems across sectors and departments. These, at times amorphous, understandings of M&E systems may make it too difficult to develop a standardised M&E system diagnostic that applies equally appropriately across the different systems"(p. vi). Furthermore, the assessment found that "While there have been advances in M&E systems in government over the past two decades, data governance remains weak … data needs are not well planned for and co-ordinated and there are huge inefficiencies and duplication in fragmented forms of data collection…."(p. iii) *[4] . Also, a core problem that has been identified is that government policies and development planning are not always sufficiently informed by credible and recent data.

4. A Study into the State and Use of Monitoring and Evaluation in Government: Synthesis Report and Strategic Proposals for Continuous Improvement, DPME, November, 2014.

5. Thembisile Serite and Andreas Wagner, "Implementing M&E in local municipalities: Experiences from Ehlanzeni District Municipality", paper to be presented at the SAMEA conference, 12-16 October 2015.

Analysis of current systems

The assessment of current systems was conducted by way of two interrelated set of activities. They are: (1) meetings with key stakeholders at national level as well as with officials in the Eastern Cape and Mpumalanga provinces – in particular with officials from the Offices of the Premier (OtP), CoGTA and selected District (DM) and Local Municipalities (LM) in these two provinces, and (2) a review of relevant documents, publications and reports of similar assessments (see also Serite, T. and Andreas Wagner, 2015) [*5].

Together these highlighted that the existing M&E systems are often characterised by:

- A diverse set of M&E systems across government that are characterised by various levels of maturity;
- M&E systems at provincial level, and especially District and Local Municipalities, are particularly weak. M&E Frameworks are either absent or poorly developed, results and indicators are poorly defined and that there is often no link between the M&E systems at the lower spheres of government and those that exist at the higher spheres, i.e. the MTEF and its Outcomes. In addition, local government M&E activities are often not integrated into the GWM&E system.
- A lack of coordination across the spheres of government as well as horizontally among sectors which often results in duplication of efforts;
- Insufficient coordination within departments – especially as it relates to the required interfaces between strategic planning, M&E, and budgeting;
- A gap exists between M&E systems and the legal mandates of departments;
- Incoherent M&E frameworks at the various spheres of government;
- A lack of credible and integrated information pertaining to service delivery implementation and its outcomes;
- Poor data governance at all spheres of government;
- Limited access to available information by officials responsible for implementation and policy makers alike;
- Fragmented and duplicated data collection and reporting and continuously changing data requirements at all spheres of government;
- Inadequate analysis of and responsiveness to available information by officials at the various spheres of government responsible for implementation;
- Inconsistent definition and measurement of indicators at all spheres of government;
- Inadequate customisation of implementation support from national and provincial government;
- Insufficient buy-in and accountability by the various stakeholders such as communities, officials and politicians (especially at local level); and
- Inadequate service delivery capacity and the associated skills within the various organs of the state. This in particular relates to weak M&E systems that are not able to provide the required management information to officials responsible for service delivery.

The current intervention was designed to address these concerns in collaboration with national and provincial partners. It was agreed that the solution will be developed and piloted in the Eastern Cape and Mpumalanga provinces around Outcome 9.

Approach and proposed solution

The assessment of current M&E systems highlighted several issues that impact negatively on the existence of efficient and effective monitoring and evaluation across government. The need for more coherence in the approach to M&E was highlighted by all stakeholders. The approach to M&E harmonisation has been designed to address the core weaknesses that were identified. These weaknesses in particular relate to data governance, M&E frameworks and systems, as well as reporting.

The six areas of change management

Our intervention is based on the recognition that in order for the harmonisation project to be successful, six key areas of change management need to be managed in a coherent approach. These areas are:

Business Processes - The desired state of operation, opportunities to improve current processes and provide business performance, how processes have to be aligned with the M&E Harmonisation, the expected results, the sequence of required activities, and the associated M&E harmonisation business rules.

Organisation - Capabilities, roles, team structures, organizational units, and training of alignment to operate the SA-department structure.

Location - Needed in terms of geographic distribution of processes, people, technical infrastructure, data, and applications. Need for physical facilities. This includes the use and the management of IT and other communication technologies.

Data – System-wide information content and structure requirements, rules & properties for M&E data, data storage, access optimization for efficiency & performance, technology, security to safeguard information confidentiality, integrity & availability.

Application - Application design to fulfil the requirements, application developed integration, testing, and deployment.

Technology - Needs on hardware, system software, and communication network components to support the IT business and its systems.

The seven work packages

It is important to recognise that the approach to harmonise M&E systems across government consists of interdependent building blocks/work packages. We will discuss each of these building blocks individually. Together these building blocks address the core issues that inhibit effective monitoring and evaluation.

The work packages identified in this approach are the following (see also Figure 2 below):

- Roles & Responsibilities and Workflows & Processes;
- The Spine;
- Strategy, Results, Indicators and Project Tracking;
- Harmonisation of M&E frameworks;
- Monitoring & Evaluation;
- M&E Systems; and
- Reporting.

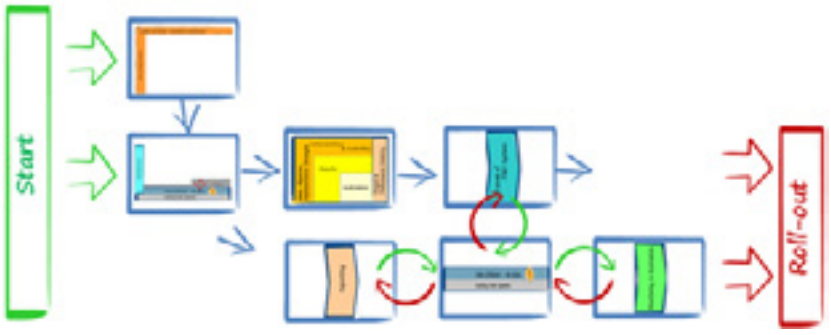


Figure 2: The work stream and flow of work packages

As is evident from Figure 2, these work streams are interrelated in a complex interaction. We will now discuss each of the work packages in greater detail.

Roles & responsibilities and workflows & processes

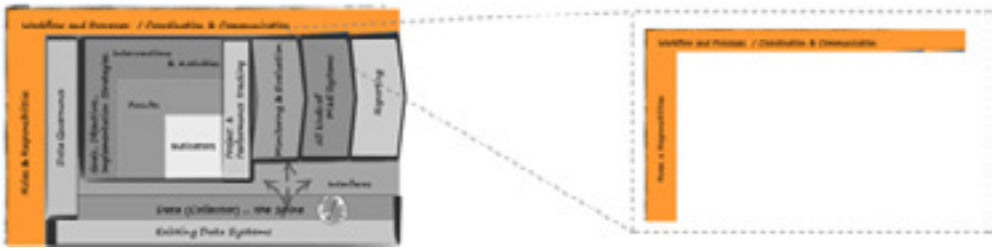


Figure 3: Roles & responsibilities and workflows & processes

A successful monitoring and evaluation system requires clearly defined roles and responsibilities. In addition, workflows and work processes, as well as their ownership have to be clearly defined. The M&E harmonisation process has to start with the analysis of roles and responsibilities, workflows and processes, because it has to be clear who is interacting with whom. The analysis of the current situation will also identify missing links, misunderstanding or processes the department is not aware of. It will highlight communication bottlenecks and identify coordination improvement potential. This is done by working through the work packages and will highlight the interdependencies during this process.

It is important to recognise that, since the analysis of the roles and responsibilities/workflows and processes impact on all the other work packages, this analysis will be an ongoing task during the whole harmonisation intervention.

The "Spine"

The so-called "Spine" is in reality a software application that stores and harmonises M&E data in a cloud database. This highly complex software assembles modules that allow all kinds of databases and systems to provide data. These data are then converted into a cloud.

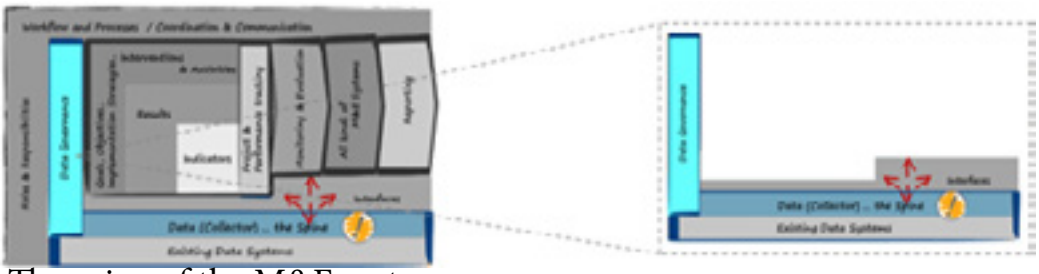


Figure 4: The spine of the M&E system

An important function of the "Spine" is the ability to provide data that are "self-aware". This means, the datasets are able to maintain themselves. The data have a clear understanding of who is responsible for this data and how sustainability is ensured.

Data Collector: The "Spine" is also self-acting to collect the correct data from the correct data sources, once defined. That is the reason why the "Spine" is also called "the data collector". The system itself is able to communicate via various interfaces with other systems, like M&E systems to provide qualified data for analysis purposes like a standard or an individual M&E report.

Existing Data Systems: The data collector will interact with the existing data systems to extract the required data. This requires an in-depth understanding of the existing systems that support the M&E functionality. A rigorous analysis of existing systems is therefore a key activity.

Interfaces: To connect existing data systems with the data collector, various interfaces are required to exchange data. The definition of the interfaces should not be a monolithic structure but rather flexible so that future M&E systems can be connected.

A vital factor for the success of M&E harmonisation is the management of the existing and future data used for M&E analysis, planning and reporting. If data governance is not implemented there will be no M&E harmonisation. The development of a data governance structure is another key activity of an M&E harmonisation intervention.

Governance Structures: The delivery of data by the Spine, the development of sustainable interfaces, and the evaluation of newly integrated data requires a strict operational procedure. This has to be managed by professionals solely responsible for data. In future there might be outposts in all departments to work closely with the data governance function.

Managing Data Quality: A key responsibility of the newly defined data governance unit/function is to provide high quality and sustainable data. To do so, data governance has to take over the responsibility to manage the Spine, the connection to other external systems and to the interface definition.

Strategy, results, indicators and project tracking

The successful implementation of an M&E system requires that the objectives and the expected results of a strategy are defined and clearly communicated with the lower level/local level. Only then can M&E take place, because it will only then be clear what has to be measured. This is related to departmental performance as well as the measurement of the success of the overall strategy. In the case of South Africa the 14 Outcomes of the MTSF represent the goals and objectives that need to cascade down to the various departments, provinces and local authprities.

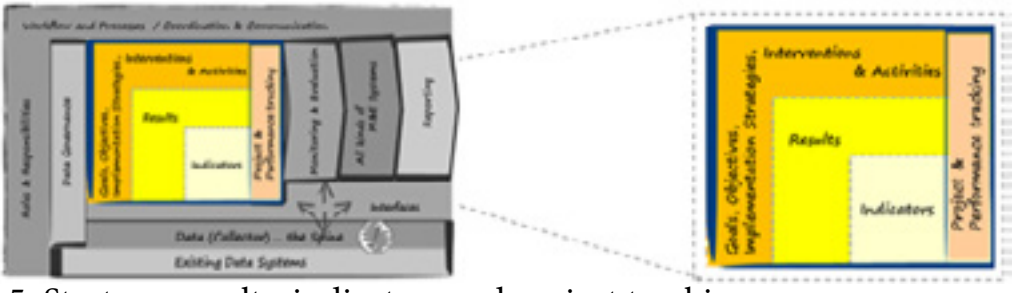


Figure 5: Strategy, results, indicators and project tracking

There is a high likelihood of success if the implementation strategies are well-designed and appropriate for the targeted operational objective. This requires also a clear understanding of roles and responsibilities and the vision, that all units contribute to a certain result to a particular objective. A successfully harmonised monitoring and evaluation system, based on clearly defined objectives, results and related indicators requires a results-based management approach.

Results-based management – the basis for M&E Harmonisation

Government Departments are responsible to fulfil the goals set by Cabinet and as expressed in the various outcomes (Outcome 1 to 14). To do so, departments develop their own long-term strategic objectives based on their Outcome responsibilities. These strategic objectives reflect the direction where departments are going to reach the outcome goals. To do so departments create Programmes and Services. These various Programmes and Services are now responsible to translate strategic objectives into reality by defining individual operational objectives and implementation strategies. Based on this, Programmes and Services are now able to deliver interventions to achieve the targeted strategic objectives of the department and in doing so to reach the desired goals as defined by Cabinet.

The Programmes and Services with their operational objectives, their implementation strategies and their interventions drive the organizational processes and resource planning and design and also the M&E requirements.

Strategy design

Strategy design needs to be done by clearly defining a set of key issues. They are the goals, the strategic objectives, the operational objectives, operational objectives, implementation strategies, interventions, and activities/operational plans.

The link between the operational plan and its detailed planned activities are the implemented projects in the project tracking system. Implemented projects may be monitored on local or national level.

Indicator management

The "spine" will not only manage a set of indicators, but also all data atoms required to produce harmonised M&E reports. Indicators are data or a combination of data collected and processed for a clearly defined analytical or policy purpose. That purpose has to be explicitly specified and taken into account when interpreting the value of an indicator. Indicators have to provide practical and cost-effective means to evaluate the state and the development of national or local issues and the effects that policy changes and interventions have on strategy or performance.

Indicators are not an end in themselves: Indicators are a tool to clarify assessments of and comparisons between developments over time. They describe in simple terms the extent to which the results of objectives set for sustainable development are being achieved. The main purpose in developing a set of sustainability indicators is to assess the performance of interventions in order to define activities to better pursue sustainability objectives.

Identification of expected results: Indicators are bound to results. They are the toolset to describe, if an objective and its results have been achieved.

Harmonisation of M&E frameworks

The harmonised M&E Framework seeks to define roles and responsibilities, toolsets and guidance, and to enhance the effective implementation of sound monitoring and evaluation activities across various departments and spheres of government.

The following a key success factors:

- Ensuring standardisation in the collection of monitoring data for improved data collation and analysis linked to the "Spine";
- Guiding results-based planning and reporting processes to utilise the "Spine";
- Enhancing the measurement and tracking of programmes and their operational objectives, implementation strategies, there interventions and activities to prove performance, impact and contribution to the Outcomes; and
- Improving programme evaluation, the drawing of lessons and utilisation of M&E information in improving the planning processes.

Monitoring & Evaluation

The optimisation and provision of timely, accurate data and information that is required for decision-making at various levels of programme implementation and management is another key success factor for effective M&E. This will form the basis for the "Spine" system design and improvements needed to satisfy information needs of the different stakeholders.

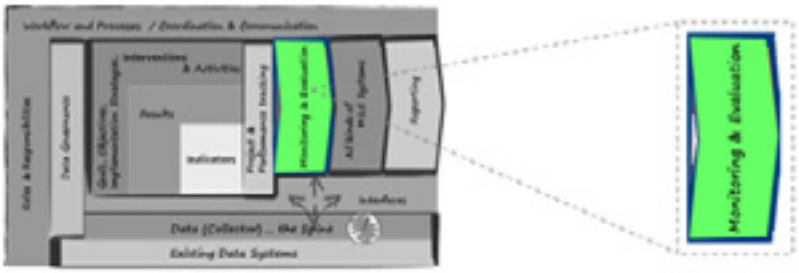


Figure 6: Monitoring & Evaluation function

Enhancing coordination between that various stakeholders with a view to harmonise M&E activities by different players and clearly delineate roles and responsibilities of the different actors in the M&E processes is another key success factor. Where appropriate, the M&E system need to be aligned to and integrated with other institutional systems and frameworks to enhance data collation, analysis and comparability, also involving hyper linking and interfacing with existing information systems. This entails alignment of programme plans, local integrated development plans with provincial, national level strategic plans and the overall government's programme of action, all of which contributes to Outcome 9.

The future M&E system which supports standardization of data collection templates, reporting formats, data storage and dissemination mechanisms, is a critical step towards alignment and M&E harmonisation. The importance of monitoring, reporting and evaluation as an integral part of planning cannot be overemphasized.

Existing M&E systems

When talking about an M&E system there is often confusion as to whether it also includes the software solution. This is not always the case.

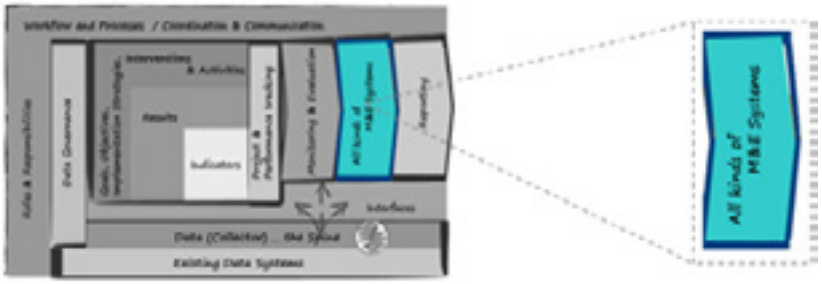


Figure 7: Existing M&E systems

The characteristics of an M&E system can be defined as a set of organisational structures, management processes, standards, strategies, plans, information systems, reporting lines and accountability relationships which enables the department, local government but also external partners, and other institutions to discharge their M&E responsibilities effectively.
Once the requirement of monitoring and evaluation is defined, the M&E system can be designed that performs the required functions.

Reporting

The final element of the M&E harmonisation is that the necessary processes are established to report on the results of performance measurement and M&E and that reporting commitments are met.

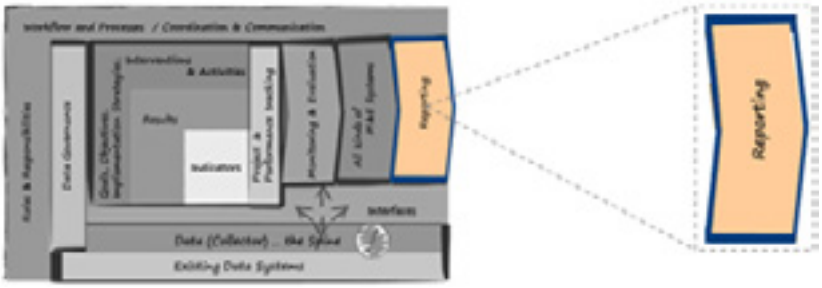


Figure 8: Reporting

"The Spine" will trigger the information on when performance and M&E reporting and periodic evaluation reporting will occur, by whom and how. This requires a clear M&E task portfolio that includes roles and responsibilities and M&E reporting processes. In order for this to be successful, the definition of two key elements are required:

- Management authority (this may involve multiple partners) responsible for reporting the performance information and the evaluation results; and
- Processes like annual progress reports, Departmental / Local Performance Reports, mid-term evaluations, summative evaluations as well as the timeframe for reporting performance information.

While a standard set of reports or views on data will be developed, this might prove to be inadequate for future use. Also, there might be the need to export data and present them in a different way. Therefore, the system will provide a set of standard interfaces to export data. These export data may then be individually used to create reports with standard tools like MS Excel.

The full work package

Figure 9 below presents the different components into the full work package. In order for M&E practises and systems to be effectively harmonised and operational, a complex set of

interdependent elements needs to be completed and tested. Furthermore, access to credible data forms the foundation of such a system.

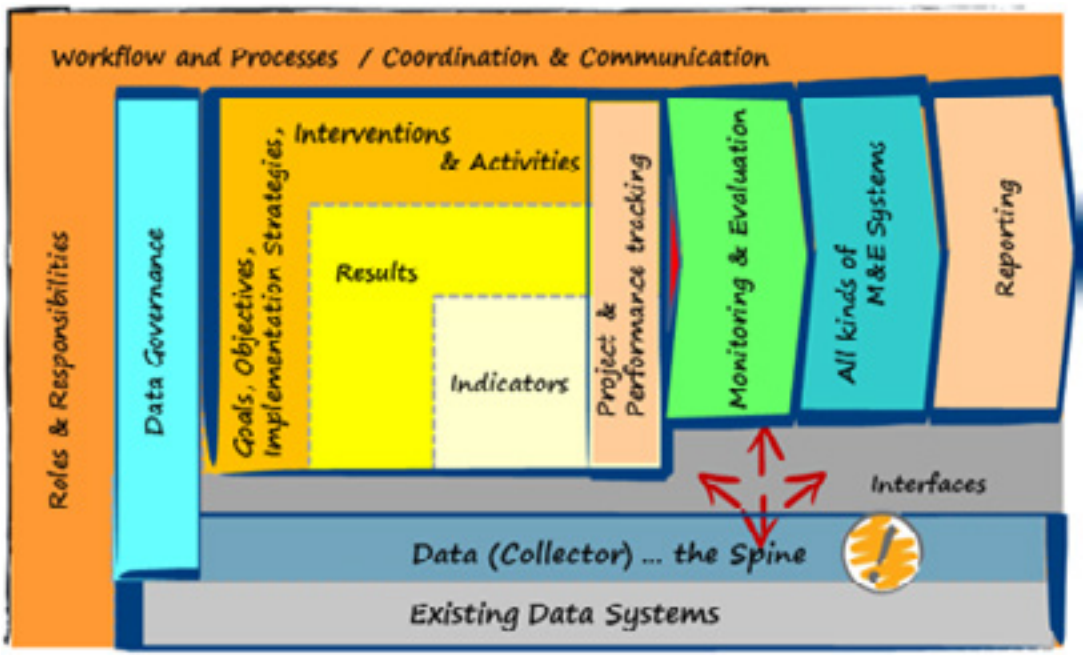


Figure 9: The M&E Harmonisation work packages

Conclusions

Our current assessment of M&E systems has shown that M&E systems have traditionally been developed and designed from the reporting requirement perspective. Our central argument is that this is the wrong approach and that such initiatives are doomed to fail. In order for M&E systems to be successful, the design starting point should be the data requirements – up to the level of the data atom, the smallest piece of information. Today's technological advances have made data alive, self-aware, and with its own rules.

Our approach in harmonising M&E systems has been designed around this realisation and consists of a number of inter-related work-package requirements where the "data" work-package is the point of departure and the "reporting" work-package is the end-product. The intermediate work packages are preconditions for a successful M&E system and include, among others, defining roles and responsibilities, data governance issues, and project/programme objectives and implementation strategies, as well as project and performance tracking. It is only when all these have been correctly executed that M&E and its associated systems, and ultimately reporting, will be successful.

A functional and effective M&E system is dependent on well-designed objectives, results and indicators. In tracking performance, strategy and impact, the M&E system will make use of sound data collection methods with multiple and comparable data sources being used to validate the findings provided by the Spine. The M&E harmonisation will contribute to good governance through enhancing transparency, accountability and community participation. This measurement process is aided by well designed, reliable, verifiable, cost-effective, appropriate and result-relevant indicators.

It is expected that valuable best practise experience pertaining to the approach as set out above and the individual work-packages will become available during the implementation period of this intervention. We will document these on an on-going basis in collaboration with the various partners. This will serve as important reference points for future M&E system development and harmonisation efforts.

Professionalising Monotoring and Evaluation for Improved Performance and Integrity: Opportunities and Unintended Consequences.

Author
Richard M Levin
Principal: National School of Governmen

Abstract

Monitoring and Evaluation (M&E) has developed into an active discipline and practice in South Africa over the past two decades. This is evident in the number of M&E activities undertaken in government, civil society, academic institutions, private sector consultancies and donor agencies. Various institutions at the centre of government have also been mandated to ensure that M&E takes place government-wide, and Parliament and Legislatures have incorporated M&E into their oversight work. Academia has responded with a variety of post-graduate and undergraduate degrees as well as short learning programmes and curricula to accommodate the increasing demand for professional capacity in this area. Research is evolving. Following international trends, there is a definite need for South Africa to consider the opportunities and challenges involved in transforming M&E into a profession and establishing frameworks and codes to take Evaluation to the highest level of integrity. There is a view, which argues that professionalisation of M&E practices will contribute to the establishment of a recognised "Evaluation Profession", through a professional body and affiliated professionals. The consequence of this will potentially be differentiation between those who hold qualifications and are 'legitimately' working in the field of evaluation and those who are 'unqualified' but who have gained experience. It may also create hierarchical divisions within the occupational classes of evaluators and a new division between 'professionals' and 'amateurs'. To avoid pitfalls and unintended consequences, it is critical to look at how the profession or occupation will be specified and classified and the consequences this would have on existing practitioners and emerging communities of practice. The integrity of evaluation practice must be protected, recognising that the discipline itself is contested, but we should be cautious of over-regulation which may prejudice the historically disadvantaged and reproduce old race and university hierarchies in new forms. Professionalisation can potentially bring benefits for those who practice the profession. However, South Africa's historical context must be taken into account if we go this route. Professionalisation must be a transformative project sensitive to the unintended consequences, which avoids elitism and exclusion.

1 Not to be quoted without the author's permission.

Introduction

Since the establishment of a democratic state in 1994, Monitoring and Evaluation has developed into an active discipline in South Africa. Successive Presidents have highlighted the importance of monitoring progress and evaluating the impact of strategies and programmes. There are a large number of M&E activities undertaken in government, civil society, academic institutions, private sector consultancies and donor agencies. During the Presidency of Thabo Mbeki, a cluster system of governance was introduced to strengthen integration and vertical and horizontal coordination across sectors, spheres and departments of government. A government-wide system of monitoring and evaluation was conceptualised and introduced to try and strengthen practice and data collection with a view to establishing a dashboard for the President and Cabinet to focus on priority areas and track progress and impact. Under the Presidency of Jacob Zuma, a new department of Performance Monitoring and Evaluation was established under a Minister in the Presidency. Concomitant with these developments, Parliament and Legislatures have steadily incorporated M&E into oversight work.

Various academic institutions and scholars have responded with courses and curricula to the increasing demand for human capabilities in this area of work, and research is growing and developing. In line with international trends, there is a definite need in South Africa to deepen Evaluation theory and practice and to increase the rigour and usefulness of monitoring information in order develop the discipline and meet the demand both inside and outside of government. Internationally, there is growing interest and debate around the question of professionalising the discipline of evaluation and setting up various forms of accreditation for professional evaluators.

There is a contention that professionalisation of Evaluation practices will contribute to the establishment and maintenance of a recognised "Evaluation Profession", through a professional body and affiliated professionals. The professional body would introduce systems and oversee the accreditation of evaluators. The consequence of this will potentially be differentiation between those who hold qualifications and are 'legitimately' working in the field of evaluation and those who are 'not'. It may create further hierarchical divisions within the occupational classes of evaluators and a new division between 'professionals' and 'amateurs'.

This paper explores the challenges of professionalisation in South Africa, with an eye on the country's history of racial discrimination and privilege and how this has manifest itself both in government and society more broadly. Unless the material context is thoroughly scrutinised, it is likely that professionalisation of Evaluation will have the unintended consequence of reproducing historical inequalities and power relations.

Evaluation as a Profession

At the outset, it is important to pose the question: Is Evaluation a Profession? A profession is defined as "body of qualified persons in a specific occupation or field" or "a vocation, a calling, especially one requiring advanced knowledge or training in some branch of learning or science."[2] Research on the emerging discipline of Evaluation, demonstrates that it has

2 These definitions are drawn from the American Heritage Dictionary of English Language (American Heritage Publishing Company:1969) and the New Shorter Oxford English Dictionary (Oxford Dictionaries: 2002).

"acquired all the characteristics of a distinct discipline … it has a assembled a body of knowledge, a set of specialized skills and a compendium of ethical guidelines that must be mastered to do what evaluation peers will consider to be good work." [3]

The standards that would guide Evaluation practice differ from country to country given the wide range of socio-political and cultural environments within which evaluators operate. Nevertheless, guidelines developed by evaluation associations have much in common, such as commitment to systematic inquiry and integrity; an abiding respect for diverse peoples and cultures; an eagerness to achieve results, and an unwavering public interest orientation.

Nevertheless, the difference in opinions amongst those regarding evaluation professionally remains wide and deep.[4] For Michael Patton evaluation is a "demanding and challenging profession"; Ernest House agrees. He views evaluation as a "new profession." However, Lincoln refers to evaluation as a "fledging profession," whilst Rossi's opinion is that "evaluation is not a profession at all." [5]

There are a number of critical features which have been identified and which are most frequently referred to when characterizing a profession. These are: the possession of skills based on theoretical knowledge; the provision of education and learning courses and curriculum; the existence of a professional body or organization; the testing of the competence of its members by a professional body and, significantly adherence to a code of conduct stipulated by the professional organization.

Evaluation For Accountability, Learning and Managment

M&E reporting fulfils important accountability requirements for states and governments, development aid donors and civil society. Some countries and role players see M&E accountability and learning as interrelated, while others see accountability and learning as discrete and even at times in contradiction with one another.[6] There is unanimity, however around the need to strengthen the usefulness of evaluation feedback.

Linda Morra Imas and Ray Rist write in the Preface to their seminal work The Road to Results:

As pressures grow across the globe for accountability by governments and organisations for the consequences of their actions for greater responsiveness to internal and external stakeholders for their performance, and most profoundly for greater development effectiveness, evaluation is emerging as a key way in which to systematically address and answer the question: "So what?" It is not enough to document that one is busy, it is now a requirement to document that one is (or is not) effective.

Evaluation also is a component of Results-Based Management (RBM). Within the context of

3 Robert Picciotto, The Logic of Evaluation Professionalism, Evaluation Journal, SAGE, (2011: 165-180).

4 Bickman and Reich cited in Picciotto (ibid.)

5 M. Patton (1990), E. House (1993), Lincoln (1985) and Rossi et al (2004) cited in Picciotto (ibid.)

6 Evaluation Feedback for Effective Learning and Accountability, (OECD Workshop Proceedings, Tokyo 26-28 September 2000).;

public programmes, it is part of a management approachwhich aims at:

- Achieving development targets,
- Clarifying the roles and responsibilities of public servants,
- Increasing transparency and accountability in public affairs and budgets, and
- Using good-quality data to improve learning and decision-making .

Evaluation is Also a Machanism of Deomcratic Development

Participatory development evaluation is part of good governance, puts people first and provides an opportunity to find democratic solutions to development challenges. Based on the successes of popular participation including the Porto Alegre participatory budgeting experience in Brazil, there is a growing belief that effective planning is dependent on the capacity of citizens to participate in development implementation and evaluation. The value of people-centred development evaluation methodologies is that effective state action is best facilitated by the evaluation of outcomes and impacts that translate the learning experience of deliberative and engaged evaluation processes into strengthened participatory policy implementation.[7]

Popular participation not a technical requirement for Good Governance; requires particular capacities, strategies and engagement to achieve developmental objectives. The state must engage with civil society at a number of levels. At a national and state or provincial level, organized formations of civil society, trade unions and business leadership organisations in a variety of forms must be engaged. At a local level, the state must engage with communities and the structures and organisations that represent them to develop strategies to create material conditions for sustainable livelihoods and entrepreneurship.

It must draw on the creative capabilities of communities to engage in new scalable forms of petty productive and trading activities in agriculture, manufacturing and micro industry. It must also engage with communities on social services and governance processes related to both social and economic development. This demands a new role for the state beyond its traditional roles of top-down government-centred delivery of social services, security and large-scale state controlled enterprises.

Based on the Porto Alegre participatory budgeting experience, Erik Olin Wright's general design principles for "empowered participatory governance" include:

- Bottom up empowered participation (face to face community meetings);
- Pragmatic orientation (concrete problem solving);
- Deliberation (how decisions are made, with the best argument winning…);
- Devolution and decentralisation (decisions made at the locus of the problem);
- Recombinant decentralisation (connecting to supportive muscular central powers);
- State centred institutionalisation - to change their traditional character and nature; and
- Organisation of countervailing powers through popular mobilization, which reduces

7 Ray Rist et al, Introduction for Influencing Change: Building Evaluation Capacity to Strengthen Governance, 2011: 02

8 E. Olin Wright, Envisioning Real Utopias, Verso (2010).

the power advantages of the traditional elites. [8]
Outcomes-based evaluation can change the way in which governments work. Public Administration is a discourse and practice through which states have developed and implemented policies, which suit the state and government's own internal workings and interests . The focus on priorities, eliminates bureaucratic turf warfare and self-interest and empowers citizens to become active participants rather than the objects of development. Critical to unlocking the realization of this "public good" discourse is the identification of the "theory of social change" which underlies a particular outcome and pinpointing, prioritizing and sequencing the building blocks to achieve the intended result. Users and the use they make of knowledge and evidence to be generated by M&E processes should determine all aspects of evaluation, including the methods and approach to dissemination. The challenge is to ensure that the poor are empowered in a context where highly differentiated communities, become the objects of government and/or donor development initiatives. The popular democratic challenge is to ensure that the marginalised and voiceless become the subjects of "development."

The foregoing discussion on the different uses of evaluation, for accountability, learning, improvement and democratic deepening and transformation illustrates that any serious contemplation of professionalisation of evaluation and evaluators must consider its differential application and the different skills, competencies, applications and attributions of evaluation and evaluators. The distinction between evaluation and evaluators in the debate on professionalisation is also critical because it could be argued that in the case of the "democratic dividend," it is the process of evaluation that must be professionalized and democratized.

Role and Purpose of a Professional Body

The primary function of a professional body is to promote and support the particular profession. This includes two critical aspects. The first relates to the protection of the interests of the professionals themselves while the second relates to the protection of the public interest.

Professional bodies set best practice standards in their area of expertise. This would include collectively agreed and established ethical standards such as codes of professional conduct. This would require the professional body to take action when a member's conduct falls below the standard expected. Membership of a professional body is dependant on achieving certification or accreditation that demonstrates that the individual is appropriately qualified to work in a field. A qualification and certification that represents the professional qualifications required in an area, is based on a recognised body of learning. Professional bodies also fulfil the role of creating platforms for learning for academic disciplines related to the profession, and may be involved in educational and knowledge building activities. [10]

In addition to initial registration, members may be required to complete continuing professional education to maintain their membership. Professional bodies also provide services to their members, for example in the area of career development. They also provide an im-

9 J. Scott, Seeing like a State, Yale (1988)

10 See for example, Chartered Quality Institute (CQI), Specialist professional bodies, The role and purpose of professional bodies, accessed from http://www.thecqi.org/Knowledge-Hub/Knowledge-portal/Compliance-and-organisations/Specialist-professional-bodies/

portant arena for social and professional networking with peer professionals often achieved through the organisation of social activities. Professional bodies in cases also facilitate access to consultancy services that are relevant to their field of expertise. Interested parties may be referred to accredited or approved consultants. Consultant fees may or may not be charged. Professional bodies normally advise consultancy service users of the conditions under which such services are provided.

Scope of Work

The scope of work for professionalising evaluation needs to be clearly defined. Currently in the public service, for example, we have officials working on performance assessment, research, policy analysis, and programme evaluations, all of which are lumped as evaluation experts. Similar differentiated expertise and competence exists in the private sector as well as in civil society. Furthermore, there are content experts in various fields in all sectors including in health and education who also occupy the space of evaluation expertise.

The questions may well be: who is an evaluation professional and what exactly do they do and what do they not do; and do we want to create a defined "occupational closure" to those termed "Evaluation Professionals" defined by professional demarcation and grade?

The point is that professionalisation of evaluation would involve "professional evaluators" with a variety of specialisations, skills and competencies. This would possibly require an umbrella body of professional evaluators with a variety generic, evaluation specific as well as sectoral specific attributes, competencies and skills.

Qualifications, Their Limitations and Experience

Linked to the issue of the scope of work are the academic qualifications required. Individuals require specific qualifications to become members of a professional body. Currently, some evaluation experts have junior degrees in particular fields, followed by a post-graduate evaluation-specific, whilst others have learnt through on-the-job evaluations with non-NQF education. An 'acceptable' or should we say 'appropriate' qualification would need to be determined for an individual to become a member of a professional evaluation body. A clear career path would need to be set by the professional body for evaluators who want to reach a career pinnacle in evaluation. This may entail rigorous processes such as board exams or submission and publication of a minimum number of articles in accredited journals on a annual or biennial basis. Such individuals may further be accredited to open evaluation practices by the professional body.

Qualification requirements for the evaluation profession would need to be balanced by experience in the South African context because of the historical legacy of racial discrimination and exclusion. It would be necessary therefore to develop a Recognition of Prior Learning policy and standard for Evaluation and evaluators.

Challenges with Professionalisation based solely on Qualifiacations

South Africa's historical context of racial discrimination, economic exploitation and exclusion must inform any strategy of the professionalisation of evaluation. Over-emphasis on specific formal qualifications that neglect other indicators of competence including relevant

experience, alternative qualifications, on-the-job training or informal education will exclude highly competent people from the profession. Professionalisation in South Africa will therefore need to be informed by transformational imperatives including:

- Ensuring the profession is accessible to all sectors of society;
- Ensuring the profession does not become the preserve of graduates from "elitist institutions";
- Recognition of Prior Learning;
- Achieving equity in the profession; and
- Providing equal opportunities for the advancement of all evaluators within the profession.

Building Human Capabilities to Advance the National Development Plan

The National Development Plan commits South Africa to building a capable developmental state. [11] The concepts of Autonomy and Embeddedness are critical determinants of the institutional character of a capable developmental state . Evans and Heller argue that the concept of autonomy refers to the ability of a developmental state to consistently pursue and achieve collective goals as opposed to the pursuit of individual or sectional interests. A key institutional requirement is strong coordination and synchronization of government's programmes and policies through a coordinating agency or super ministry. Embeddedness refers to the ability of the state's organizational structures to penetrate and interact with non-state actors – the ability of the state to elicit broad-based co-operation and promote its development goals.

The outcomes based monitoring and evaluation system in the Presidency seeks to transform the way in which government works and the way in which the state interacts with society more broadly. It creates the potential for participatory people-centred development to be achieved in practice. It requires competencies and skills that many working in the evaluation field do not have. Thomas Schwant has argued that:

To possess a life of the mind for evaluation practice is to recognise that professional practice is a matter of designing means for investigating and rendering a judgement of value that will be effective in light of contextual circumstances, political and ethical considerations, client expectations, and resource constraints. It is to realise that assumptions that evaluation practitioners and the public make about the purpose and roles of evaluation, how and why it is of service to society, the uses to which evaluation knowledge is put, and the evidence employed in arguments about value are almost always contested or contestable and must constantly be revisited in professional practice. A "life of mind for practice" is required bevause central to the nature of what we call professional work is the notion that it lacks uniformity in the problems it must contend with and the solutions it must invent." [12]

Schwandt's seminal book and its analysis of the professionalisation issue points to the hetero-

11 "Human Development, State Transformation and the Politics of the Developmental State," in The Oxford Handbook of Transformations of the State, S. Liebfried et. al., Oxford, (Forthcoming)

12 T. Schwandt, Evaluation foundations revisited: cultivating a life of the mind of the practice, Stanford Business books (2015).

geneity of evaluation practice and evaluators and their specialisations in terms of scale, unit of analysis, skill, competence and methodology. His work also evokes questions about the ostensible "objectivity" and "independence" of evaluation and evaluators. The establishment of Independent Offices of Evaluation in major multilateral organisations such as the World Bank and the UNDP implies that evaluation is ultimately evidence-based and objective. But the opening shot in Schwandt's book states: "Evaluation is the act of judging the value, merit, worth, or significance of things." [13] Value judgement is difficult to encapsulate within a qualification, skill, competence or experience. A recent Study by Gilens and Page in the US examined public-opinion polls on 1,779 policy issues between 1981 and 2002, and concluded:

…economic elites and organized groups representing business interests have substantial independent impacts on US government policy, while average citizens and mass-based interests have little or no independent influence. [14]

Ernest House argued in 2003 how George Bush's neo-fundamentalism influenced a methodological fundamentalism in evaluation focused on "evidence-based" methodological fundamentalism rooted in quantitative pseudo-scientific randomised trials, through which, "Government officials often yearn for certitude in evaluation findings as a way of bolstering their authority." [15] Notwithstanding the problems identified above with the discourses of objectivity, neutrality and independence, the manipulation of evaluation findings specifically and evidence more generally to bolster authority, provides a strong basis for the establishment of "Independent Offices of Evaluation."

These are no doubt critical considerations for South Africans as evaluation has become central to the transformative project in South African society. More recently the application of the Malaysian "Big Fast Results" methodology through Operation Phakisa has seen evaluation being used to develop policy and implementation plans for whole sectors of the economy such as the Ocean Economy and Mineral Value Chains. This may be a far cry from solving specific local challenges related to human settlements and sanitation, but either way, expanding human capabilities in the field of evaluation has become a key foundation for development and growth in South Africa.

Developing Evaluation Capacities

Building human evaluation capabilities is central to development and progress. It must be recognised that capacity development goes beyond an individual or an organisation. Evaluation principles need to be integrated with measures that go beyond the individual to span the institutional framework and the enabling environment for evaluation. [16] Currently the South African Government has made building the human capabilities for evaluation and monitoring a national priority. There is a huge opportunity for evaluation and evaluators to contrib-

13 Schwandt, op.cit., p.1.

14 M. Gilens and B. Page, "Testing Theories of American Politics: Elites, Interest Groups and Average Citizens, Perspectives on Politics (Forthcoming: 2014).

15 E. House, "Bush's Neo-Fundamentalism and the New Politics of Evaluation," Studies in Policy and Educational Philosophy, E-tidskrift (2003:2)

16 C. Heider, "A conceptual framework for developing evaluation capacities: Building on good practice," Chapter 5 in R.C. Rist, M. Boily, and F. Martin, (eds.), Influencing Change: Building Evaluation Capacity to Strengthen Governance, 2011: The World Bank

ute to socio-economic change at a variety of level and in various contexts, so the enabling environment is there. The institutional context requires further capacity and opportunities for Higher Education Institutions to develop the relevant programmes and for public, private and civil society Institutions and organisations to develop the institutional capacity for evaluation execution.

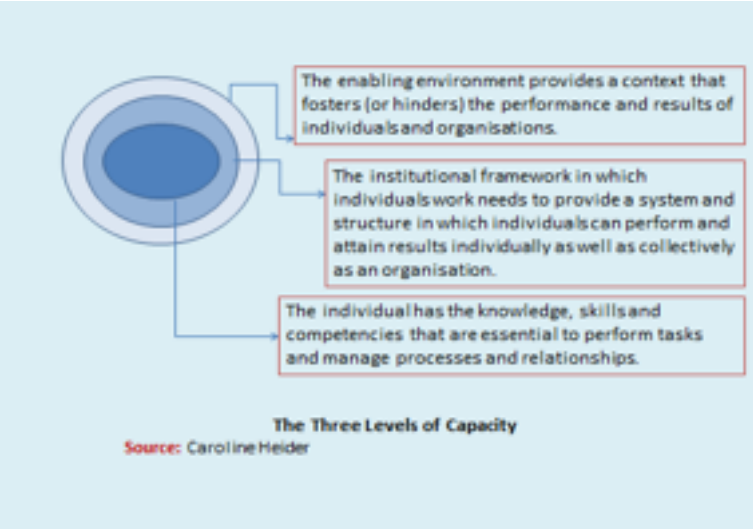


Figure 1: The Three Levels of Capacity

Effective and efficient educational development programme aimed at building evaluation capacities, should be developed based on a skills and competency profile for evaluators, as a developmental path benchmark. According to the Canadian Evaluation Society (CES - 2010) competencies are "the background, knowledge, skills and dispositions programme evaluators needs to achieve standards that constitute sound evaluations" . The CES goes further in arguing that 'competencies are not static'; skills and knowledge in any profession grow and evolve over time, influenced by new research and changing environments.

The context in which we embark on professionalisation of evaluation and evaluators in South Africa is one characterised by diversity of sectors and individuals as well as uneven geographical spread, which can be seen in sectoral and spatial representations of SAMEA's membership.

Type	Number
Academic/Research Institution	33
Foreign government	4
Government	145
International development organisation	14
NGO/Civil society	43
Non-affiliated students	13
Affilitated students	24
Private	125
Total	401

Figure 2: SAMEA Membership Profile

17 The Canadian Evaluation Society, Competencies for Canadian Evaluation Practice, (CES: 2010, p.2). The competencies are adapted from L. Stevahn, J. King, G. Ghere, J. Minnema, Establishing Essential Competenciesfor Program Evaluators, American Journal of Evaluation, Vo. 256, No. 1 (2005: 43-59).

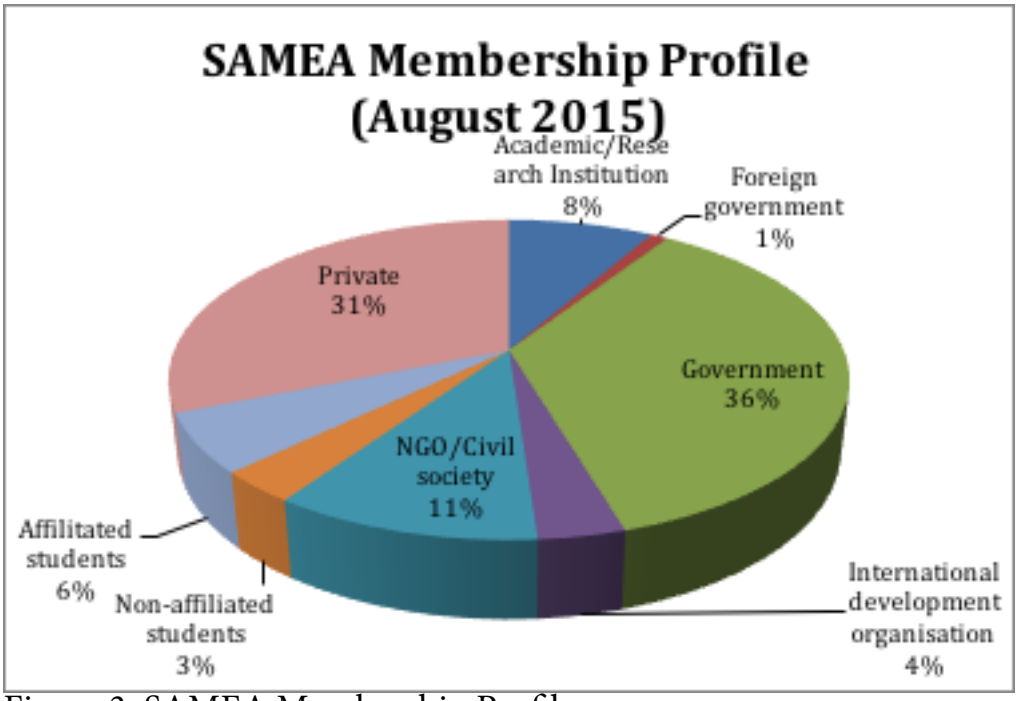


Figure 3: SAMEA Membership Profile

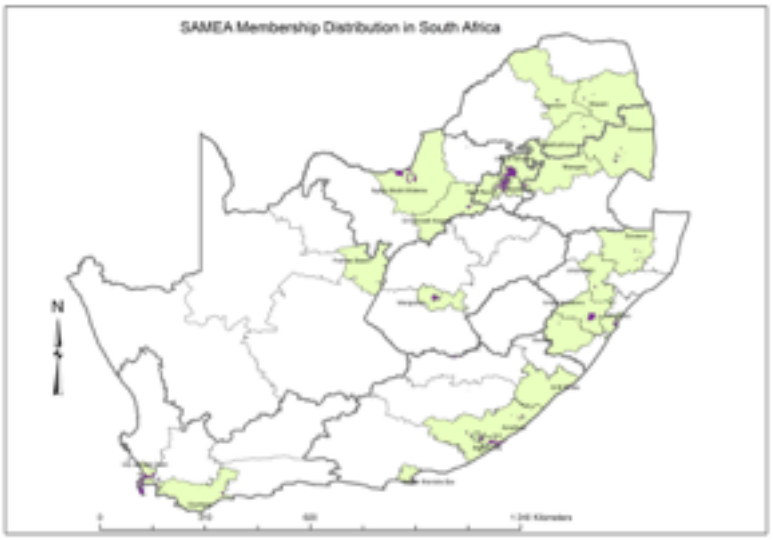


Figure 4: Spatial Distribution of SAMEA Membership

Professionalisation People-Centred Development Evaluation and Participatory Development

In keeping with the overarching themes and resolve of the International Year of Evaluation, the SAMEA Board elected to focus its 10th Conference on the theme: Using Evaluation to Improve People's Lives. Across the globe, hundreds of millions are marginalized, but the big question is: will these voices be heard? Are there alternatives to comprador parasitic accumulation from above in developing countries, characterized by people-driven accumulation from below? The answer to these sets of questions creates a unique space, where people-centred development evaluation can make its contribution. Through its interdisciplinary approach development evaluation has the potential, utilising participatory methodologies, to provide an integrated articulation "of an alternative "cultural and economic paradigm" of development to the "Western modelo civilizatorio." [18]

18 A. Escobar, Encountering Development: The Making and Unmaking of the Third World, Princeton and Oxford, (2012 Edition, Preface: xxiv).

The concept of 'development', which many of us use so glibly and with an unflinching certainty, has been intensely contested over the last two decades. Wolfgang Sachs has argued that "development is much more than just a socio-economic endeavor; it is a perception which models reality, a myth which comforts societies, and a fantasy which unleashes passions." [19] Interestingly enough, the critique of development and the post-development discourse argues that the concept of development manifest in post-World War II Truman doctrine presents the western world as idealized and hegemonic. This is because the trajectory and development destination of a developing society has over the past seventy years all too frequently been cast in the frame of an "ideal-typical" western country. [20]

"Underdeveloped" countries have been exhorted to aspire to this "ideal-type". During these decades of massive transformations of community, society, nation state and national identity variously characterized as modernization, underdevelopment, post-modernism, post-structuralism, globalization and post-globalization, the state through massive investment in infrastructure and social welfare was at times the epicentre of development and at times its antithesis. In the era of Margaret Thatcher and Ronald Reagan, the market was seen as a more efficient allocator of resources. The 'Washington Consensus' formed during this period, and its free market policy reforms ushered in neo-liberalism with its efforts to reshape the globe in the image and fantasy of market fundamentalism. The onset of the global financial crisis of 2007 and the great recession provided a rude awakening. The net effect of this was a re-legitimization of state intervention in the economy without necessarily negating the role of the market. In the meantime, developmental states across the globe inspired by the robust success of the east Asian tigers used the power of the state to reshape their economic growth and development trajectories and made significant inroads on the challenges of unemployment, poverty and social inequality.

This is the global context in which South Africans, confronted by the apartheid legacy of extreme social and income inequality grapple with the construction of a capable, people-centred developmental state. This context is confounded by the post-colonial conundrum of limited pathways of accumulation and the importance of state power, state procurement and contracting, including the outsourcing of evaluation by the state as critical levers of socio-economic transformation and empowerment. This is the context in which evaluation and evaluators in South Africa contend with the challenges of professionalisation of the discipline. Professionalisation can potentially protect the integrity of evaluation, but it can also lead to exclusion and the inadvertent reproduction of historical inequalities both of individuals and institutions.

19 W. Sachs, The Development Dictionary – a guide to knowledge as power, Zed Books, Second Edition (2010).

20 G. Esteva, S. Babones and P. Babcicky, The Future of Development: A radical manifesto, Policy Press, (2013: Chapter 2).

Conclusion

The issues presented provide some thoughts around the implications of professionalising evaluation in South Africa. We would need to look at how the profession or occupation would be specified, classified, accredited and qualified as well as the consequences this would have on existing practitioners and emerging communities of practice. We must protect the integrity of evaluation practice and evaluators recognising that the discipline itself is ubiquitous, highly differentiated across numerous axes and contested. Nevertheless, we should be cautious of over-regulation, which may prejudice the historically disadvantaged and reproduce old race and university hierarchies in new forms. Professionalisation can potentially bring benefits for both evaluation and evaluators. However, South Africa's historical context must be taken into account if we go this route. Furthermore, we need to understand the unintended consequences, which may arise out of a process of professionalisation, which ideally should be a transformative project. The primary risk lies in the real possibility of producing and reproducing new forms of elitism and exclusion.

SIMPLIFYING THE LOGIC OF LOGIC MODELS

Lauren Wildschut
CREST, Stellenbosch University October 2015

Introduction

Programme managers and evaluators have a few things in common: they are both concerned with appropriate intervention designs, improving interventions and the need to know if the programme is making a difference to the lives of beneficiaries. In their endeavours they are surrounded by a myriad of tools, particularly models which focus on the design and logic of interventions. These models which can be used for designing, planning, monitoring, reviewing or evaluating programmes can be broadly categorized as logic models. Some of these models are associated more closely with Theory-based evaluation (Fitz-Gibbon & Morris (1975); Friedman (2001); Weiss (1995, 1997a) as they attempt to represent the theory (causal effects) of an intervention while others are not, as they represent the logic (sequence of components) of an intervention. As the popularity of Theory-based evaluation has grown, so has the popularity of graphics that represent the theory or logic of interventions such as theory of change models, logic models and logframes. Despite the distinctive matrix of the logframe, it is regarded by some as a type of logic model (Funnell &Rogers, 2011: 395; Roduner, Schläppi & Egli, 2008:5). Funnel and Rogers (2011:391) claim that "a logframe is a pipeline model" which is used in international development. "Logic model" can therefore be viewed as an umbrella term for a host of models that are closely related in purpose and function as, regardless of the type of model, it would seem that all have some overlap in their use, i.e. description, strategic planning, monitoring and evaluation and learning.

Patton (2002:162) makes it clear that a logic model's only function is to show the connections between programme components - logic models must have a "reasonable, defensible and sequential order" of programme components. This is different to a theory of change, which "bears the burden of specifying and explaining assumed, hypothesized, or tested causal links". The logic model does not have to indicate what the mechanism for change for an intervention is, only the logical order of programme components. Astbury and Leeuw (2010:365), like Patton, blame the interchangeable use of terms such as logic model and program theory model for the current confusion in evaluation literature and practice, but indicate that "there appears to be growing recognition that they actually serve different functions" (Chen, 2005a; Leeuw, 2003; Rogers, 2007; Scheirer, 1987; Weiss, 1997a). It is not simply a case of looking at a graphic and deciding that it focuses on the logic or a theory of a programme, as it depends on how the model was developed, how it is used in the evaluation process and what other documents accompany the model. Just as the definitions of logic models contain a mix of a reference to programme components and theory, evaluation literature also contains visual representations which are labelled "logic models" but which may contain either programme components and/or a representation of the causal mechanism or theory of the programme.

As "models" they are more or less accurate reconstructions of the essential logic of interventions. As is the case with any model in science, there is usually not one true or correct representation of the underlying reality that it attempts to capture. In fact, constructivists would argue that all models are constructions of multiple realities (Hummelbrunner 2010). In addition, social interventions range from being very simple to complicated, to complex

(Rogers 2008). Consequently, models which represent a wide range of complexity of interventions usually display a similarly wide range of representations. If one adds to this the fact that the different organisations and agencies who have contributed to the range of available logic models - United Way (1996a); Centre for Disease Control (2004); Cyfernet (2000); Purdue University (2000); University of Wisconsin Extension (2008); W.K.

Kellogg Foundation (2004) - amongst others, also have their own interests and stakeholders in mind when developing their specific models, it is even more understandable that we end up with a wide array of seemingly disparate and incommensurable models. For programme managers and evaluators alike, this myriad of models can seem overwhelming and confusing. In a perplexing world of what seems to be countless types of models, there is in fact a typology of models which can help both programme managers and evaluators identify a model which works for their particular intervention and context. This article shares the findings of a systematic review of a sample of logic models (120) in order to bring some order and clarity to a diverse collection of models to a more manageable and comprehensible set of main and subsidiary types. This forms part of a larger study which included a typology for logframes but the purpose of this article is to focus on all models labelled as "logic models" and excludes models which utilize the conspicuous logframe matrix.
By having knowledge of and utilization of the typology, programme managers, planners and evaluators will be better able to choose a model which suits their context and need. The typology will also make the sometimes daunting world of logic modelling more accessible. The article follows the stages of a framework for typology development proposed by Kluge (2001) and concludes with a discussion of the theoretical, programme management and evaluation implications of the typology and directions for future research.

Development of typology

Typologies are constructed in order to understand and explain complex concepts and social realities. They are the result of a grouping process and the elements within a type have to be as similar as possible (internal heterogeneity at the "level of the type") and the differences between the types have to be as strong as possible (external heterogeneity at the "level of the typology" (Kluge 1999, pp.26). Typology development, as with other qualitative processes, is iterative in nature and Kluge (2001:6) provides a useful depiction of this iterative process which is shown in Figure 1 and used to structure this article.

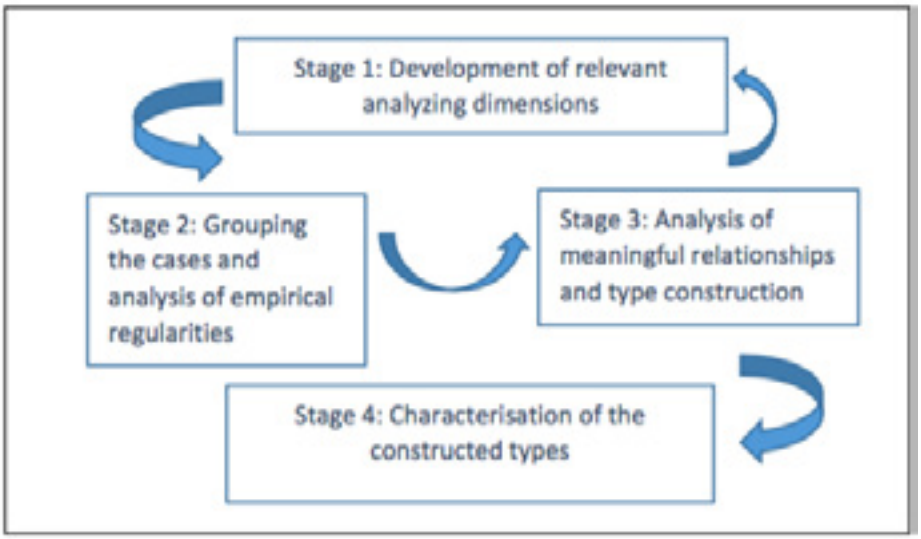


Figure 1: Stages in typology development (Kluge 2001:6)

Stage 1: Development of relevant analyzing dimensions

One of the key issues of developing a typology is having clear, set dimensions group types. When reviewing models to develop the typology under discussion, the interchangeable use of terms in articles and manuals made it difficult to decide what to include under the label "logic model". For example, a logic model is sometimes labelled as a programme theory, programme model, theory of change or action theory among other terms (Donaldson 2007, Funnel and Rogers 2011). This undifferentiated use of terms when describing the models associated with presenting of the logic of the implementation of interventions has led to a fair amount of confusion amongst donors, evaluators and programme staff (Astbury & Leeuw, 2010; Brousselle, Lamothe, Mercier & Perreault, 2007; Coryn, Noakes, Westine & Schröter, 2011; Davidson, 2000; Gasper, 2001; Hansen & Vedung, 2010; Weiss, 1997b). The function of the models i.e. whether they are used to describe implementation logic or theory is not used as a criteria in the typology, so if a graphic was labelled "logic model", "programme logic" or "programme theory" it was included in the typology of logic models. The difficulty with logic models depicting the logic of interventions is the variability of so many dimensions:

1. Shape or format – models can be flowcharts, tabular, narrative or have geometric shapes
2. Terminology –components can be called goals or overall objective and mean the same thing
3. Definitions of components – these vary from model to model e.g. outputs may in some cases
include activities whilst in others not
4. Number of components included – these vary from no labelled components at all to eight components
5. Display features – these can include arrows, connecting lines, blocks, feedback loops and colour
6. Direction of flow – this can be left to right, top to bottom, bottom to top or even circular
7. Level of detail – some models have very little detail while others have multiple components, strands and feedback loops (Wyatt Knowlton and Phillips 2013; Funnel and Rogers 2011).

The variation in the format of models is often due to the differences in intended purpose, the nature of the organisation for which the model is developed, and the level of complexity of the programme that is depicted.

Stage 2: Grouping the cases and analysis of empirical regularities

This stage in the development of a typology involves classification of the cases – in this instance the models - by means of the defined properties and dimensions and then identified groups can be analysed with regard to regularities. "Cases which are assigned to a combination of attributes must be compared with each other, in order to check the internal homogeneity of the constructed groups—which form the basis for the later types" (Kluge, 2001:5). The typology of logic models is based on the United Way model (1996a:vii) as the paradigm case. The United Way model is defined as the "paradigm case' as it is generally recognised as one of the earliest and most widely used logic models (Hendricks, Plantz, & Pritchard, 2008:13).

The model consists of four key components: inputs, activities, outputs and outcomes and captures the logic of interventions as a linear flow (with no feedback loops) from left to right. The model has been classified in the review as a flowchart due to the arrows between components

(contained in blocks).

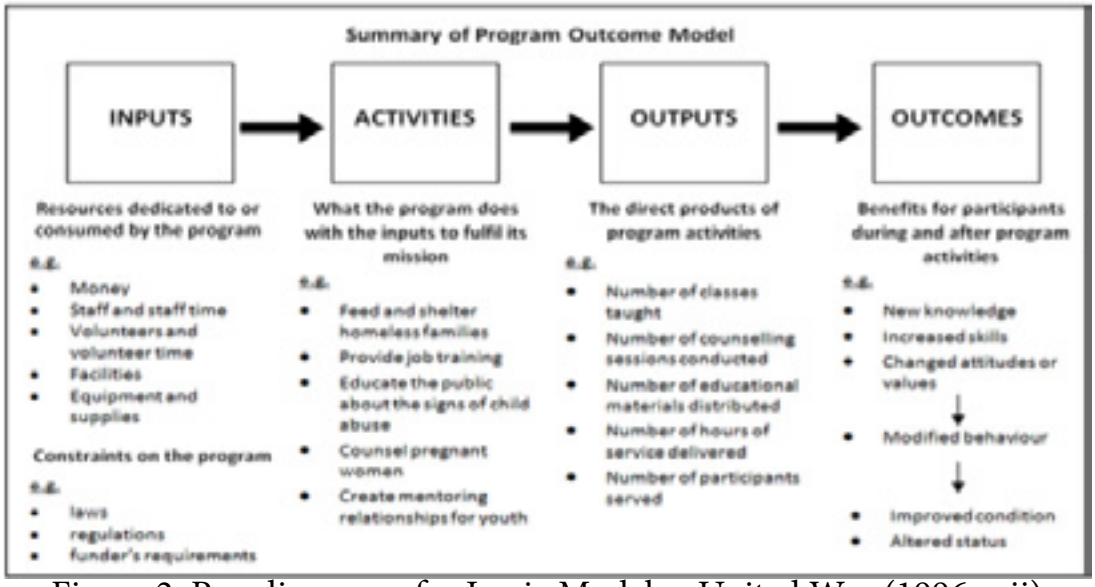


Figure 2: Paradigm case for Logic Models - United Way (1996a:vii)

In this early version of the logic model, the outcomes component has not been labelled in terms of time periods but the arrows from one outcome to another hint at a progression. It is interesting to note that this early version of the logic model already includes a "framing" of the model, as the left- hand corner has a section labelled "Constraints on the programme", which lies outside the structure of the model and describes external factors. The term "framing" used here is to refer to those components that lie "outside" the internal structure or logic of the intervention. These "framing components" situate the intervention in a specific context through reference to external factors, assumptions or conditions. These components are not extensions of the core logic of the intervention but are rather external to the intervention, enclosing or contextualising the intervention.

Using the United Way model (1996a:vii) as the standard or reference exemplar of a logic model, a typology of ten types of logic models was subsequently developed. The development process is described in the next section.

Stage 3: Analysis of meaningful relationships and type construction

The models utilized in the development of the typology were collected from 198 sources – 155 articles, 27 logic model manuals and 16 theory of change guides. The articles collected came from a range of sources – mostly evaluation journals, but also from discipline-specific articles from the health and development sector particularly. ATLAS.ti, a computer-aided qualitative data analysis software (CAQDAS) programme was used to compare and code ("applying a short-hand label to sort, syntheisize, and conceptualize data", Charmaz 1983: 110) the 120 collected models against the paradigm case. The coding approach used for the review of models was both inductive and deductive, as once a set of particular model components had emerged from the inductive coding of 20 models, deductive coding was used on the rest of the models.

Stage 4: Categorisation of the types

The three classification principles that underpin the logic model typology are

adherence - matching components of paradigm case;

addition of components through a) extension - adding to the causal chain and b) elaboration
- adding a dimension to one of the components;
linearity – the flow of the logic, whether the model was linear or recursive.
These classification principles are shown in Table 1 together with the symbols used in the typology to depict the nature of the model components and the flow of the model.

Principle	Criteria for categoristaion of type	Key
1. Adherence	The model **adheres** to the standard (as per paradigm case) number of four components.	
2. Addition	Change to the components constitute an **extension** to the core causal chain of components	
	A specific component is **elaborated** on	
3. Linearity	Intervention is represented in a **linear** (left to right) or **non-linear** format (recursive)	

Table 1: Criteria for grouping of logic models

Essentially there were 4 major groupings within the typology which are based on the number of paradigm components contained in the model and the nature of the additions. These are:
Standard models – with 4 components (matching United Way model)
Truncated models - 3 or fewer standard components
Hybrid models - 3 or fewer standard components with additional components
No labelled standard components – these are generally theories of change

Table 2 lists the ten types of models found in the typology. The table shows a simple grouping of models based on adherence, addition and linearity. Types 1 to 4 contain the standard United Way components but differ because of extension (Type 2), elaboration (Type 3) and extension and elaboration (Type 4). Types 5 to 8 all have 3 of the paradigm components (Type 5) but again differ based on extension (Type 6), elaboration and non-linearity (Type 7) and extension and elaboration (Type 8). Types 9 and 10 have no labelled components and differ on the basis of linearity.

Type	Category Name	Description	Graphical representation	Examples
1.	Standard linear	4 paradigm case components in a linear format		Carman, J.G. (2010: 265) Cooksy, L.J., Gill, P. and Kelly, P.A. (2001: 122) Cozzens, S.E. (1997: 86)
2.	Extended linear	4 paradigm case components with additional components which extend the causal chain in a linear format		Centre for Disease Control (2004:8) Cyfernet(2000:2) Department for Transport UK (2010:5)
3.	Elaborated linear	4 paradigm case components with additional components which provide more detail on case components in a linear format		Dyehouse M Bennett D (2009:188) Otto, A.K., Novelli, K. and Mohoran, P.S (2006: 282)
4	Extended AND elaborated linear	4 paradigm case components with additional components which provide more detail on case components and extend the causal chain in a linear format		Centre for Effective Services (2011: 16) Flex Monitoring Team (2006: 6) United Way (2008: 39)
5.	Truncated linear	3 or fewer paradigm case components with a linear format		Anderson, L.A. Gwaltney, M.K., Sundra, D.L., Brownson, R.C. (2006:9) Goodson, P., Pruitt, B.E., Suther, S., Wilson, K. and Rubi, F. (2006:265) University of Wisconsin Extension (2008:44)
6.	Extended hybrid linear	3 or fewer paradigm case components with additional components which extend the causal chain in a linear format		Adler M.A. (2002:207) Cooksy, L.J., Gill, P. and Kelly, P.A. (2001: 121) Golman, K.G. and Smelz, K.J. (2006: 10)
7.	Elaborated hybrid non linear	3 or fewer paradigm case components with additional components which provide more detail on case components in a non- linear format		Livingood, W.C., Winterbauer, N.L., McCaskill, O. and Wood, D. (2007:E4)
8.	Extended AND elaborated hybrid linear	3 or fewer paradigm case components with additional components which provide more detail on case components and extend the causal chain in a linear format		Alter C & Egan M (1997: 98) Den Heyer, M. (2001: 4) Humphreys, J. et al (2010 p9)
9.	No designated component linear	No paradigm case components have been identified/labelled and are presented in a linear format		Adams J. and Dickinson P (2010:424) Bickman, L (1996:113) Brousselle, A. and Champagne F, (2004: 300) Camasso M.J., Jagannathan R. and Walker C. (2004:47)
10.	No designated component non-linear	No paradigm case components have been identified/labelled and are presented in a non-linear format		Brousselle, A., Lamothe, L., Mercier, C. and Perreault, M (2007:101) Friedman, J. (2001:168)

Table 2: Logic Model Typology

The spread of the 120 models identified in this study across the ten types is shown in Figure 3.

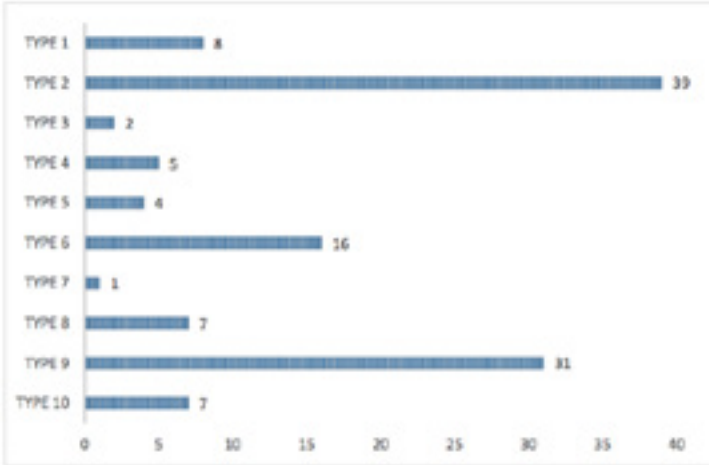


Figure 3: Number of models per Type (n=120)

Figure 3 shows that through the classification of 120 models it was found that the most frequent logic model types are:

Type 2 – Extended linear models which consist of 4 United Way core components with additional components which extend the causal chain in a linear format (n=39)

Type 9 - No designated component linear models which contain no labelled paradigm case components in a linear format (n=31)

Type 6 - Extended hybrid linear which have 3 or fewer paradigm case components with additional components which extend the causal chain in a linear format (n=16)

The review revealed that the first version of United Way model was very specific and components were fairly limited and so components were added to and disaggregated by evaluators and other agencies to make the model more user-friendly and a better representation of how interventions unfold over time. Extensions in the models occur mainly through disaggregation of outcomes into initial/immediate, intermediate and long term and the addition of a target group or participants. The second key cluster of modifications to the paradigm case began when the model began to be used

for purposes other than design of interventions. Inclusion of components such as indicators, sources of data, external factors suggest that the models were being used for monitoring and evaluation purpose as well.

The Second version of the United Way model (2008:26) incorporates both extension and elaboration and presents the most comprehensive version of a logic model. In this second version of the United Way model the concept of programme theory is also separated from the model and highlighted as an issue to be addressed. This can be seen as critical in a field where models often try and do both or claim to do one and actually do the other.

In the course of the review of the models it was also found that models are typically represented graphically in four formats: - flowcharts, tabular formats, shapes (such as triangles or circles) and narratives.

Although the four formats of representation are not always distinct (and sometimes presented in a hybrid format) the reviewed models were categorised as

Flowcharts – if the text of the model is separated into blocks (outlined or not) with arrows showing the direction of the causal chain flow

Tabular – if the text of the model is arranged in a table with columns and rows (with no arrows included)

Shapes – if the text of the model is contained in a geometric or graphic shape

Narrative – if the text is a set of statements linked by the phrase "if... then".

While this distinction was not utilized in the development of models types, the format was used as a discussion point for each of the types.

What follows is a brief discussion of the two most popular types of models, Types 2 and 9. The discussion describes the category overall (in a summary table) and then presents some of the variations within the category and examples of the models. Type 2 models (labelled Extended Linear) are described first.

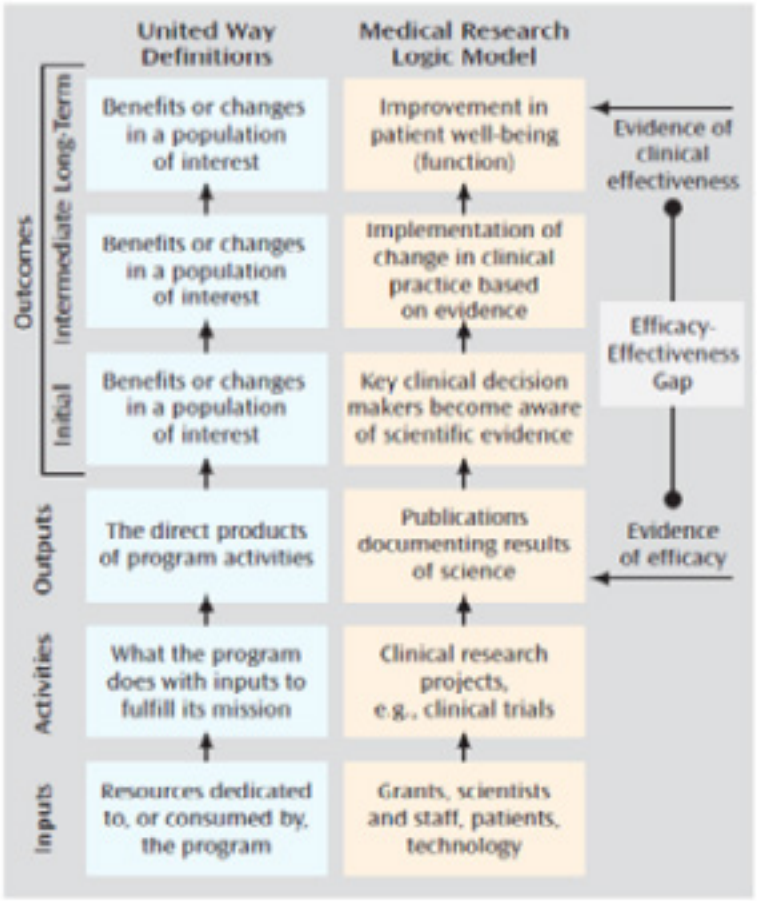
TYPE 2: EXTENDED LINEAR MODEL

Number of models	39
Number of flowcharts	29
Number of tables	10
Number of narratives	0
Sources	Articles (20) and manuals (18)
Main format	It contains two types of formats - flowcharts and tables but flow charts are more prevalent

Table 3: Overview of Type 2 category

The Extended Linear Model type still retains the fairly simple original format of the paradigm case but the teasing out of components in this category indicates that the developers of this group of models found the four core United Way components to be too limiting, and therefore unpacked components to include a greater level of detail.

Extensions are carried out in four ways (three which involve disaggregating the outcome component) and the addition of the problem/situation or target group. The most frequent way (11 models) of extending models is by disaggregating the outcomes – as shown in Figure 4 - followed by the disaggregation of outcomes and addition of situation and target group (8) as per the UWEX model. The last two ways of extending the causal chain occur by adding impact (5) and target group/participants (5).



[a] Adapted from United Way (12). Used by permission, United Way of America.

Figure 4: Type 2 _Weiss, A.P. (2007: 207)

The common principle that underpins the extension (whether through disaggregation of outcomes or addition of impact) is in fact the same, namely that outcomes take time to materialize. Outcomes and impact are time-bound, and more so with complex interventions. It takes time for the outcomes of an intervention to be visible. Some outcomes are dependent on prior outcomes being achieved. Type 2 models improves on the original Type 1 as it foregrounds the notion of change over time.

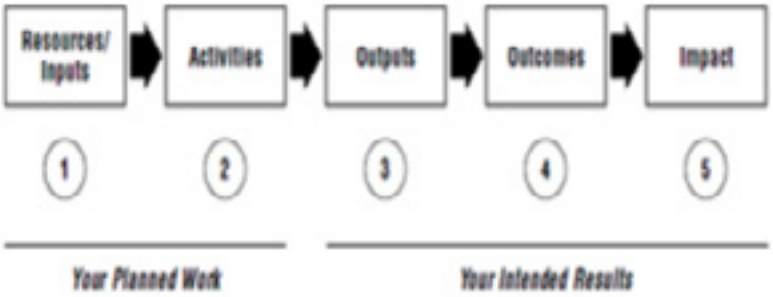


Figure 5: Type 2_W.K.Kellogg (2004:1)

In two other models, Goodstadt, M (2005:3) shown in Figure 6 and the Department for Transport United Kingdom (2010:5), the outcomes have been disaggregated and impact has been added. This format makes a distinction between time-bound outcomes and impact. Funnel & Rogers (2011:27- 30) describe the range of meanings of the term "impact" in eight organisations in five different countries (Australia, South Africa, Canada, India and the United States of America). From the descriptions that they provide, it is clear that in some instances impact is linked to time - as in "long term" (as in AusAid), or specifically within a time frame of "seven to ten years" (W.K. Kellogg Foundation) or simply " after program activities" (United Way). Thus it is critical when providing a graphic that the definitions of terms are provided. Without these definitions, models can easily be misinterpreted.

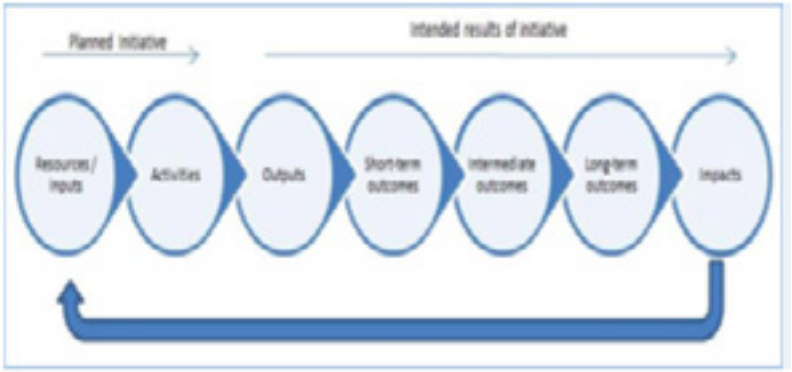


Figure 6: Type 2_Goodstadt M (2005:3)

In other exemplars not shown, Crane (2011:911) and Purdue University (2000:5) outcomes and impact are equated – that is the long-term outcome is shown as synonomous with impact. In Lindgern (2001:292) the model has outcomes which are disaggregated by level, i.e. for individuals and society rather than timeframes.

Another common type of extension was the disaggregation of outcomes and addition of target group or participants. An example of this is shown in Figure 7 (Medeiros et al. 2005:198).

Figure 7: Type 2_Medeiros et al. (2005:198)

The splitting of outputs into activities and participants is a key element of the logic models that were developed by the University of Wisconsin Extension (UWEX) in 2008.

There are eight flowcharts (out of twenty nine in this category) that show disaggregation of outcomes and addition of problem (situation) and target group. The structure of the model from UWEX is shown in Figure 8.

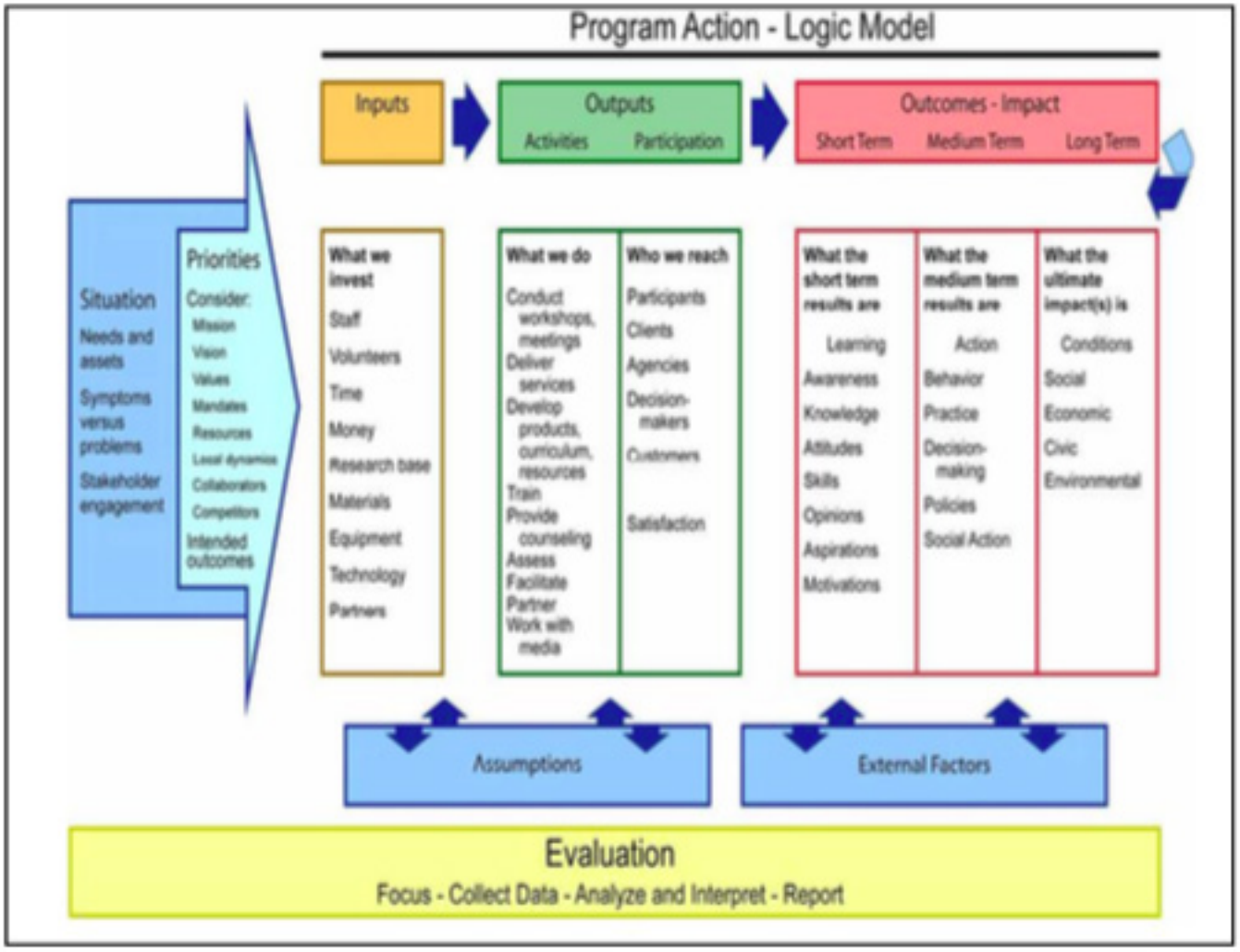


Figure 8: Type 2_University of Wisconsin Extension Manual (2008: Handout 14)

This model has the identifiable branding of the UWEX model:

the use of the term "Situation", which refers to the problem the intervention is addressing
the splitting of the Outputs component into Activities and Participation
assumptions and external factors framing the model
use of bright colours in the model

Two of the models in this category simply replicate the UWEX model – Medeiros et al. (2006:199) and UNESCO (2009:26) and acknowledge the original source, while others remove the colour branding and typical format and utilise the identical components - Arnold M (2006:261), Evaluation Support Manual (n.d.:3), Trinity College, Dublin (2008:8); University of Idaho (n.d.:1) and Medical Reserve Corps (2007:10). The impact of both the United Way and UWEX models can be seen in this grouping. Type 9 models which is the second largest grouping of models (after Type 2) and contains what is often termed as outcome chains is discussed next.

TYPE 9: NO DESIGNATED COMPONENT LINEAR

Number of models in this category	31
Number of flowcharts	31
Number of tables	0
Number of narratives	0
Sources	All articles except one manual (Aspen Institute)
Main format	Flowcharts

Table 4: Overview of Type 9 category

These models do not use any labelled components of the United Way case although these concepts may be included in the model. A few models use labelled components other than the core paradigm components and the general format of the models is to simply show the linkages between an intervention and the reactions to it.
Type 9 contains outcome chain models which are described by Funnell and Rogers (201:242) as showing "a sequence of results leading to the ultimate outcomes or impacts of interest". An outcome chain model (the format of which is shown in Figure 9) was first developed by Weiss (1972:50) and is either linear (classified as Type 9) or non-linear (classified as Type 10).

Figure 9: Type 9_Weiss, C. (1997a:504)

Coryn et al. (2008:201-202) when describing models as either "linear" or "non-linear" explain that the latter are intended "to integrate systems thinking in postulating program theory, taking contextual and other factors that sometimes influence and operate on program processes and outcomes into account" (Coryn et al. 2003:202). Virtually all models in Type 9 do not have labelled or identified paradigm components but one model by Tucker et al. (2006:2) labels some components of the model and includes a reference to the external environment. Figure 10, a logic model used by Adams & Dickinson (2010:424), uses a similar format to that of Weiss – thirteen years later. This is an indication of Weiss' influence on logic modelling which is still evident today.

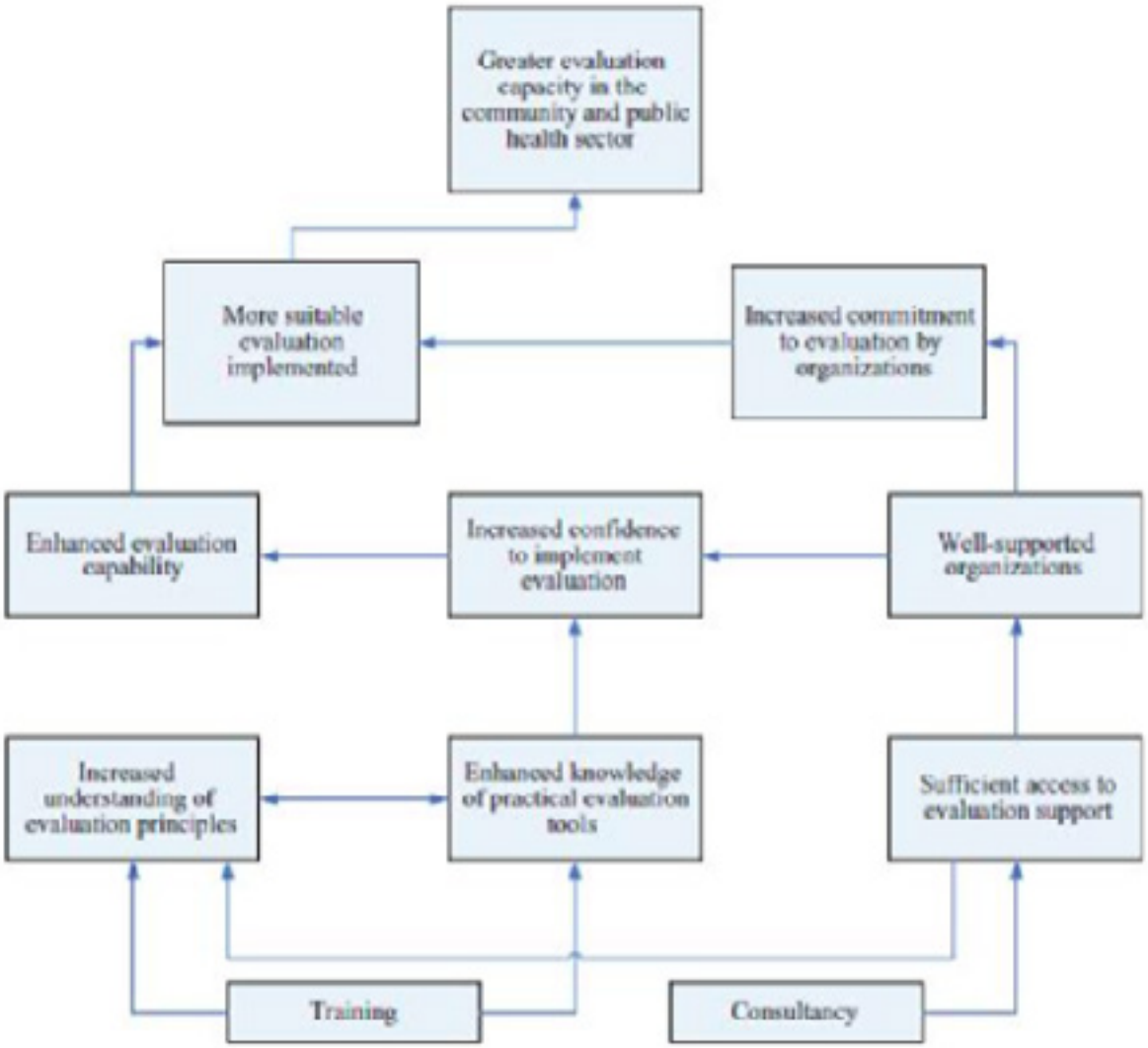


Figure 10 : Type 9_Adams J. and Dickinson P (2010:424)

The models in this category were called a variety of names. For example:

Logic model - Hawkins, et al. (2009:32); Tucker et al. (2006:2); Reed and Brown (2001:292)

Program theory - Weiss. (1997a:504); Donaldson and Gooler (2003:358); Mercier et al. (2000:2)

Program logic – Leeuw (2003:10)

Theory driven model – Turnbull (2002:277)

Chain of reasoning - Torvatn (1998:76)

Logic model map - Page (2009:79).

The names are not the only elements of these models that differ from one another. They also differ greatly in terms of level of detail, elements, structure and direction of flow. Table 5 shows some of the range of formats:

Very little detail	**Detailed**
Monroe *et al.* (2005:64) Camasso *et al.* (2004:47) Julian and Clapp (2000:206) Donaldson and Gooler (2003:358)	Leeuw (2003:10) Weiss (1997a:504) Hawkins *et al.* (2009:34) Page M (2009:79)
Inclusion of geometric shape	**Text and arrows**
Tucker *et al.* (2006:2) Mc Graw *et al.* 1996 in Weiss (1997b:50)	Duignan, P. (2004:9) Weiss, C. (1997a:504) Reed, C.S. and Brown, R.B. (2001:292)
Some labelling within models	**No labels within models**
Goodson *et al.* (2006:264) Tucker *et al.* (2006:2) - phases of projects Carvalho, S. and White, H (2004:145) - level of project	Aspen Institute (no date p 32) Adams and Dickinson (2010:424) Torvatn, H. (1998:79).

Table 5: Range of formats within Type 9 models

It is interesting to note that the two most popular types of models, Type 2 and 9 are those either with extended paradigm case components or very few labelled components (if labelled at all). Whilst Type 2 represent implementation logic, Type 9 are much closer to programme theory as they do not focus on the detailed components of an intervention, but far more on the causal mechanism which results in a ripple of outcomes.

Conclusion

There has been a huge growth in the use of models representing the logic underpinning interventions - whether this focus on the "implementation" or "program theory" (Weiss 1998:58) - for a range of reasons. Organisations of all types (governments, government development agencies, foundations and the corporate sector) have incorporated logic modelling into their planning, monitoring, evaluation and reporting demands. There has been a range of influences on the development of logic models over the years from forces in the academic, evaluation, government, donor and management fields. These influences have spawned a range of depictions that have attempted to represent both the logic of implementation and the theory of interventions. Some of these representations are more closely linked to Theory-based evaluation (logic models and theories of change) and others are not so clearly linked (logframes). This article has described a typology which will help programme planners, managers and evaluators as they develop models of interventions. The reduction in the types of models will assist with making choices for the use of a particular type of models and also provide for a clearer understanding of the multitude of models available. During the course of the study that underpins this article, it also became evident that the plethora of model formats and components cannot always be plausibly explained. An interest in promoting one's own "brand" may actually be the best explanation – at least of some of the variations that were identified. A comprehensive study of donor manuals, ethos' and philosophies could shed more light on this issue.

It is unlikely that any of the international donor organisations will give up their particular demands around logic modelling very easily but programme staff and evaluators, armed with

the understanding of the 10 key types of logic models, will be better equipped to argue for a more standardized approach to representing the logic of programmes.

Conclusion

There has been a huge growth in the use of models representing the logic underpinning interventions - whether this focus on the "implementation" or "program theory" (Weiss 1998:58) - for a range of reasons. Organisations of all types (governments, government development agencies, foundations and the corporate sector) have incorporated logic modelling into their planning, monitoring, evaluation and reporting demands. There has been a range of influences on the development of logic models over the years from forces in the academic, evaluation, government, donor and management fields. These influences have spawned a range of depictions that have attempted to represent both the logic of implementation and the theory of interventions. Some of these representations are more closely linked to Theory-based evaluation (logic models and theories of change) and others are not so clearly linked (logframes). This article has described a typology which will help programme planners, managers and evaluators as they develop models of interventions. The reduction in the types of models will assist with making choices for the use of a particular type of models and also provide for a clearer understanding of the multitude of models available. During the course of the study that underpins this article, it also became evident that the plethora of model formats and components cannot always be plausibly explained. An interest in promoting one's own "brand" may actually be the best explanation – at least of some of the variations that were identified. A comprehensive study of donor manuals, ethos' and philosophies could shed more light on this issue.

It is unlikely that any of the international donor organisations will give up their particular demands around logic modelling very easily but programme staff and evaluators, armed with the understanding of the 10 key types of logic models, will be better equipped to argue for a more standardized approach to representing the logic of programmes.

References

Adams, J. & Dickinson, P. (2010). Evaluation training to build capability in the community and public health. American Journal of Evaluation, 31:421-433.

Adler, M.A. (2002). The utility of modeling in evaluation planning: the case of the coordination of domestic violence services in Maryland. Evaluation and Program Planning, 25:203- 213.

Alter, C. & Egan, M. (1997). Logic modeling: a tool for teaching critical thinking in social work practice. Journal of Social Work Education, 33(1):85-102.

Anderson, L.A. Gwaltney, M.K., Sundra, D.L., Brownson, R.C., Kane, M., Cross, A.W., Mack, R., Schwartz, R., Sims, T. & White, C.R. (2006). Using concept mapping to develop a logic model for the prevention research centers program. Preventing Chronic Disease: Public Health Research, Practice and Policy, 3(1):1-9.

Arnold M.E. (2006). Developing evaluation capacity in extension 4-H field faculty: a framework for Success. American Journal of Evaluation, 27:257.

Aspen Institute. (n.d.). The community builder's approach to theory of change. Available from: http://www.dochas.ie/Shared/Files/4/TOC_fac_guide.pdf [Accessed 7 March 2013].

Astbury, B. & Leeuw, F.L. (2010). Unpacking black boxes: mechanisms and theory building in evaluation. American Journal of Evaluation, 31:363.

Bickman, L. (1996). The application of program theory to the evaluation of a managed mental health care system. Evaluation and Program Planning, 19(2):111-119.

Brousselle, A. & Champagne, F. (2004). How was the UNAIDS drug access initiative implemented in Chile? Evaluation and Program Planning, 27:295-308.

Brousselle, A., Lamothe, L., Mercier, C. & Perreault, M. (2007). Beyond the limitations of best practices: how logic analysis helped reinterpret dual diagnosis guidelines. Evaluation and Program Planning, 30:94-104.

Camasso, M.J., Jagannathan, R. & Walker, C. (2004). New Jersey's transitional housing demonstration program: the relationship of service delivery structure and process to the attainment of more permanent forms of housing. Evaluation and Program Planning, 27:45-58.

Carman J.G. (2010). The accountability movement: what's wrong with this theory of change? Non- profit and Voluntary Sector Quarterly, 39: 256-266.

Carvalho, S. & White, H. (2004). Theory-based evaluation: the case of social funds. American Journal of Evaluation, 25(2):141-160.

CDC, (Centre for Disease Control). (2003). A guide on logic model development for CDCs Prevention Research Centers. Available from: https://www.bja.gov/evaluation/guide/documents/cdc-logic-model- development.pdf [Accessed 22 October 2012].

CES, (Charities Evaluation Service). (2011). Building a work related logic model. Available from: http://www.effectiveservices.org/images/uploads/file/projects/P012/Workbook%20Two%2028_2_11%20v5.pdf [Accessed 19 March 2013].

Charmaz, K. (1983). The grounded theory method: an explication and interpretation. In Emerson, R.N. (ed.). Contemporary field research: a collection of readings. Boston, MA::little Brown: 109-126.

Chen, H.T. (2005a). Practical program evaluation: assessing and improving program planning, implementation, and effectiveness. Thousand Oaks, CA: Sage.

Cooksy, L.J., Gill, P. & Kelly, P.A. (2001). The program logic model as an integrative framework for a multimethod evaluation. Evaluation and Program Planning,

24:119-128.
Coryn, C.L.S., Noakes, L.A., Westine, C.D. & Schröter, D.C. (2011). A systematic review of theory- driven evaluation practice from 1990 to 2009. American Journal of Evaluation, 32(2):199-226.
Cozzens, S.E. (1997). The knowledge pool: measurement challenges in evaluating fundamental research programs. Evaluation and Program Planning, 20(1):77-89.
Crane, B. (2010). Using qualitative data to refine a logic model for the Cornell family development credential program. The Qualitative Report, 15(4):899-931, July.
Cyfernet. (2000). Using logic models to evaluation parenting skills programmes: a "How-to" guide. Available from: http://ag.arizona.edu/sfcs/cyfernet/cyfar/nowg_bul_pf_4.pdf [Accessed 11 October 2012].
Davidson, E.J. (2000). Ascertaining causality in theory-based evaluation. New Directions for Evaluation, 87:17-26.
Den Heyer, M. (2001). The temporal logic model concept paper evaluation unit IDRC. Available from: http://www.idrc.ca/en/ev-27877-201-1-DO_TOPIC.html [Accessed 13 July 2012].
DFID, (Department for International Development, UK). (2010). Logic mapping: hints and tips. Available from:
http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/38 17/logic-mapping.pdf [Accessed 15 November 2012].
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, 26:355-366.
Donaldson, S.I. (2007). Program theory-driven evaluation science: strategies and applications. Mahwah, NJ: Erlbaum.
Duignan, P. (2004). Intervention logic: how to build outcomes hierarchy diagrams using the OH Diagramming Approach. Available from:
http://www.strategicevaluation.info/se/documents/124pdff.html [Accessed 31 September 2012].
Dyehouse, M., Bennett, D., Harbor, J., Childress, A. & Dark, M. (2009). A comparison of linear and systems thinking approaches for program evaluation illustrated using the Indiana Interdisciplinary GK-12. Evaluation and Program Planning, 32:187-196.
ESS, (Evaluation Support Scotland). (n.d.). Developing a logic model. Available from: http://www.evaluationsupportscotland.org.uk/media/uploads/resources/supportgui de1.2logicmodelsjul09.pdf [Accessed 7 March 2013].
Fitz-Gibbon, C. & Morris, L. (1975). Theory-based evaluation. Evaluation Comment. The Journal of Educational Evaluation, 5(10):1-4. Reprinted in 1996.
Flex. (2006). Creating program logic models – a toolkit for state Flex programs. Available from: http://www.flexmonitoring.org/documents/PLMToolkit.pdf [Accessed 18 April 2013].
Friedman, J. (2001). Designed blindness: an action science perspective on program theory evaluation. American Journal of Evaluation, 22:161- 181.
Funnell, S.C. & Rogers, P.J. (2011).
and logic models.
Purposeful program theory: effective use of theories of change
San Francisco, CA: Jossey-Bass.
Gasper, D. (2001). Logical frameworks: problems and potentials. Unpublished paper of Institute of Social Studies, The Hague, Netherlands. [Online].
Available from: http://www.ecode.es/marco_logico/pdf/10.pdf [Accessed 26 April 2013].

Golman, K.G. & Smalz, K.J. (2006). Logic models: the picture worth ten thousand words. Health Promotion Practice, 7:8-12.
Goodson, P., Pruitt, B.E., Suther, S., Wilson, K. & Buhi, E. (2006). Abstinence education theory based? The underlying logic of abstinence education programs in Texas. Health Education & Behavior, 33(2):252-271, April.
Goodstadt, M. (2005). The use of logic models in health promotion practice. Available from: http://www.course-readings-and-resources.bestpractices-healthpromotion.com/attachments/File/Goodstadt%20Logic%20Model%20Paper/G oodstadt_Introduction_to_logic_models_paper.pdf [Accessed 18 April 2013].
Hansen, M.B. & Vedung, E. (2010). Theory-based stakeholder evaluation. American Journal of Evaluation, 31(3):295-313.
Hawkins, S.R., Clinton-Sherrod, A.M., Irvin, N., Hart, L. & Russell, S.J. (2009). Logic models as a tool for sexual violence prevention program development. Health Promotion Practice, 10:29-37.
Hendricks, M., Plantz, M.C., & Pritchard, K.J. (2008). Measuring outcomes of United Way-funded programs: expectations and reality. In Carman, J.G. & Fredericks, K.A. (eds.). Nonprofits and evaluation. New Directions for Evaluation, 119:13-35.
Hummelbrunner, R 2010a, 'Beyond logframe: critique, variations and alternatives', in N Fujita (ed.), Beyond logframe: using systems concepts in evaluation, Foundation for Advanced Studies on International Development, Tokyo.
Humphreys, D.K., Eisner, M.P. (2010). Evaluating a natural experiment in alcohol policy - The Licensing Act (2003) and the requirement for attention to implementation. Criminology & Public Policy, 9(1): 41-67.
Julian, D.A. & Clapp, J. (2000). Planning, investment and evaluation procedures to support coordination and outcomes based funding in a local United Way system. Evaluation and Program Planning, 23:231-240.
Kluge, S. (2000).Empirically Grounded Construction of Types and Typologies in Qualitative Social Research. Forum Qualitative Research Vol 1. No 1.
Leeuw, F.L. (2003). Reconstructing program theories: methods available and problems to be solved. American Journal of Evaluation, 24(1):5-20.
Lindgern, L. (2001). The non-profit sector meets the performance-management movement. Evaluation, 7(3):285–303.
Livingood, W.C., Winterbauer, N.L., McCaskill, Q. & Wood, D. (2007). Evaluating medical home constructs for children with special needs: integrating theory and logic models. Fam Community Health, 30(5):E1-15.
Medical Reserve Corps. (2007). Training guide. Available from: http://www.medicalreserve-corps.gov/File/MRC_Resources/Evaluation_Toolkit_Final _Logic_Model_7-5%281%29.pdf [Accessed 15 November 2012].
Mercier, C., Piat, M., Peladeau, N. & Deganais C. (2000). An application of theory-driven evaluation to a drop-in youth center. Evaluation Review, 24:73-91.
Monroe, M.C., Fleming, M.L., Bowman, R.A., Zimmer, J.F., Marcinkowski, T., Washburn, J. & Mitchell, N.J. (2005). Evaluators as educators: articulating program theory and building evaluation capacity. New Directions for Evaluation, 108:57-71, Winter.
Page, M., Parker, S.H. & Renger, R. (2009). How using a logic model refined our program to ensure success. Health Promotion Practice, 10:76-82.
Patton, M.Q. (2002). Qualitative research and evaluation methods. 3rd ed. London: Sage. Reed,

C.S. & Brown, R.B. (2001). Outcome-asset impact model: linking outcomes and assets. Evaluation and Program Planning, 24:287-295.
[Online]. Available from: http://www.agridea- international.ch/fileadmin/10_International/PDF/Concepts_and_approaches/OM_LF A_Synthesis_2008.pdf [Accessed 26 September 2013].
Rogers, P.J. (2007). Theory- based evaluation: reflections ten years on. New Directions for Evaluation, 114:63- 81.
Scheirer, M.A. (1987). Program theory and implementation theory: implications for evaluators. New Directions for Program Evaluation, 53:59-76.
Torvatn, H. (1998). 'Evaluation of Industrial Development in Norway'. New Directions for Evaluation, 77: 71-87.
Trinity College, Dublin. (2008). Public health nutrition intervention management. Available from: http://medicine.tcd.ie/nutrition-dietetics/assets/pdf/1-Intelligence-module/Unit-7- Intervention-Research-090128.pdf [Accessed 18 April 2013].
Tucker, P., Liao, Y., Giles, W.H., & Liburd, L. (2006). The REACH 2010 logic model: an illustration of expected performance. Preventing Chronic Disease, 3(1):1-6.
Turnbull, B. (2002). Program theory building: a strategy for deriving cumulative evaluation knowledge. American Journal of Evaluation, 23(3):275-290.
UNESCO, (United Nations Educational Scientific and Cultural Organization). (2009). A guide for evaluating community based projects. Available from:
http://unesdoc.unesco.org/images/0018/001862/186231e.pdf [Accessed 19 March 2013].
United Way of America. (1996a). Focusing on program outcomes: a summary guide. . Available from: http://uwnea.org/wp-content/uploads/2012/07/Focusing-on-Program- Outcomes.pdf [Accessed 7 March 2013].
Roduner, D., Schläppi, W. & Egli, W. (2008). Logical framework approach and outcome mapping: a
constructive attempt of synthesis. A discussion paper, ETH, Zurich, Switzerland.
United Way of America. (1996b). Measuring program outcomes: a practical approach. Alexandria,VA: United Way of America.
United Way of America. (2008). Logic model handbook. Available from: http://vsuw.org/file/logic_model_handbook_updated_2008.pdf [Accessed 12 September 2012].
University of Idaho. (n.d.). The logic model for program planning and evaluation. Available from: http://www.uiweb.uidaho.edu/extension/LogicModel.pdf [Accessed 18 April 2013].
University of Purdue. (2000). Utilizing the logic model for programme design and evaluation. Available from:
http://www.humanserviceresearch.com/youthlifeskillsevaluation/LogicModel.pdf [Accessed 7 April 2013].
University of Wisconsin (2008). Developing a logic model: teaching and training guide. Available from: http://www.uwex.edu/ces/pdande/evaluation/pdf/lmguidecomplete.pdf [Accessed 22 October 2012].
Weiss, A.P. (2007). Measuring the impact of medical research: moving from outputs to outcomes. American Journal of Psychiatry, 164:206-214.
Weiss, C.H. (1972). Evaluation research: methods for assessing program effectiveness. Englewood Cliffs, NJ: Prentice Hall.
Weiss, C.H. (1995). Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In Connell, J.P, Kubisch,

A.C., Schorr, L.B. & Weiss, C.H. (eds.). New approaches to evaluating community initiatives: concepts, methods, and context. Washington, DC: The Aspen: 65-92.
Weiss, C.H. (1997a). How can theory-based evaluation make greater headway? Evaluation Review, 21:501.
WKKF, (W.K. Kellogg Foundation). (2004). Logic model development guide. Available from: http://cpmcnet.columbia.edu/dept/pi/ppf/LogicModel.pdf [Accessed 7 April 2013].

The impact of a remedial reading programme on second language grade 4 students in KZN: Evidence from a randomised impact evaluation.

First author
Stephen Taylor

Second author
Brahm Fleisch

Third author
Volker Schöer

Forth author
Volker Schöer

Abstract

The majority of South African children do not speak English as their first language yet are taught in English from Grade 4 onwards. This represents one of the various educational disadvantages that are contributing to the low levels of learning observed amongst the majority of poor children in South Africa. Finding ways to reduce the learning deficits amongst these children is therefore an important policy priority.

This paper reports on a randomised controlled trial of a remedial programme designed to boost the English reading and literacy skills of grade 4 students, for whom English is a First Additional Language. The study randomly assigned 100 initially low-performing public schools in the Pinetown district of KwaZulu-Natal to treatment and control groups. The intervention lasted for 11 weeks, was administered within normal school time and consisted of three components: the provision of scripted lesson plans, additional reading resources and on-site instructional coaching for teachers.

The intervention had no statistically significant impact on the overall reading achievement of learners. However, treatment schools improved more than control schools in the spelling and grammar subcomponents of the test. The programme impact was larger for learners who initially had a basic minimum of English skills and for those whose teachers participated actively in the programme. The paper describes some of the challenges involved in implementing a randomised controlled trial in the context of the South African school system. The paper also reflects on how this sort of impact evaluation presents a challenge to the conventional research in education policy, but also creates valuable opportunities for inter-disciplinary collaboration to take knowledge further.

Funding for this study was provided by the Zenex Foundation

Acknowledgements

We would like to acknowledge the financial support of the Zenex Foundation, without which this study would not have been possible. That said, the authors take full and sole responsibility for the views and content of this report. There are a number of other organisations and individuals that need to be acknowledge and thanked for their ongoing support and assistance. We would like to thank the Wits Education Ethics Committee (non-medical) for its scrutiny and flexibility. To Mary Metcalfe, a big thanks for facilitating our discussions with the Kwazulu-Natal Department of Education. We need to thank the top management of that organisation for having the foresight to allow the study to take place. We owe a real debt of gratitude to the Pinetown district staff and above all to the teachers both in the control and treatment schools for their ongoing commitment. Finally to our partners at Class Act and the JET Education Services, thank you for your sterling work.

1.Introduction

Over the past decade there has been a growing recognition that a substantial portion of South African schoolchildren are one or more years below the acceptable achievement levels, particularly in key subjects like English First Additional Language and Mathematics (Taylor, 2014, NEEDU, 2014, Spaull and Kotze, 2015). Spaull and Kotze (2015) makes the compelling case that schoolchildren that are academically behind the acceptable levels of performance in the Foundation Phase, are likely to fall further and further behind their counterparts as they progress up the school system. This is clearly not a conventional 'remedial' problem, i.e. a small number of individuals within a class that have specific learning barriers or challenges, but rather the learning deficits are systemic, often effecting almost all learners across the majority of disadvantaged schools.

How can education departments address these systemic learning backlogs? There are a growing number of specialized programmes, particularly at the Grade 12 levels that focus on providing additional instruction. Although the systemic achievement gap often begins at the Foundation Phase, fewer programmes have been developed specifically to address the systemic problem early in learners' school careers. One exception is the Intermediate Phase Catch-Up Programme that was developed as a component of the Gauteng Primary Literacy and Mathematics Strategy in 2012. The eleven week programme that focused on re-teaching Foundation Phase English First Additional Language skills and content to learners in underachieving primary schools was designed to replace the curriculum for a single term to ensure that learners in these schools had an opportunity to master basic of English language literacy. Hellman's (2012) interval evaluation suggested that the Intermediate Phase Catch-Up programme was effective at-scale in helping the majority of learners in Grades 4 to 6 to gain basic literacy proficiency. But while the results were clearly encouraging, the design of the internal evaluation was not rigorous. The impact evaluation was administered by the service provider that designed the intervention, the pre and post-instruments were administered by the teachers themselves, and the study did not have an estimate of a counterfactual.

Given the importance in the education space of systemic catch-up programmes and the need for robust evidence of their effectiveness, a research team designed a robust impact evaluation of the Gauteng Intermediate Catch-up Programme. Although evidence generated from small scale studies (such as Pretorius, 2014) has the potential to contribute to the knowledge base, there is a clear need for studies with a sufficient sample size and with plausible method for identifying the causal impact to allow policy makers and researchers to establish with greater certainty the efficacy of education initiatives and/or specific programme interventions.

The impact evaluation of what came to be called the Reading Catch-Up Programme (RCUP) had a number of design features to ensure robustness. The research team, which designed the study and analyses and reports on the findings, separated the study into a learner data collection component and an implementation component. Class Act, the agency originally involved in the development of the intervention was tasked with implementing the intervention in treatment schools. JET Education Services was responsible for collecting learner information from pre- and post-tests in both treatment and control schools. The intervention took place in April to June 2014 in Pinetown, Kwazulu-Natal.

This paper is structured into four sections. Following this brief introduction, the paper provides a detailed description of the study method focusing on a description of the intervention, the Randomised Control Trial (RCT) methodology, the rationale for the selection of the study site, and the data collection processes. The third section presents the major findings including both information from a qualitative case-study undertaken during the intervention and the results of the pre and post-testing. While the focus here is on the main findings of the impact evaluation, this section also provides insights about other related findings. The discussion section explores explanations for the main finding. The final section considers the implications of the study.

2. Research Design and Methods

2.1 Background to the Reading Catch-Up Programme

In 2011 the Gauteng Primary Language and Mathematics Strategy developed and implemented an Intermediate Catch-Up Programme to remediate the learning gaps in underperforming Grades 4 to 6 classrooms. The Catch-Up Programme contains three key elements, i.e. scripted lesson plans, provision of high quality learning materials, and on-site coaching. The scripted lesson plans divided the term into 11 weeks, with each week designated with a number, e.g. Week 8, and each numerical week was linked to a particular calendar week, e.g. Week 8 Monday 5 March 2012-Friday 9 March 2012. Each calendar week for assessment was specified. These seemingly simple weekly plans signalled to teachers that they would need to keep up and that work assigned for the specific work week would have to be completed by the end of the calendar week so as to ensure that the learners were prepared for the assessment on the specific designated dates.

The original programme used six different learning resources for the classroom. The first was the printed A4 black and white lesson plan guide itself. The second was two A4 learner exercise books for each learner, one to write in during the regular class time and a second specifically for tests. The guide prescribed that the class exercise book is to be sent home every day, and the test book only to be sent home at the end of the term. The four listening and teaching posters provided to each class cover four themes: In the Classroom, At the Zoo, On the Beach, and At the Hospital. The key learning resource provided to all Intersen classrooms (Grades 4 to 7) was a set of 'reading' books, what could best be referred to as graded class readers. The guide lists the book title and the week that they are to be used in. The selected titles were listed on the Gauteng Department of Education approved book list as Grade 2 and 3 books for English home language learners. The use of Foundation Phase readers for Grade 6 and 7 learners was based on research (PIRLS, SACMEQ and ANA) that suggested that most learners in disadvantaged schools are three or more years behind the appropriate grade level in reading in English.

In addition to the A4 exercise books and the 240 reading books, the teachers received a set of 'reading sheets', sufficient for one set per learner. The reading sheets contained 'look and say' words that learners were expected to know the meaning of and commit to memory for the formal assessment. The 'look and say' words were derived from the reading books and constitute the core vocabulary and spelling words for the programme. The 'look and say' technique however did not dominate the Catch-Up programme's systematic reading approach, but formed one of three distinct interconnected components along with a phonics

programme and the graded class readers, what teachers called the 'thin books'. The last learning and teaching resource was a mark book, what the programmed called 'the Assessment Record Book'. The designers of the Catch-Up programme prescribed a strict and consistent weekly teaching routine to be followed in the same sequence every week. The teaching week was divided into seven half hour teaching periods. The teaching and the homework for each period was specified. Every week was to begin with a 'listening and speaking' task during which teachers teach ten sentences using the posters. The second period was for phonics and spelling; two new sounds and related words as well as specific high frequency words were introduced. Period 3 was devoted to teaching the 'look and say' words that would appear in the class reader for that week. During fourth period, the teachers were expected to begin using the class reader assigned for the week. The tasks for the period included reading aloud, shared reading and an oral comprehension exercise around the class reader. Period 5 was used for consolidation, the sixth period for reading and writing. The final period of the week had two main activities, writing and assessment. The assessment took the same form every week, a spelling test and a comprehension task. For each period, the guide specified the required homework. Save for the week during which there was to be formal assessment, each week would follow exactly the same format as the teacher worked systematically through the twelve graded class readers, the four posters, and twelve 'look and say' word sheets.

The daily lesson plan guide provided a comprehensive description of each of the 70 lesson periods. A typical daily lesson plan began with a heading which specified the week number, day of the week and the date. The lesson time (number of minutes), lesson outcomes and lesson resources were all shown at the top of the page. The 30 minute lessons have either one or two activities. The bulk of the daily plans consist of descriptions of these activities. The activities provide fairly detailed tasks per activity. The lesson plan specified the questions that teachers must write on the chalkboard and provided the answers (but tells the teachers not to write these on the board.) The ten questions on the graded class reader vary. Some were simple recall questions from the text (e.g. name the fruits that they use to make the fruit salad?); others required the learner's own response (which is your favourite?); a few required slightly higher order engagement (why do they add sugar over the fruit salad?).
The scripted lesson plans and the high quality learning and teaching resources, are regarded as a necessary but not sufficient condition for instructional change at scale in this model. The other component, the "just-in-time" training at the start of the programme and the ongoing in-class coaching is viewed by the programme designers as pivotal in shifting habits and routines of daily teaching practice. The deployment of instructional coaches was an essential ingredient. The coaches played a number of roles in the programme. They provided training to teachers in small groups, they visited classrooms to model teaching practice, to observe, support and encourage teachers as they work on the lesson plans and they monitored and tracked compliance. In the original programme, all coaches were themselves trained in the use of prescriptive protocols for coaching practice.

2.2 The Theory of Change

How do whole-class remedial programmes consisting of the scripted lesson plans, prescribed learner resources, just-in-time training and in-class coaching change instructional practices and improve learning outcomes? The theory of change embedded in the intervention assumes that these types of interventions, when they are tightly aligned, act to disrupt and

re-engineer three core elements of practice. First the lesson plans and the coaching change how time is understood and used. The first page of the lesson plan guidelines clearly link particular lessons to specific calendar days, thus specifying the pace at which the learning programme is to unfold. The pace remains the same even if teachers are absent or the day is interrupted for any reasons. The responsibility or burden shifts to the teacher to keep up with the pre-specified timeframes. Within the lesson, teachers need to increase their stamina to keep pace with the relentless forward motion of the lesson plans. The role of the coaches is to assist teachers, and once trust is established, to push them harder to remain on track and to keep up. What the new use of time does is to increase both the amount of time on learning tasks and intensify work on the tasks, thus allows for increased opportunities-to-learn and curriculum coverage. The prescribed weekly lesson routine provided a defined structure to school and lesson time. It is the routine and rhythms of that structure that would allow teachers to cope with the increased pace.

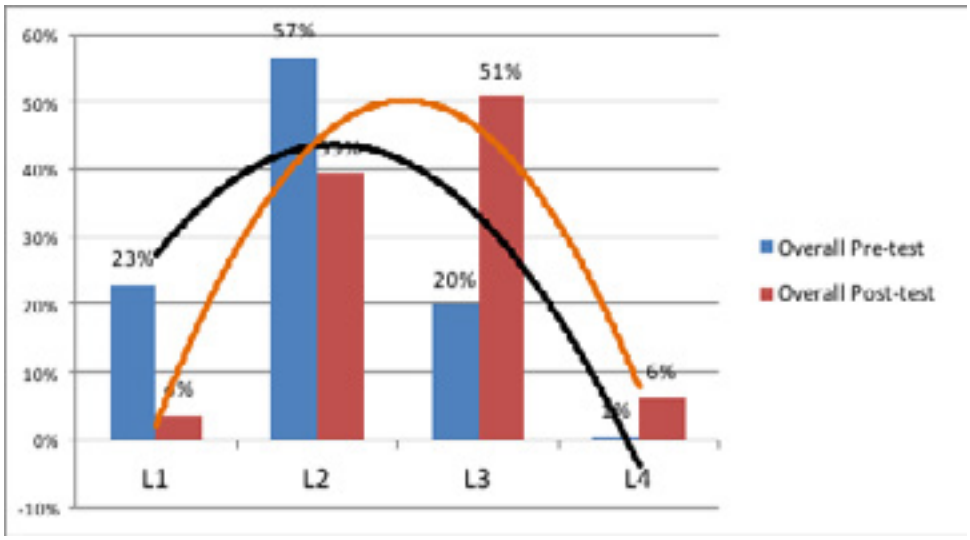
Second, the lesson plans and the learning resources, complemented by the work of the coaches, expand the teachers' pedagogic techniques and classroom management repertoire. One of the consistent findings in the literature (Fleisch 2008; Carnoy 2012; Taylor 2012) is the narrow range of activities and tasks teachers tend to use. The Catch-Up Programme lesson plans mandate a range of instructional methods and techniques. These included vocabulary development using the wall chart, graded reading using self-contained single theme readers, systematic phonics, 'look and say' words lists, and writing and comprehension strategies. While teachers may have made use of some or even all of the methods or techniques at one time or another, the lesson plans provide a systematic and integrated framework within which each method or technique is deployed sequentially and developmentally over time through the carefully structured framework. Not only did teachers experience how the learning tasks embedded in each lesson built on each other, but how the various methods and techniques, e.g. phonics and class reading, reinforced the learning pathway. The lesson plans also provide tangible instruction on the organisation of time, resources and classroom management.
The third way it disrupts and re-engineers practice is that it links instruction more directly to the reading levels of most of the learners in the class. An emerging finding in international literature on large-scale reform is the negative consequences of the overambitious curriculum (Pritchett, 2012). By beginning with the average actual reading levels of learners and moving them systematically along, the intervention ensures that a large proportion of learners will be able to benefit from reading instruction and reading materials at the appropriate grade level by the end of the intervention.

2.3 Results of the 2012 Catch-Up Programme Pre- and Post-Test Study

A preliminary "pre and post" test evaluation suggested that the programme is effective (Hellman, 2012). That study was conducted internally by those responsible for the programme as it was administered in the Gauteng province. It focused on learner performance, assessing the extent to which the Catch-Up Programme improved four distinct literacy skills, i.e. spelling, language, comprehension and writing. Two assessment tools were developed, one for learners in Grades 4 and 5, and a second for learners in Grades 6 and 7. The final non-randomly selected sample consisted of 1570 classes, which was about 45% of English teachers covered by the programme. Hellman (2012) found that while not all learners were on the same level of achievement at the start of the intervention, across skills, NGOs and districts,

the magnitude of gains made by learners was roughly of the same order. Overall, the programme's striking characteristic is that irrespective of grade, NGO and district, it seemed to have had a strong, positive and consistent effect.
Figure 1 is taken from Hellman (2012) and demonstrates the test score gains made by the children exposed to the Catch Up programme over the period of the programme. On the pre-test only about 21% of children scored above 50%, whereas after the programme about 57% of children scored higher than 50% on the test. The study had no control group with which to estimate what learning gains would have been in the absence of the programme.

Figure 1 Catch-up programme distribution across the four levels, 2012 (Source: Hellman (2012))



2.4 Experimental Design

The core question that animated this study involved the efficacy and cost-effectiveness of the Catch-Up Programme in improving learner performance on four components of reading. At an educational theory level, the study has the potential to contribute to an understanding of the effectiveness of combining scripted lesson plans, high quality materials and instructional coaching.

Until recently, RCT studies were uncommon in developing country contexts. While the findings of these randomised experiments are clearly important, given the high-stakes consequences of their findings, it is necessary to expand the number of studies using these approaches and compare findings. One of the problems with some of the existing South African studies is that the evaluations have often been undertaken by the programme developers, potentially compromising the independence of the investigations.
2.5 Sampling frame and rationale

The Pinetown District of KwaZulu-Natal province was the research site for the study. It has the advantage of containing a range of poor schools of different types (rural, urban, informal settlements and formal settlements). The study was conducted amongst grade 4 children in this district in schools where the dominant home language was not English. In the majority of cases the home language of children was isiZulu.
A detailed report on the sampling procedure is available online in a Pre-Analysis Plan on the

RCT registry of the American Economic Association (https://www.socialscienceregistry.org/trials/405). Particular care was taken in designing the most appropriate sampling frame and sample size for the study, to ensure optimal statistical power, as well as to satisfy ethical and cost concerns. As the intervention is designed to improve English reading achievement in underperforming primary schools, we selected only those primary schools where English is the Language of Learning and Teaching (LOLT) from Grade 4 onward. The second criterion is that only schools that scored at 55% or below on the Grade 4 First Additional Language (FAL) test in both 2012 and 2013 ANA tests in the Pinetown district were eligible for inclusion. The third criterion is that selected schools must have entered between 15 and 120 learners on the FAL Grade 4 ANA test in 2013 (in practice this number was much higher). This was justified on the grounds of cost. One of the two biggest cost-drivers in this intervention is learner support materials, particularly the graded readers which are determined by learner numbers and coaches. It is expensive to provide coaching services to schools with fewer than 15 learners in Grade 4. We also excluded schools classified as Quintile 5 schools, which is the most affluent category of schools according to the official school poverty classification system. Using these criteria, we selected 100 schools to qualify for participation in the study.

For ethical and practical reasons, we sampled intact classrooms within the treatment and control schools. In other words, all learners in a particular grade in a selected school were included in the study. The ethical reason is that sampling classrooms within schools would mean that some schoolchildren would receive the benefits of the treatment or control within a single school and grade, others will not. The practical reason was that if the study had a sub-sample for treatment or control within a school, the language teacher would have be required to teach two different methods simultaneously, which would substantially add to the workload. We assumed, possibly incorrectly that given the size of the province and the relative isolation of many rural schools, there would be little danger of a spill-over effect from the treatment to the control schools.

The study team made the following assumptions when planning the sample:

1. Each school is regarded as an intact cluster for the purposes of calculating standard errors.
2. Only schools that performed below 55% on the FAL Language 2013 ANA are included.
3. Only schools with between 15 and 120 learners (based on 2013 ANA) are included.
4. Only public ordinary schools are included.
5. 80% power level and 5% significance levels .
6. Testing restricted to a random sample within a single grade.
7. ICC value (between-school variance as a proportion of total variance) of 0.20 .
8. Oversampling of control schools relative to intervention schools.
9. A correlation between pre-tests and post-tests of 0.7.
10. Attrition amongst learners would not pose problems to the integrity of the study. Since the pre and post testing occurs within a 12-week period, absenteeism was probably going to be the main cause of attrition, and this would not likely to be systematically different between treatment and control groups. Consequently attrition would not bias the estimated treatment effect.
11. Minimum detectable effects (MDE) set at 0.2 standard deviations .

Given these assumptions, a sample size of 40 treatment schools and 60 control schools was adequate. A computerised lottery was used to randomly allocate schools in the final sampling frame into the treatment and control groups.

Ultimately, these sampling assumptions proved to be conservative – a particularly low intra-class correlation coefficient (0.15) and a high correlation between baseline test scores and endline test scores (0.8) meant that the study was actually powered to identify a minimum detectable effect size of 0.15 standard deviations, which turned out to be about 3.5 percentage points in the reading test. This means that if the true impact of the intervention was to improve reading test scores by 3.5 percentage points (relative to the control group) then we will be 80% sure to obtain a statistically significant estimate of the treatment effect.
In addition to measuring the short-term effect of the intervention on average grade reading performance, we also planned on using official data from Annual National Assessments to measure the longer-run impact of the programme on language achievement. This would provide important evidence on the extent to which short-term remedial interventions, such as the Catch-up Programme, can lead to improvements in educational outcomes.

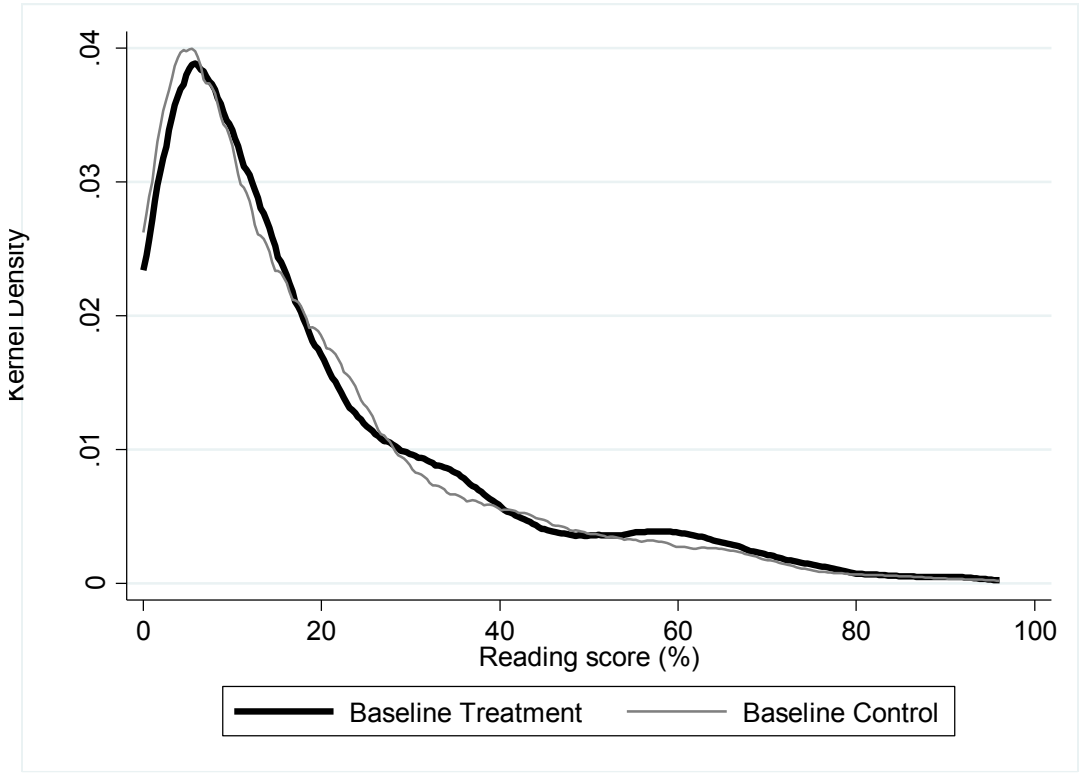
2.6 Pre-Test Learner Results

This sub-section begins with descriptive information on the intervention in Pinetown. This is followed by the presentation of key data from the pre-test.
The original intention was to have a balance of 40 treatment schools and 60 schools in the control group. One problem that occurred was the need to replace three control schools just before the pre-testing began. These schools were replaced at the request of the district office and the reasons provided were legitimate and would have applied equally to treatment schools had it been necessary. This meant that the remaining 57 control schools still serve as a valid comparison group to the treatment schools. For the calculation of results we thus used only these 57 control schools and did not use the three new control schools, because these were non-randomly added by the district office, therefore potentially compromising the validity of the control group. A further challenge was that one control school did not participate in the baseline testing, but did participate in the endline testing. We therefore did not have baseline data for this school.

We obtained data on the pre-test for 2663 learners from 96 schools. For purposes of analysis, however, we only used data from the 2543 learners who also wrote the post-test. The focus of the data analysis of the pre-test was on the effectiveness of test items and to check the balance between the treatment and control schools.
There were 36 numbered test items, a few items with multiple components. As such the total test score was out of 51. The first analysis was designed to ascertain the number of learners with non-responses on items. Non-response could have been due to no answer provided or more than one response provided. 75% of children had six or fewer items with no response. This was positive. Our plan for calculating test scores was to regard non-response as incorrect. Figure 2 shows the distribution of baseline scores (expressed as percentage scores) for both learners in treatment schools and control schools. The figure indicates how similar the distributions of achievement were between treatment and control schools, confirming that the randomisation was successful in producing adequate balance between the two groups. Figure 2 also shows that the vast majority of the learners scored below 20% on the pre-test.

Given the very low scores on the pre-test, concerns were raised about a possible 'floor effect'. This may have had the unfortunate effect of making it harder to identify improvements in learning at the bottom end of the distribution.

Figure 2 Kernel Density of Pre-Test Scores, Percentage



The questions on the cover of the test instrument allowed the research team to analyze some of the characteristics of the study population. Tables 1 and 2 and figure 3 show the performance averages and distributions by age and gender.

Table 1 Baseline Performance by Age of Grade 4 Learners

Age	Mean reading score	Number of learners
8	27.09	11
9	21.80	1072
10	17.41	832
11	13.56	324
12	10.69	97
13	9.59	46
14 and older	17.29	148
Age not specified	16.89	13
Total	18.41	2543

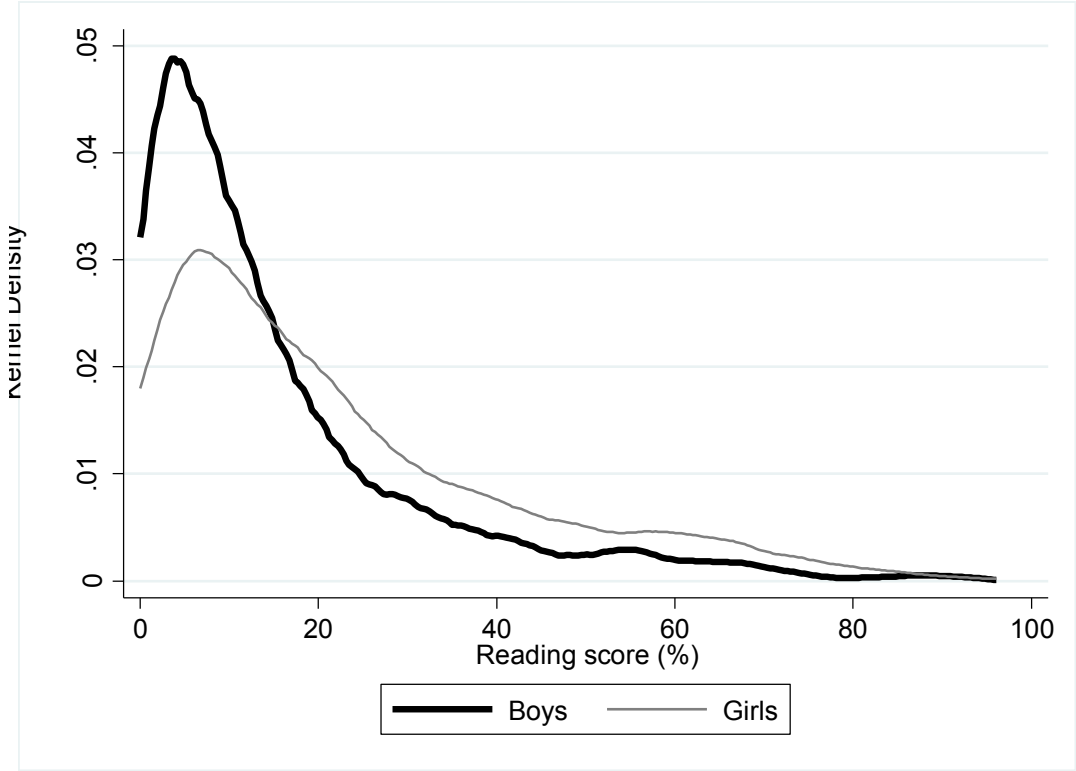
Table 1 reveals that on average, schoolchildren at the 'correct' age to grade had the highest mean scores, with the scores dropping substantially for older learners. What is of concern is the relatively large number of learners (148 out of 2543) who reported their age as 14 years or older, five full years beyond the norm for the Grade 4.

Table 2 Baseline Performance by Gender

Gender	Mean reading score
Boys	14.89
Girls	22.05
Total	18.40

Table 2 and Figure 3 reveal the gender imbalance in performance with girls substantially outperforming boys in the overall sample. This is in line with other South African test results, such as PIRLS 2011 and the Annual National Assessments of recent years, which all show a significant test score advantage for girls especially in literacy.

Figure 3 Distribution of Baseline Performance by Gender



2.7 Implementation

The service provider collected information about the enactment of the programme by teachers or, put differently, the levels of compliance with the programme. Altogether, 79 lesson plans could have been implemented over the period of the intervention, and on average, teachers completed 66% of these lessons. Five teachers completed fewer than half the lessons. Eight teachers completed at least 75% of lessons. We also have information on the number of class assessments (that were provided as part of the intervention programme) that each teacher completed. Twenty seven teachers completed 12 assessments and 13 teachers completed fewer than 12 assessments. The service provider also recorded attendance at afternoon workshops held in small clusters in intervention schools. As Table 3 indicates, some teachers only attended one or two afternoon workshops while others attended five or six.

Table 3 Teacher Attendance at Afternoon Workshops

Number of training sessions attended	Number of teachers
1	9
2	16
3	11
4	8
5	4
6	7
Total	55

The implementing agency reported the following challenges:

- Teachers felt that the pace required by the project was too fast, and they were not used to preparing for or implementing 10 English lessons per week, despite CAPS requirements.
- The second major challenge was related to compliance: preparation; planning and implementation. The afternoon workshops addressed this to some extent, but teachers who did not attend did not get the benefit of these planning sessions. The response to this was initially to offer additional support to non-compliant teachers. However, from mid-May, a decision was made to focus coaching attention on more committed teachers. Non-compliant teachers and principals were aware that post-testing would be implemented.
- Teachers needed support with the technical process of working out average test scores for reporting purposes. In response to this, the implementing agency introduced a 'reward system'. Once teachers were up to date with submissions, and the submissions have been verified against learners' books, they received a pack of stamps / stickers to use when marking the learners' books.
- The poor quality of written work was identified as an ongoing challenge. Teachers generally gave poor instructions, and did not give enough support with regards to written work.
- The management of classroom resources by teachers was another challenge. Teachers did not display the flashcards and other resources in a meaningful way, to reinforce learning that had taken place.
- The use of code switching was pervasive. Some teachers taught the entire English lesson in isiZulu, using English only for key words or phrases.
- Most teachers appeared to welcome the structure, routines, standardised methodologies and content of this project. There was some evidence of improved time on task and work rate, despite the constant tension around pacing.

3. Results

3.1 Attrition

From the perspective of the study design, one of the most positive outcomes of the post-test was the low level of attrition between the pre-test and post-test. No entire schools were lost on follow up. Table 4 shows that attrition amongst learners appears to have been low and not particularly skewed across treatment and control groups.

Table 4 Attrition between Pre-test and post-test, RCUP 2014

	Present at Endline	Not present at Endline	Total
Control	1423	127	1550
	(91.81%)	(8.19%)	(100%)
Treatment	1043	70	1113
	(93.71%)	(6.29%)	(100%)
Total	2466	197	2663
	(92.6%)	(7.4%)	(100%)

Overall, of the 2663 learners who wrote the pre-test, 2466 completed the post-test, which represents a 7.4% attrition rate. The attrition rate was slightly higher in the control group compared to the treatment group. When running a regression to test whether allocation to treatment group predicts attrition it is evident that Treatment does not predict attrition at all once controlling for variables such as baseline scores. Therefore, we delete learners that were absent from the dataset and proceed to analyse the data using only learners present in both the pre-test and post-test.

3.2 Main results

The core question that animated this study focuses on the extent to which learners' achievement in English literacy improved as a result of exposure to the Reading Catch-up Programme. The data show only a very small difference in post-test means between control and treatment school groups. A comparison of the trend lines in the pre and post tests for the treatment and control schools, shows that while both groups improved substantially between the pre and post tests, the improvement is only marginally better in the treatment group. In other words, while the base-line trends were very similar, so were the end-line trends.

Figure 4 Post-test Score Distributions for Treatment and Control Schools

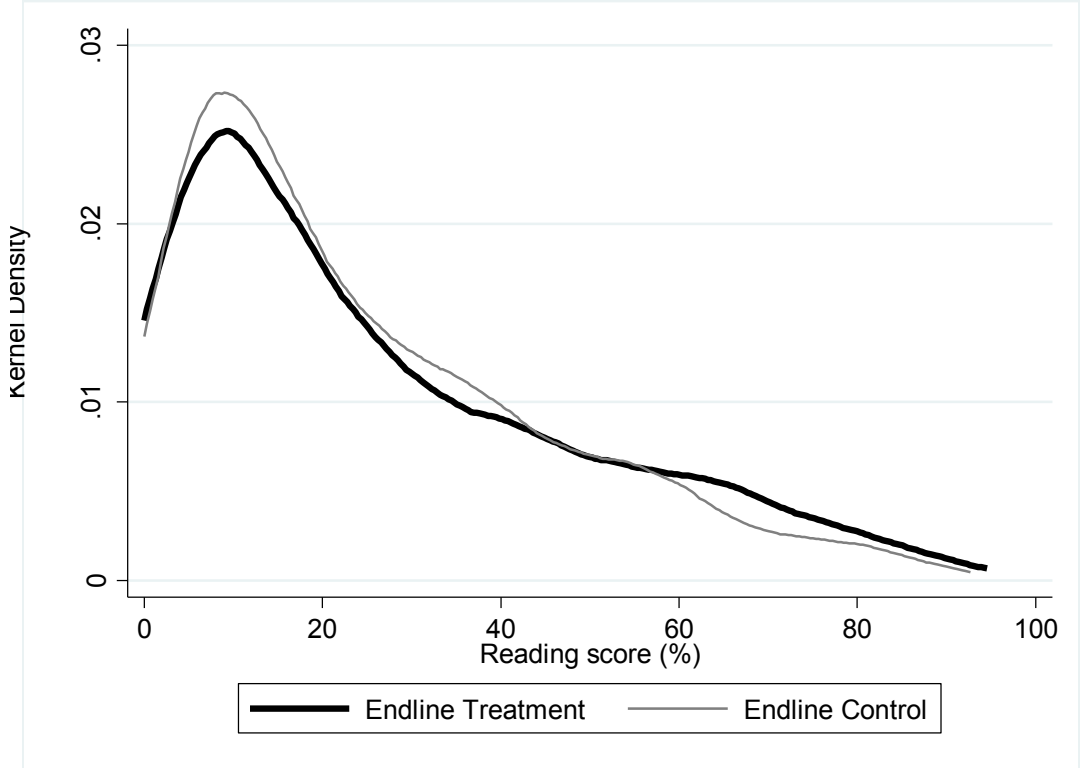
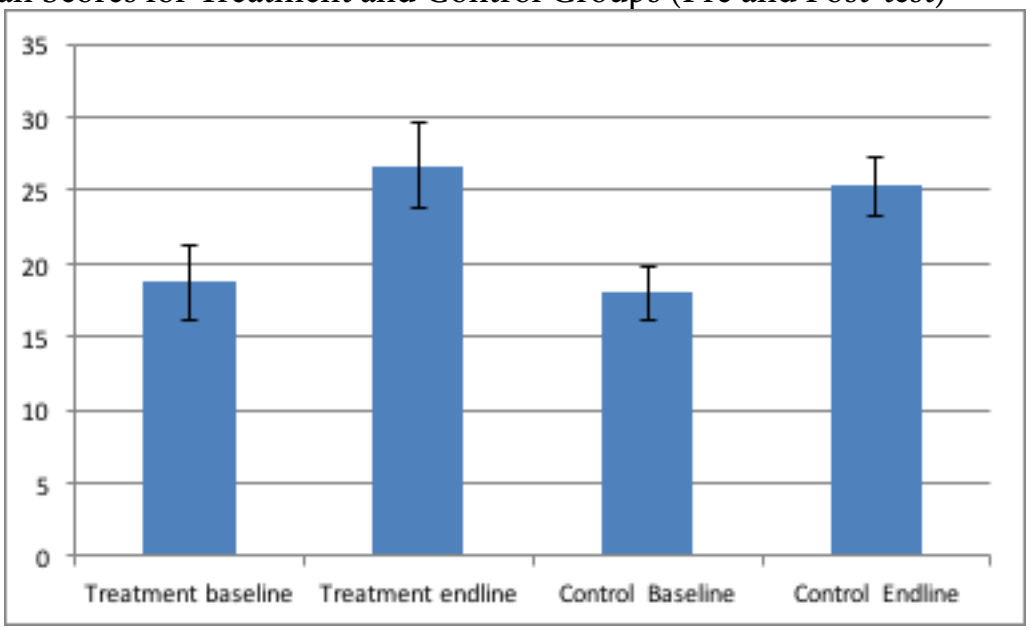


Figure 5 Mean Scores for Treatment and Control Groups (Pre and Post-test)



Note: 95% Confidence Intervals are indicated

The relative small difference between the improvement in the treatment and control schools is clearly evident in figure 5. In statistical terms, although the treatment schools mean post test score was higher than the control group, the difference is not statistically significant.
Table 5 shows the results of five regression models, which represent the most robust methods for estimating the impact of the programme. Column 1 shows the model where the outcome variable is the overall score on the post-test or end-line literacy test. The main explanatory variable of interest is a variable indicating whether the school is a treatment school or a control school. Other variables included in the regression model are the learner's baseline or pre-test score, stratification dummies, learner gender, age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications and school size. Although there is no reason to expect endline test scores to be different between treatment and control schools other than because of the intervention, it is still worth including these other control variables in order to enhance the statistical precision of the estimated treatment effect. Only the coefficient on the treatment variable and the standard error of the estimate are reported in Table 5, but all the above-mentioned controls were included. Columns (2)-(5) on the right of the table refer to models with the same set of explanatory variables but the outcome variables are learner scores for each of the four literacy domains which formed part of the reading test.

Table 5 Main Regression Results

	(1)	(2)	(3)	(4)	(5)
	Overall score	Spelling	Language	Comprehension	Writing
Treatment	0.49	1.27**	3.96***	-1.40	1.14
SE	(0.67)	(0.61)	(1.07)	(1.34)	(1.40)
Observations	2466	2466	2466	2466	2466
R-squared	0.77	0.77	0.46	0.53	0.28

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

The estimated treatment effect on the overall literacy score is an additional 0.49 percentage points relative to the control group. However, we are unable to conclude with any level of statistical confidence that the true effect is different from zero. On the other hand, we are able to conclude with high levels of statistical confidence that the intervention improved spelling outcomes and language outcomes for learners in treatment schools. We estimate that spelling improved by 1.27 percentage points relative to the control group and that language improved by 3.96 percentage points. The estimated impact on comprehension and writing items was not statistically different from zero.

3.3 Heterogeneous treatment effects

We also investigated so-called 'heterogeneous effects' – whether the impact of the programme was different depending on various learner, school or teacher characteristics. There was no evidence of heterogeneous effects based on learner gender, learner age, learner' exposure to English at home or class size (full results not reported here). In planned forthcoming analysis we will continue to investigate heterogeneous effects according to other characteristics as outlined in the Pre-Analysis Plan.

The following analysis (table 6), however, points to the possibility that the impact was larger for children who initially performed better on the baseline test. The result is statistically significant for spelling. Although not significant in language the size of the coefficient is actually larger than that for spelling so it may be that the same was true for language and we are simply unable to conclude so with statistical confidence. For spelling, there was effectively no impact on those who had initially scored poorly (and there were indeed many zero scores). The coefficient on the interaction term indicates that every additional 10 percentage points on the baseline test was associated with an increased treatment effect of 0.5 percentage points.

Table 6 Impact by Baseline Performance of Learners

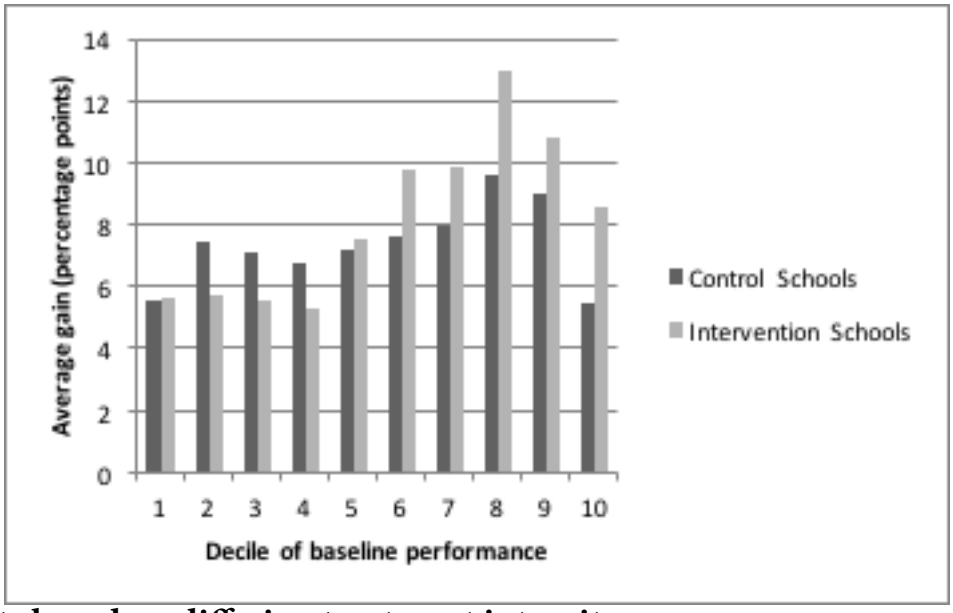
	Combined score	Spelling	Language	Comprehension	Writing
Treatment	-0.44	0.32	2.92**	-1.96	1.39
	(0.86)	(0.7)	(1.15)	(1.54)	(1.47)
Baseline percentage score	0.97	0.93	0.66	0.79	0.49
	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)
Treatment X Baseline	0.05	0.05*	0.07	0.02	-0.03
	(0.04)	(0.03)	(0.04)	(0.04)	(0.06)
N	2466	2466	2466	2466	2466
r2	0.77	0.77	0.46	0.53	0.28

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for baseline score, stratification dummies, learner gender, age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

Although the regression analysis above did not conclusively indicate that programme impact varied according to baseline learner performance on the combined test score, some

descriptive analysis points to the strong possibility that it did. The result in Table 4 may be a functional form issue. The following graph shows the average gain score for learners in treatment and control schools by each decile of baseline performance. Deciles are ten equal sized groups of learners split according to baseline performance. So, Decile 1 includes the bottom 10% of learners on baseline performance.



3.4 Effects based on differing treatment intensity

The main estimate of the programme impact as reported in Table 5 is conventionally referred to as the "Intent to Treat" (ITT) estimate, where allocation into the treatment group indicates an intention that these schools receive the intervention. However, when compliance with the intervention is not uniform we are also interested to measure what is called the "Treatment on the Treated Effect" (TTE), i.e. the effect of the intervention for those who complied with the intervention. In our particular situation we are not able to retrieve the TTE since compliance is not a zero or 1 categorization but rather there were varying levels of compliance. Therefore, we are only able to show descriptive statistics of the average learner gains depending on how many training sessions teachers attended (Table 7). Note that zero category includes the control schools. The gains were highest when teachers attended at least three training sessions, pointing to the possibility that the success of an intervention such as RCUP may depend on the extent to which teachers engage with it.

Table 7 Average learner gains by number of training sessions attended by the teacher

Number of sessions	Language	Spelling	No learners
0	5.37	4.63	1606
1	1.22	2.45	74
2	4.23	4.40	189
3	8.58	6.46	254
4	8.87	6.18	142
5	6.73	2.95	101

Did the impact of the intervention depend on which coach the school was allocated? The service provider used two coaches to implement the programme. Each coach was allocated 20 schools. Therefore, one can estimate two separate treatment effects, one for each coach. Table 8 shows the results when running the exact same regression models as reported above but instead of including a single treatment dummy variable, we include two dummy variables

(one for each coach), still relative to the reference category of control schools. There are two main limitations in this analysis. Firstly, the coaches were not randomly assigned to schools. However, the fact that we have baseline scores for each learners and can control for stratification and other learner, school and teacher characteristics reduces the likelihood of omitted variables bias. Secondly, the effective sample size is cut in half – instead of a treatment group of 40 schools we now compare each treatment group of 20 schools to each other and to the control group. This means that standard errors will be larger and therefore we are less likely to observe a statistically significant treatment effect.

Table 10 shows no significant impact for coach B on any of the outcomes. For Coach A, however, there were statistically significant effects on both spelling and language. The coefficients for Coach "A" are all larger than in the overall treatment effects as reported in Table 5 (though we cannot conclude with statistical certainty that the effects are larger). Therefore, this provides suggestive evidence that the success of an intervention that uses coaches to support teachers may depend on the particular person doing the coaching. If indeed, this was the case, we are not able to determine what characteristics of Coach "A" led to a larger impact.

Table 8 Impact of Coaches

	Combined score	Spelling	Language
Coach A	1.42	1.98**	5.87***
	(0.93)	(0.76)	(1.42)
Coach B	-0.42	0.56	2.09
	(0.89)	(0.88)	(1.41)
N	2466	2466	2466
r2	0.7698	0.7692	0.4606

Note: *p<0.1 **p<0.05 ***p<0.01

The reference category for both coaches is the control group. All models include controls for baseline score, stratification dummies, learner gender, age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

3.5 Impact on Annual National Assessments

The Annual National Assessments (ANA) of 2014 were written during the week of 16 – 19 September across South African schools. This was about three months after the RCUP intervention was finished. All children in grades 1 to 6 and 9 wrote a mathematics test. Children in grades 1 to 3 wrote a Home Language test (in Pinetown it was in isiZulu). Children in grades 4 to 6 and 9 wrote one of the following language subjects: English Home Language, Afrikaans Home Language, English as First Additional Language or Afrikaans as First Additional Language. In Pinetown, 94% of learners in our sample of treatment and control schools wrote English as First Additional Language.

There are several hypotheses which the availability of ANA data allowed us to investigate:

i. The treatment effect for intervention schools relative to control schools may diminish

over time or it may grow through continued use of the new materials and pedagogies.
ii. An improvement in literacy may benefit other learning areas, such as mathematics.
iii. Although the intervention targeted grade 4 teachers in a school, there may be spill over benefits to other grades.

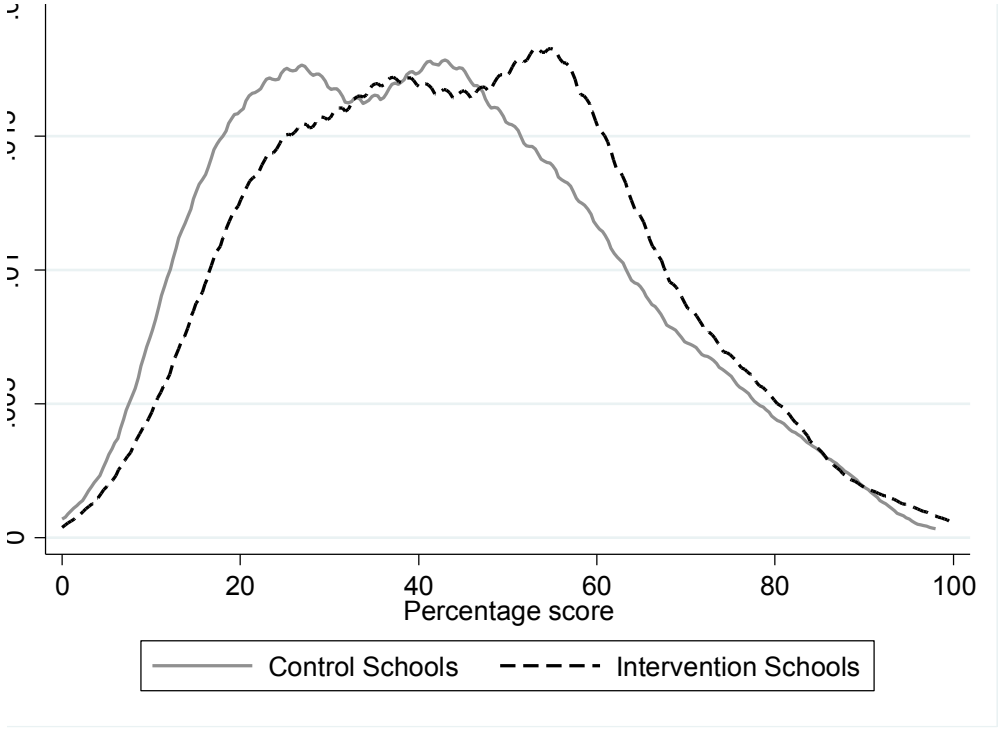
The third hypothesis is especially possible since the majority of grade 4 teachers in South Africa also teach in another grade. We used ANA data for literacy in grades 1, 2, 3, 5, and 6, to see whether students in untreated grades in intervention schools improved relative to students in control schools. We also used ANA data for mathematics in Grade 4 to ascertain possible impact of the treatment on other subjects.

There are, however, several limitations of the ANA data for our purposes. The data quality is not expected to be as high as that collected by our service provider. This is because the ANA tests were locally administrated and marked by teachers within each school. Differences in the conditions of testing and in marking standards across schools should make the data a somewhat noisy signal of learner proficiency. This is confirmed by the respective correlations between our baseline test score, our end-line test score and the ANA language scores of learners. In a sample of 1928 learners who we were able to match between the RCUP and ANA datasets, the correlation coefficient between the baseline test score and the end-line test score was 0.86. However, the correlation coefficient between baseline score and ANA English score was only 0.53 and between endline score and ANA English score was 0.56. Noisy data would be expected to cause a degree of attenuation bias in the estimated treatment effects (where the estimated effect is biased towards zero). Fortunately though, there is no reason to expect differences in marking or the quality of ANA information to be correlated with assignment to treatment.

In the first analysis using ANA data we use all learners in treatment and control schools, i.e. not only those learners who were sampled for our own independent testing. This provides us with a dataset of 6419 learners across our treatment and control schools. While this improves the statistical power for identifying a treatment effect, the disadvantage of this approach is that we do not have a baseline score for each learner. The best we can do is to control for each school's average ANA score in previous years.

The average score in Grade 4 English as a First Additional Language within our sample of schools was 43.0%. As was the case in our independently administered tests, girls (average score of 46.8%) substantially outperformed boys (average score of 39.4%). Importantly, the male disadvantage was still large (about 6 percentage points) in all our multivariate regression models even after controlling for other characteristics such as age (boys are noticeably older than girls on average). Although this finding is not central to this paper, it confirms an increasingly clear pattern of a large learning disadvantage for males in South African schools. Figure 7 presents Kernel Density curves showing the distributions of test scores for those in intervention schools and those in control schools. This indicates that learners in intervention schools had a somewhat better distribution of achievement than those in control schools. This is a preliminary indication of a positive treatment effect.

Figure 7 Kernel Density Curves of Test Scores for Grade 4 English as First Additional Language (ANA, 2014)



The first hypothesis to test is whether learners in intervention schools performed better in the Grade 4 English ANA test than those in control schools. When no attempt is made to control for baseline differences in achievement, the estimated treatment effect is 3.35 percentage points and this is statistically significant at the 90% level. Models 2, 3, 4 and 5 in Table 9 show the estimated treatment effect when different ways of controlling for prior school performance are used (controlling for school mean language score in ANA 2013, controlling for school mean language score in ANA 2012, controlling for both school mean language score in ANA 2012 and in 2013). In all cases, the estimated treatment effect is somewhere between 3 and 4 percentage points but in models 4 and 5 it is not statistically significant.

Table 9 Treatment Effect on Grade 4 English First Additional Language (ANA)

	Model 1	Model 2	Model 3	Model 4	Model 5
Treatment	3.35*	3.83**	3.49**	3.13	3.26
	(1.93)	(1.88)	(1.72)	(1.95)	(2.14)
School mean grade 4 LANG 2013	No	Yes	Yes	No	Yes
School mean grade 4 LANG 2012	No	No	Yes	No	No
School mean RCUP baseline	No	No	No	Yes	No
School mean grade 3 LANG 2013	No	No	No	No	Yes
N	6419	6419	6419	6419	6055
r2	0.1731	0.1914	0.2072	0.1753	0.2042

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender and age. Standard errors are adjusted for the fact that learners are clustered in schools. Since a few schools wrote English as Home Language, only 91 schools are represented in the models (37 treatment and 54 control). The results are robust to an alternative specification where the outcome variable is percentage score irrespective of whether this was from the English as Home Language test or the English as First Additional Language test.

Was there a spillover benefit observed in mathematics scores of learners who had been exposed to the catch-up programme? Since the mathematics test is written in English it is plausible that an improved English proficiency thanks to the RCUP intervention would have led to improved mathematics scores. As reported in Table 10, although the estimated treatment effect on mathematics scores was positive it was not statistically significant. Therefore, we cannot conclude that the intervention led to improved mathematics performance.

Table 10 Effect of Treatment on Grade 4 mathematics (ANA)

	Model 1
Treatment	2.38
	(2.58)
Baseline school average 2013	Yes
Baseline school average 2012	Yes
N	6687
r2	0.2153

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender and age. Standard errors are adjusted for the fact that learners are clustered in schools.

Was there a spillover benefit observed in language performance for other grades at treatment schools? The results in Table 11 indicate that there was a positive effect for the grades either side of the treated group, i.e. grade 3 and grade 5. The fact that the majority of grade 4 teachers in South Africa teach in another grade strengthens the plausibility of this result. On the other hand, it seems less likely that grade 3 Home Language (isiZulu) would improve through an English intervention at Grade 4. Therefore, we recommend that no strong conclusions be made on the basis of this result.

Table 4 Treatment Effect on Language across Untreated Grades (ANA)

	Grade 1	Grade 2	Grade 3	Grade 5	Grade 6
Treatment	1.55	-1.01	5.80***	3.49**	3.26
	(1.82)	(1.50)	(1.80)	(1.60)	(2.07)
Baseline school average 2013	Yes	Yes	Yes	Yes	Yes
Baseline school average 2012	Yes	Yes	Yes	Yes	Yes
N	9144	7673	7089	5341	4963
r2	0.1131	0.0958	0.1577	0.2083	0.2407

Note: *p<0.1 **p<0.05 ***p<0.01

All models include controls for stratification dummies, learner gender and age. Standard errors are adjusted for the fact that learners are clustered in schools. Since a few schools wrote English as Home Language, only 91 schools are represented in the models (37 treatment and 54 control). For grades 1, 2 and 3 the test was a home language test whereas for grades 5 and 6 the test was English as a First Additional Language.

As before, we test whether there was a different treatment effect for each coach. The results are very similar to those observed when using the independently administered test data. For Coach "A" there was a fairly large and statistically significant treatment effect, whereas no significant effect was observed for Coach "B". However, as before we cannot actually say with statistical certainty that the effect for Coach "A" was larger than that for Coach "B".

Table 12 Impact of Coaches (ANA)

	ANA language
Coach A	5.38**
	(2.45)
Coach B	0.85
	(1.70)
N	6419
r2	0.2106

Note: *p<0.1 **p<0.05 ***p<0.01

The model includes controls for both school mean language score in ANA 2012 and in 2013, stratification dummies, learner gender and age. Standard errors are adjusted for the fact that learners are clustered in schools.

3.6 Analysis of sub-sample of individuals participating in both RCUP testing and ANA

Out of the 2466 learners with valid pre and post-test scores in the final RCUP dataset we were able to identify 1928 learners in the Universal ANA dataset of 2014. We matched learners using the first three letters of their first names, the first three letters of their surnames, their gender, their school and their grade. This led to some duplicates where individuals were identical based on these variables. We therefore dropped all such individuals to avoid the possibility of false matches. There are several other possible reasons why we would have not identified all learners in the ANA data. It may have been that some learners were absent on the day of the ANA testing. Some learners may have participated in the Verification ANA testing, in which case their ANA marks would not be present in the Universal ANA dataset. Some learners may have participated in Universal ANA but due to incomplete data capturing their results were not uploaded onto the national dataset. There may have been errors in the information used to match learners across the two datasets, i.e. they may have misspelt their name or surname in one of the datasets.

The advantage of using individuals with both RCUP information and ANA test scores is that we can control for a baseline score for each learner, namely the baseline score on the RCUP test. We ran a regression to check whether treatment status predicts being successfully matched in the ANA data. This indicated no statistically significant relationship between

being in a treatment school and being found in the ANA dataset. Therefore, we can analyse the results on the ANA tests for treatment and control schools without fear of any selection bias that might influence the estimated treatment effect. This is further confirmed by the fact that when we run the exact same regression as the main model in Table 6 (i.e. predicting RCUP endline scores) but on the sub-sample of 1928 matched learners we obtain essentially the same estimated treatment effect (a coefficient of 0.48 as opposed to 0.49).
Table 13 reports the results of the two models we ran on the individually matched sub-sample. The outcome variable is percentage score in grade 4 English as First Additional Language. The magnitude of the coefficients observed in Table 15 are broadly consistent with earlier results throughout the paper – namely a relatively small positive effect of being in the treatment group, a larger positive effect for Coach "A" and a negligible effect for Coach "B". However, all coefficients of interest in these two models are not statistically significantly different from zero. The effect sizes are non-negligible which means we were somewhat underpowered, especially in the case of the coach-specific models. The overall conclusion to draw from this analysis remains as follows: there is tentative evidence of a fairly small effect of the intervention on performance and this effect appears to have been larger for Coach "A", but we cannot make these conclusions with a high level of statistical certainty.

Table 5 Impact of Treatment on ANA language scores (for individually matched sample)

	Model 1	Model 2
Treatment	2.40	
	(2.20)	
Coach A		4.89
		(3.39)
Coach B		0.14
		(2.38)
N	1928	1928
r2	0.4643	0.4676

Note: *p<0.1 **p<0.05 ***p<0.01

The reference category for both coaches is the control group. All models include controls for the baseline score of the learner in the RCUP testing, stratification dummies, learner gender, age, exposure to English at home, frequency of an adult reading at home, class size, teacher age, teacher gender, teacher qualifications and school size. Standard errors are adjusted for the fact that learners are clustered in schools.

4. Discussion

Even though the increases in the learner spelling and language scores in the treatment schools are statistically significant, and the ANA scores show statistically significant relative gains compared to the control schools, the gains may have limited educational significance. The effect sizes as measured by standardized scores were relatively small compared to the gains suggested in the original 2012 Reading Catch-up study and in Pretorius' (2014) new study . A scan of a sample of learners' post-test scripts from amongst treatment schools clearly shows that most of the Grade 4 learners continue to be very weak spellers with limited command of basic structures of the language, comprehension and writing. The gap between these learners' literacy performance and the demands of the curriculum remains large.

The core hypothesis that intermediate phase learners' literacy proficiency could be 'caught-up' across a 'sub-system' using a well-designed ten week intervention is simply not supported by the evidence from this randomised control trial. That said, there is evidence to suggest that with higher levels of implementation intensity and/or extended duration and with strong coaching, interventions like the Reading Catch-up Programme could indeed enable learners to narrow the gap between their actual literacy performance and the expectations of the official curriculum, particularly around domains such as spelling and language. The potential for improvement through this sort of programme appears larger for those learners who are not at the very bottom of the performance distribution.

Before exploring substantive reasons for the low estimated impact of the programme on reading outcomes, it is worth highlighting a few possible measurement limitations that may have contributed to this outcome. While there was a substantial increase between the pre-test and post test, the gains was very similar for treatment and control schools. Why would there be such a dramatic gain in the control group? A number of explanations can be offered. Firstly, it may simply be that soon after beginning with English as the language of instruction (as occurs in grade 4) learners typically demonstrate quick gains in basic vocabulary. If this is the case, then the large gain in the control group is perfectly legitimate and in no way biases the results of this study.

Another possibility relates to the Hawthorne effect, that irrespective of whether a school was assigned to the control or the treatment group, they were all subject to external scrutiny particularly around learner performance testing (i.e. pre and post testing). The very fact of being tested by an external agency in and of itself might have been the impetus for more engaged teaching and learning, particularly as schools are increasingly concerned about possible high stakes consequences of the new annual national testing policy. If a Hawthorne effect was present for the control schools then this is not a problem for the study design since treatment schools would also have experienced a Hawthorne effect through having been tested and these effects would cancel each other out. We are precisely interested in the effect of the programme over and above any effects of testing.

A potentially problematic possibility is that there was an unanticipated spill-over effect, where schools that were part of the control group received some of the benefits of the RCUP intervention through informal sharing between schools. Further analysis of the data will be conducted to investigate whether this may have occurred but it seems unlikely that this would have occurred to any great extent since the main aspects of the programme were not easily transferable (coaches and materials).

A third explanation may be found in the 'floor-effect' evident in the pre-test results. While the decision to employ the identical instrument used in the original Gauteng study was deliberate and would theoretically have allowed for precise comparison of gain scores, the context in Kwazulu-Natal might mean that learners in that province have considerably lower access to English vocabulary and literacy in English in general than counterparts in Gauteng. A different instrument, one that emphasized Grade 1 English FAL questions might have provided results more closely resembling a normal distribution. Such an instrument might have revealed gains at the lowest levels of literacy.

Notwithstanding the above questions, the statistically significant findings of gains in two domains, spelling and language (grammar), are important. These are clearly the domains most likely to change as they have the lowest cognitive load associated with them. Should learners have encountered the words directly during the ten weeks of lessons or mastered some aspect of English phonics, it would be reasonable to expect that this learning would be evident in the post-test and a few months later in the ANA test. Similarly, explicit teaching of basic language structures, such as capital letters at the beginning of the sentence and full stop at the end, would carry through to improved scores on the language section of the post-test. In contrast, the fact that comprehension scores did not change, which requires a much wider and more complex range of knowledge and skills to be taught and learnt, is not surprising given the relative brevity of the intervention.

While the main finding shows little real difference in gains between the treatment and control groups of schools overall, the more nuanced analyses provide important insights into the possible conditions under which meaningful change, what Hopkins (2003) described as 'improvement for real', could occur. The analysis suggests that the more extensively teachers participated in the intervention (as measured by the number of training sessions attended), and the higher their commitment or enthusiasm (as measured by the number of lessons covered and assessments administered), the stronger was the programme's effect on their learners' spelling and language performance. An added insight that emerges, one that will require new studies to confirm, is the differential impact of individual coaches. The RCUP findings suggest that while instructional infrastructure (Cohen, 2011) in the form of lesson plans, learner resources and coaches may be necessary conditions for improvement, the quality and effectiveness of individual coaches may be a often hidden but powerful factor.
Notwithstanding this strong finding, the study has also provided substantial evidence around a range of themes. These included further evidence of the serious under-performance in English as a first additional language at the start of the Intermediate Phase and the scale of the gender performance gap.

The study pre-test dataset suggested that the Grade 4 learners' English language knowledge and skills is very weak. Pinetown was selected as one of the higher performing districts in the province as indicated by the ANA scores. Our findings, however, suggest that there is a significant discrepancy between the performance levels indicated by ANA scores and proficiency levels as measured by our test. The divergent performance measures may be a function of the different test instruments or of the different conditions under which the tests were administrated and marked.

Another major insight from the pre-test analysis is the large performance gap between boys and girls. This gap is evident both in the pre and post tests and is consistent between the study tests and the ANA results. This trend, identified by Perry (2006) in the early 2000s and recently confirmed by Zuze (2014), is not adequately understood.

5. Conclusion

The Reading Catch-up Programme has been shown to have little educationally significant impact. The results of this study, robust as they are, do not suggest any specific policy or programme warrants. The lesson, however, for policy makers is that policy or programme effectiveness claims can and should be tested using robust counterfactual studies prior to system-wide rollout.

The study demonstrates the value of counterfactual research. If this study had used a simple pre- and post-test design (as was used in the initial study), the conclusion would be a false positive, namely that the intervention was highly effective. Having a randomly selected control group to provide a valid estimate of the counterfactual allowed us to observe similar gains for the control group and by extension, that improved performance cannot be simply ascribed to the intervention. The study also shows the value of replication studies to address questions of external validity. Assuming the results of the Gauteng study were reliable and valid, this study demonstrates policy transfer cannot automatically be assumed. This may be because of substantive differences in language context and language practice across provinces.

Finally, while the study was explicitly designed as an impact evaluation, the data collected for the study are likely to be fertile ground for a number of additional secondary studies. Thanks to a generous grant from the National Research Foundation, a number of graduate students are likely to undertake more fine-grained analyses of first additional language literacy acquisition in our schools.

References

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics. An empiricist's companion. Princeton, New Jersey: Princeton University Press.
Cohen, D. K. (2011). Teaching and its predicaments. Harvard University Press.
Educational Evaluation and Policy Analysis. (2007) 29/1.
Fleisch, B (2008) Primary Education in Crisis: Why South African Schoolchildren Underachieve in Reading and Mathematics. Cape Town: Juta.
Fleisch, B., Taylor, N., Herholdt, R., & Sapire, I. (2011). Evaluation of Back to Basics mathematics workbooks: a randomised control trial of the Primary Mathematics Research Project1. South African Journal of Education, 31(4), 488-504.
Fleisch B (2013a). Change at the Instructional Core: Insights from the Intersen English Catch-Up Programme. Paper Presented at the South African Education Research Association Meeting, January 2013.
Fleisch, B (2013b). System Reform and Primary Literacy: Implication for School Leadership. In I R Haslam, M S Khine & I M Saleh (eds.) Large Scale School Reform and Social Capital Building. Routledge.
Hellman, L (2012) GPLMS Intersen Catch-up Programme: Analysis of Results. Memo.
Hopkins, D. (2003). School improvement for real. Routledge.
Kremer, M., Brannen, C., & Glennerster, R. (2013). The Challenge of Education and Learning in the Developing World. Science , 340, 297-300.
Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-Based Practices in a Changing World Reconsidering the Counterfactual in Education Research. Educational Researcher,
National Education Evaluation and Development Unit (2013)
Perry, H., & Fleisch, B. (2006). Gender and educational achievement in South Africa. In V. Reddy (Ed.), Marking matric: Colloquium proceedings (pp. 107-26). Cape Town: HSRC Press.
Pretorius, E. J. (2014). Supporting transition or playing catch-up in Grade 4? Implications for standards in education and training. Perspectives in Education, 32(1), 51-76.
Pritchett, L., & Beatty, A. (2012). The negative consequences of overambitious curricula in developing countries. Center for Global Development Working Paper, (293).
Raudenbush, S (2005), Learning from Attempts to Improve Schooling: The Contribution of Methodological Diversity Educational Researcher, 34/1 25-31
Robinson, D & Levin J (1997) Reflections on Statistical and Substantive Significance, with a Slice of Replication Educational Researcher, 26/5 21-26
Spaull, N. & Kotze, J. (2015). Starting Behind and Staying Behind in South Africa: The Case of Insurmountable Learning Deficits in Mathematics. International Journal of Educational Development.
Taylor, N (2007) Equity, Efficiency and Development in South African Schools. In T Townsend (ed.) International Handbook of School Effectiveness and Improvement. Dorchrecht, Netherlands: Springer
Taylor, N, van der Berg, S & Mabogoane T (eds.) (2013) Creating Effective School. Pearson.
Van Staden, S., & Bosker, R. (2014). Factors that affect South African Reading Literacy Achievement: evidence from prePIRLS 2011. South African Journal of Education, 34(3), 01-09.
Zimmerman, L., & Smit, B. (2014). Profiling classroom reading comprehension development

practices from the PIRLS 2006 in South Africa. South African Journal of Education, 34(3), 01-09.
Zuze, T. L., & Reddy, V. (2014). School resources and the gender reading literacy gap in South African schools. International Journal of Educational Development, 36, 100-107.

Endnotes

[1] We use the terms reading and literacy interchangeably. While the RCUP programme was clearly geared to improving reading proficiency, the test instrument was oriented toward the measurement of certain literacy skills rather than oral reading fluency and comprehension.

[2] The terminology of "treatment" and "control" groups originates from the literature on medical trials, where a particular drug or "treatment" was being trialled. The terminology is now widely used across fields in impact evaluations. We use "intervention" group and "treatment" group interchangeably.

[3] A paper by Kremer, Brannen and Glennerster (2013) provides a concise review of international RCT studies focussing on education.

[4] Initially, we tried to select schools based on the original ANA 50% and 30 and 90 learners criteria. But in order to find 100 schools we had to start relaxing some of these criteria. Read the full sampling report in the Pre-Analysis Plan to see exactly what we did.

[5] The power of the statistical test refers to the probability of avoiding a Type II error (i.e. incorrectly rejecting a null hypothesis). Therefore it represents the likelihood of drawing the correct conclusions about the significance of differences between groups. Typically, a power level of 80% is considered high enough to detect differences while keeping sample sizes reasonable.

[6] The ICC is the proportion of the total variation in test scores that is accounted for by between-school variation; the remainder is accounted for by with-school variation amongst students. It describes the level of inequality between schools. The higher the ICC, the larger are the systematic differences in achievement scores between schools and the more groups required in the sample.

[7] In order to determine appropriate sample size, it is necessary to have some prior knowledge of expected size of the intervention effect. In much of the contemporary US based literature this is has been standardized to a common effect size unit, i.e. percentage of the standard deviation of the outcome measure. This allows for comparison across studies using different scales. While the original PRMP study did not report results in percentage of the standard deviation of the outcome measures, the percentage point gains reported were very high. The use of 0.2 standard deviations can be regarded as a moderate effect size relative to those typically observed in the international literature on school interventions.

[8] Given this core finding, the question of cost-effectiveness is of no consequence.

[9] An internal DBE analysis of the Annual Survey of Schools indicated this.

[10] Although no baseline score is inserted as a control variable, there is no reason to expect substantial baseline differences between treatment and control schools because of randomisation. Moreover, we include the strata dummies in the regression to further control for differences in school characteristics, including prior ANA achievement, which was one of the dimensions influencing stratification.

[11] Although Pretorius' study used only one school and therefore should not be considered as a benchmark for typical effect sizes.

[12] Lemon (2014) describes similar patterns with strong gains in counterfactual study groups. Their account however stresses the shifts in the entire school system as a result of improved early reading teaching.

Theory of Change Development for South African Education Programmes: A means to increase ownership and utilisation of evaluation findings.

First author
Ms Eleanor Hazell, Acting Executive Manager: Monitoring and Evaluation, JET Education Services
Email: eleanorhazell@gmail.com or ehazell@jet.org.za Cell 084 8333880

Second author
Ms Hazel Mugo, Researcher: Monitoring and Evaluation, JET Education Services
Email: hmugo@jet.org.za Office 021 686 3351

Third author
Ms Carmel Marock, Singizi Consulting
Email: carmelmarock@singizi.co.za Office (011) 728 2889

Abstract

Developing or documenting "theory of change" (toc) forms the basis for designing and undertaking a theory-based evaluation (TBE). Ideally toc should be developed in the design or inception stages of a programme and updated at regular intervals, but in reality toc development often happens only when an evaluation is commissioned. A variety of methods can be used and the process may range from highly participatory to evaluator driven (Coryn et al, 2011).

This paper reflects on toc development for three evaluations of education programmes, to discuss the advantages and pitfalls of different methods, how toc development can stimulate dialogue and consensus building and may increase ownership and utilisation of evaluation findings.

A toc was developed for the Funza Lushaka Bursary Programme (FLBP) evaluation. A participatory workshop was facilitated with 60+ stakeholders to document key project events, identify blockages in business processes and refine the programme goal, objectives and intended outcomes. Delegates left the workshop energised, with a better understanding of their contribution to the bigger picture and how the success of the programme would be measured, through evaluation and other means.
An evaluation was conducted at the close-out of the Bojanala Systemic School Improvement Programme (BSSIP), to assess the extent to which it achieved its aims. A logframe was developed to guide implementation of the BSSIP; however, this did not specify the outcomes that were expected to occur. The "plausibility" of the logframe was tested and a detailed toc developed. Hence, a toc was developed retrospectively, through a programme document review and literature review. The process was evaluator driven. Presentation of findings stimulated discussion on what is possible and plausible for a school improvement programme to achieve and may lead to a review of JET's School Improvement model.
The toc was documented at the beginning of a complex College Improvement Project and revised during Project close-out; workshops were facilitated with the Project team, but the process ran parallel-to rather than integrated with other Project processes. At the end of the project the toc was modified to reflect the evolution of the project and the revised scope which reflected what was considered possible to do and achieve. The toc guided the summative evaluation, ensuring the focus was on what the Project attempted to achieve – which remained consistent - and on what was actually done – which changed.

Introduction

"Theory of change" (toc) is used in evaluation, to refer to an explanation (or theory) of how a project or programme works to bring about an intended change. Theory–based evaluation (TBE) is premised on evaluating a project or programme in relation to its stated toc. This can involve assessing the extent to which the toc sound, if the project/programme implemented as intended, whether the anticipated changes outlined in the toc occurred and what internal and external factors influenced the extent to which the toc occurred. If the toc has not been documented, or it has not been adequately documented – as is often the case – then this is usually one of the first tasks of the evaluator. However, toc, can have a number of uses beyond evaluation, it can support programme design and guide planning, implementation and internal review/reflection (Vogel, 2012).

This paper – written by monitoring and evaluation staff and a consultant working for JET Education Services (JET) – discusses JET's recent experiences in developing toc for the evaluation of South African education programmes. Additionally, we present lessons which can be drawn for improving the utilisation of toc in and beyond evaluation and for building an evaluation culture in South Africa.

Background to JET and JET's use of toc

The Joint Education Trust (JET) was established in 1992. As South Africa prepared to transition from apartheid to democracy, 14 major South African companies pledged to provide R500 million to restructure and rebuild the national education system and address injustices of the past. A total of R560 million was raised from South African companies, which, in turn, leveraged R680 million from international donors. JET was established to disburse funds to non-governmental organisations (NGOs) working in the education sector in five focus areas: early childhood development; adult basic education and training; vocational and further education; in-service teacher training and development; and youth development (JET, 2015).

Being responsible for the disbursement of substantial funding for education improvement, led JET to invest in evaluation - which was in its infancy in South Africa at the time (Mouton, 2010) - as a mechanism for improving accountability (Robbins, 2001). Three initiatives were undertaken by JET: 1) a survey of each education focal area which was supported, to map the landscape and understand the strengths and weaknesses of stakeholders working in it; 2) development of mechanisms which linked NGO service delivery to the public education system, for example certification and accreditation systems; 3) a review of evaluations (i.e. a meta-evaluation) of NGO teacher development programmes (Ibid). The review of teacher development evaluations commissioned by JET was the first time that a "meta evaluation" had been conducted in South Africa. It generated substantial information about the extent, nature and rigour of evaluation of education programmes (Mouton, 2010). The findings were discussed at a 1996 conference titled Quality and Validity, hosted by JET and attended by local and international evaluators (Ibid).

After executing its founding mandate, JET's focus shifted to managing education and development projects, and then to providing education delivery support and conducting education evaluation and research (Ibid). At least 32 evaluations were carried out by JET between 1995 and 2014.

A review of a sample of JET evaluations going back as far as 1995 (Zauyamakando, 2012) found limited use of logic models and/or documented toc - to clarify the causal relationships and assumptions critical for successful implementation and guide evaluation. However, a number of more recent evaluations conducted by JET have been theory-based and were guided by a documented toc.

The origins of toc

The use of toc in evaluation is a form of TBE. The origins and development of TBE has been extensively documented and discussed by Wildschut (2014). Ralph Tyler, Don Kirkpatrick, Edward Suchman and Carol Weiss are cited by various authors as having made the first contributions to TBE, publishing the first articles on TBE and being the first to make links between programme theory and evaluation. TBE has evolved over several decades and important contributions have been made by individuals from a wide range of disciplines, including evaluation (Ibid). Many terms are used to describe TBE and a range of tools to develop/document it have subsequently evolved, including programme theory, toc, logic models and outcome mapping etc.

In addition to being rooted evaluation, and specifically TBE, the toc "approach" is grounded in community development practice. James (2011, unpaginated) points out that: "people have long explored theories of social change, debating what leads to development and how that influences the approach organisations take… development organisations and practitioners are all, consciously or unconsciously, development theorists, drawing on macro theories of development as frameworks for action".

The Aspen Institute in the United States undertook pioneering work on toc, with the aim of improving the evaluation of complex community development programmes and published a book in 1995 which proposed New Approaches to Evaluating Comprehensive Community Initiatives. Carol Weiss, a member of the Aspen Institute's Steering Committee on Evaluation, later popularised the term "toc" as a means to describe the assumptions and steps required to achieve a long-term goal and connections between activities and outcomes which are required to achieve the expected changes (Anderson, 2004). The Aspen Institute joined forces with ActKnowledge to produce the first toc guidelines and these organisations have, since then, developed tools, user guides on how to develop and utilise a toc and published lessons learned from toc development (Anderson, 2004; James, 2011).

Methods of toc development

Although much has been published on toc and evaluation, publications often say little about how tocs are developed and who s involved (the latter is discussed later). Little has been written about toc development and evaluation in South Africa. A variety of methods can be used to document the implied toc underlying a programme/project and the process may range from highly participatory to evaluator driven (Coryn et al, 2011). Some examples of toc development which have been documented in other country contexts are discussed below.

Mason & Barnes (2007), writing about a National Evaluation of the Children's Fund in the United Kingdom, discuss a process which involved researchers and practitioners "co-con-

structing" a toc from various sources. Based on their experience and a brief literature review, they propose that data sources for developing a toc can include: advocacy (which can be based on experiential and/or research evidence), research findings, professional experience and the objectives and aims of funders. In the example discussed, the researcher/evaluators role was to facilitate discussion and subsequently construct a toc, which was checked with participants and revised based on their feedback. Mason & Barnes (op cite) believe that their process is compatible with the "collaborative approach" which underpins the concept of toc.

Describing an alternative approach - relating to an evaluation of Scottish Health Demonstration projects - Mackenzie & Blamely (2005) propose an evaluator-led process which builds on existing programme documents. The evaluator may produce a tentative toc for discussion or facilitate a meeting at which stakeholder views are summarised, critiquesd and iterated. Mackenzie & Blameley recommend that a variety of methods be used to develop the toc and a wide range of stakeholders involved, to ensure ownership.

Stein & Valters (2012) - who undertook a review of toc in international development - posit that, at its best, a plausible toc draws on social science theory, research findings, evidence of what works from previous programmes and local/contextual knowledge. The extent to which toc practitioners engage with these sources will reflect in "whether toc approaches ultimately reveal or oversimplify the complexity of processes of change".

However, Stein & Valters (op cite) remark that there is likely to be a tension between different (and often opposing) approaches and data sources when developing a toc. I.e. if a toc is based on practitioner experience, it may succeed in surfacing 'lived experience', but not relate to social science theory. Similarly, over reliance on theory may generate a toc that is theortically sound, but divorced from context and practical experience.

In closing, Stein & Valters (op cite), propose that a 'best practice' toc should be grounded in accurate analysis of the context (i.e. of the intervention and programme stakeholders), draw on a variety of forms of realistic evidence, consider the plausability of the programme goal and the extent to which it is realistic, and define the boundaries, scope and level of complexity required. This will ensure that the toc is useable and that the resources, expertise and externalities necessary for change are identified and put in place.

Van Stolk et al (2011, cited in Stein & Valters, 2012), note that a strong evidence base will enable a toc to capture "multiple dimensions of change", which can include social, economic and political dimensions. They propose that data be collected via anthropological or sociological methods and corroborated by triangulation. The robustness of toc data collection processes – i.e. the extent to which they are replicable, systematic, and considered legitimate by programme stakeholders – is considered to be particularly important (ibid).

Social science theory may be useful in contextualising and understanding the evidence basis for a toc. According to Eyeben et al (2009, cited in Vogel, 2012), social science theory - including political economy, rights-based approaches and power analysis etc – can be useful to clarify "key points in theory of change thinking", however, it is important to understand the history and circumstances of and critique theories, which are predominently rooted in

"western scientific and social science traditions" (Bronstein et al, 2009).

Several other toc practitioners and writers, including James (2011), Vogel, (2012) and Funnell & Rogers (2011) discuss the importance of triangulating data from multiple sources including: academic, evaluation and grey literature, practioners "mental models", information about the context, and stakeholder and project/programme experience.

The purpose of the toc development should be considered from the outset, to ensure that it is fit for purpose (the use of toc is discussed subsequently). A range of methods may be used and data sources drawn on to develop/document a toc. Most authors recommend that multiple methods and sources are used, to develop a stronger and richer toc, which integrates understanding of context, practical experience and research/social science theory. There may however be some conflict between different sources (this is discussed further below).

Factors which influence toc development

We have seen that there is no standard agreed upon methodology to develop a toc. Similarly, the components of a toc may vary, but will usually include a big picture goal, objectives, expected/intended outcomes and information on the steps required to achieve them, and information about the assumptions which inform the toc. Tocs may take different formats, from linear/simple to complex, and may or may not include a visual representation of change. Some of the factors which influence the development and formats of toc are discussed below.

Levels

Typically a toc is developed for a project or programme, but a toc can be developed at various other levels. Mason & Barnes (2007) make a distinction between toc's at programme design (national/policy) and implementation (local) levels. Similarly, Mackenzie & Blamey (2005) propose that toc's need to be articulated at both strategic and operational levels, and linkages between them should be made explicit. James, (2011, p7) explains that people think about toc at different levels: macro (development perspectives and thinking); sector; organisation; and project or programme. Most toc's focus on a specific level (usually project or programme), but may make links to other levels. This is illustrated visually below:

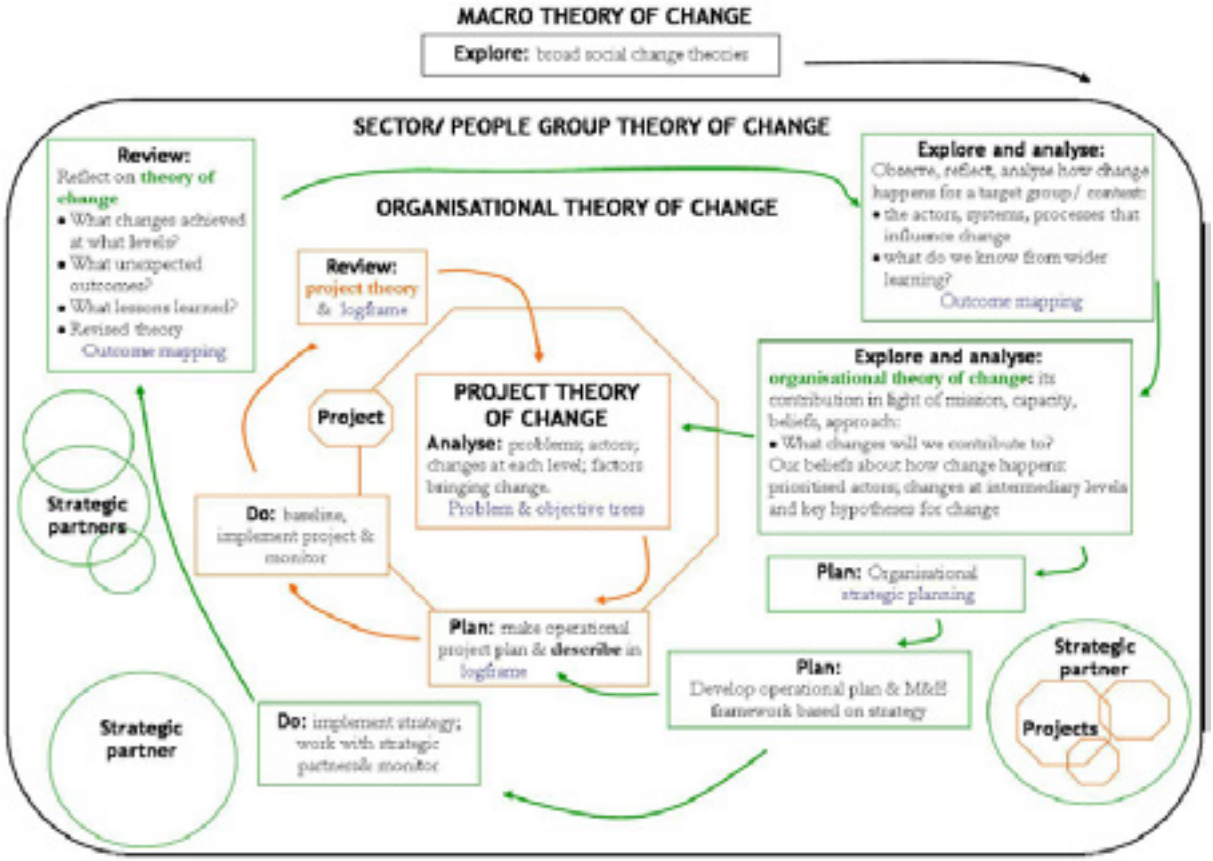


Figure 1: 'Levels' for theories of change, source: James, 2011,p26

Stein & Valters (2012, p8) explain that there are different levels of analysis when thinking about social change, from societal and conceptual to organisational and implementation. These levels are interconnected, but it is important to define the primary level which a toc is expected to function, to ensure that it is specific and clear. The concept of levels is also useful in considering how toc is used. For example, there may be implementation toc's (for a specific interventions) as well as an organisational toc which guides programming decisions.

The concept of levels is also useful in understanding who is targeted and what type of change is expected. Shapiro (cited in Stein & Valters, op cite, p9) explains:

- "Changing individuals involves strategies that shift attitudes and perceptions, feelings, behaviours and motivations of participants in an intervention."
- "Programmes that focus on changing relationships often suggest that new networks, coalitions, alliances and other cooperative relationships between members of conflicting groups not only positively change the individuals directly involved, but can be a powerful force for fostering social changes that help resolve conflicts. These meso-level change strategies aim to effect both individuals and social structures."
- "Structural, institutional and systemic changes are the primary focus for some conflict intervention programmes…These efforts are often directly aimed at legislative, electoral and judicial reform, establishing new mediating mechanisms and forums within society, economic development initiatives (e.g. micro-finance, job training) and infrastructure support for basic human necessities (e.g. water, food, health care)."

Timing

Another factor which influences toc development is the stage in the project life cycle when the toc is developed. For example, whether the toc is developed when a programme or project is being conceptualised, planned, when implementation is underway, or retrospectively (e.g. such as for a summative evaluation).

Mason & Barnes (op cite) explain that it may not be possible to accurately map the planned activities, expected outcomes and their contribution towards a goal/objective(s) during programme/project conceptualisation, as programmes and projects develop and may evolve over time. In our own experience, this is especially true of complex programmes which are characterised by "emergence" and "feedback loops" (see Funnell and Rogers). Developing a detailed toc may only be possible when implementation is underway, and the operational detail and causal links are revealed. Thus, a toc may be developed at the outset but should be reviewed and refined during implementation.

Similarly, Mackenzie & Blamey (op cite, p155-156) emphasise that it is important to know when to "draw a line" around an initial toc. There may be pressure initially to develop a detailed toc for the inception phase, but there will be opportunity later, to capture more nuanced linkages and unexpected outcomes associated with complexity. They advise that it is important to capture the changes from an initial to a final toc, so that understanding of what has occurred and why is based on what actually happened.

Stein & Valters (2012) propose that toc should be an iterative and evolving tool, "a set of theories relevant to a specific setting, which are articulated, tested, and improved over time".

James (2011) remarks that, whilst toc should be an on-going learning process, it can be very useful to have an initial "event" to kick-start the process and develop a draft toc. This could happen prospectively (i.e. at the beginning of a programme) or retrospectively (i.e. prior to a summative evaluation).

In relation to the project life cycle, Vogel (2012) notes that a toc is most effective when it is aligned to applied alongside existing planning processes. However, it is possible to effectively apply toc to an existing project/programme, if there are opportunities to align the toc approach to other processes and adjust and revise project/programme strategies.

Participants and power dynamics in the toc development process

Stakeholder participation is one of the tenets of the toc approach, but participation is costly – in terms of time and other resources – and power dynamics influence who participates and the effectiveness of their participation. The stakeholders involved, extent and quality of participation influence the type of toc which is developed.

Sullivan & Stewart (2006) describe a "typology" of toc ownership, from "total ownership" (the democratic ideal) to elite, principal, evaluator and community ownership. They conclude, from an analysis of how toc has been applied in the United Kingdom, that "total" and even "community" ownership" is very difficult to achieve as power relations tend to give rise to "elite" or "principal" toc ownership.

Power dynamics are an inherent - but often ignored or undocumented - undercurrent in the toc development process (Stein & Valters, 2012; Vogel, 2012). Certain stakeholders may be excluded from the process; for example, programme/project beneficiaries. Although it is often difficult to involve beneficiaries in toc development, Stein & Valters (op cite) propose that finding ways to "foreground their views" ? is important for strengthening toc's. Similarly, Van Stolk et al (op cite) emphasise the importance of grounding toc in "the perceptions and behaviour of local indidviduals and organisations" as their views are more relevant than those of external stakeholders.

Mason & Barnes (op cite) note that there may be tensions and differences in the toc's generated by different stakeholders which it is not be possible to resolve: "our approach is not to work to consensus… but to clarify whether there are any differences, and if so determine the reasons for these and consider the consequences".

Vogel (2012) finds that there are a range of implicit and explicit toc's which influence expectations and how projects/programmes are actually implemented, for example: organisational norms (which may or may not be documented); programme stakeholders and beneficiaries may have differing aims/goals/objectives for implementation; and toc's may be "imposed"by funders or other external stakeholders; programme implementers may then have different ways of acting and reacting to the toc's outlined above to further their own agendas. Vogel proposes that out-of-the box thinking be encouraged and different toc's be explored, as they can suggest different pathways to influence and achieve outcomes and thereby enrich programme strategies. In the case of donor funded development projects, there may be tensions between the toc's of donor organisations and the organisations and programmes/projects they fund (Stein & Valters, 2012).

Vogel (2012) notes that toc's are developed for different purposes (discussed further below). Thus, the toc developed for an evaluation will differ from that developed by an implementation team. It may be useful for an evaluator to have the toc developed by implementers - to test, critique or validate - but ultimately it is a subjective interpretation which was developed for a purpose other than evaluation.

Restrictions

The participants and power dynamics discussed above may restrict the scope and perspective of a toc. There can be other parameters. Mason & Barnes (2007) note that broad aims and possible activities for a programme/project may be set by funding parameters. Stein & Valters (2012) warn that desire to be "policy-relevant" may deter the investigation of alternative assumptions and outcome pathways. Weiss has cautioned that this may give rise to situations in which toc approaches allow us "to know more but understand less" (Weiss, 1995, cited in Stein & Valters, 2012).

As noted by Stein & Valters (op cite), developing a toc requires significant time and resources. This may be more difficult for small organisations, meaning that large organisations are more able and likely to engage with toc. Similarly, participation will be more costly for people not participating in a professional capacity.

In summary, tocs may be formulated at different levels - depending on the purpose - and it can be useful to identify the linkages between a toc's primary level of operation (e.g. project/ programme) and other (e.g. organisational and sector) levels. A range of toc's are possible, depending on standpoint, perspective, beliefs about social change and so forth. Who participates will determine whose perspective is documented and becomes the official, accepted toc. Toc processes are influenced by power relations and powerful interest groups can create barriers to dialogue. Thus, power relations can limit/restrict a toc process from challenging norms, but a well-facilitated toc process can encourage out-of-the box thinking.

Uses of toc in and beyond evaluation

The primary use of a toc by evaluators is for evaluation, but various process uses and uses of toc beyond evaluation are reported. The development of a toc leads to a document, which may be referred to as a "product", but the process of developing a toc is also valued. Furthermore, toc's may be developed for purposes other than evaluation; for example, to assist with programme conceptualisation and planning.

Stein & Valters (2012, p5-6) suggest that toc approaches can be viewed as falling along a continuum from technical – e.g. for use as a planning tool - to political – e.g. to explore how change occurs, to develop appropriate responses. "toc thinking" -understood as an informal way of thinking about how an intervention is expected to work - is positioned in the middle of the spectrum. In relation to the project/programme life cycle, they identify four main functions of toc: strategic planning; monitoring and evaluation; description and communication; and learning.

James (2011) has developed a typology of the purpose and use of toc, along several axes:

- Formative/evaluative - toc is used extensively by evaluators, but is also used in more formative ways, i.e. for learning;
- Explanatory/exploratory: – toc may be used to explain an existing theory, to test whether it works, but it can be utilised in a more exploratory manner, to construct and build theory;
- Linear/complex: toc may be used to develop a cause and effect chain and document steps that can be tested, or it may be used more systemically, to document the complexity of change, illustrate linkages between actors, processes and learning loops.

Vogel (2012) reports that some people view toc as a tool or methodology whereas others see it as a reflective process, she believes that, toc is "at its best" when it combines elements which reflect the strengths of both approaches: e.g. planning is strengthened by reflection about motivations and assumptions about how and why change occurs. She also notes that the uses of toc vary depending on the user. For example, implementation organisations may use toc for programme scoping and design, planning and results-based management. Donors and foundations may use toc for programme design and commissioning and to guide large-scale evaluations. Civil society organisations and NGOs may use toc to develop monitoring, evaluation and learning frameworks, and support empowerment by linking their work to broader forms of change. On the other hand, evaluators usually use toc to guide the monitoring and evaluation of programmes, from simple and small-scale to complex.

Drawing on the work of Connell and Kubisch, MacKenzie & Blamely (2005, op cite, p152) identify three key strengths of toc when compared to other evaluation approaches: 1) focus-

ing and sharpening programme planning; 2) enabling the prioritisation of evaluation questions and methods; and 3) reducing problems associated with attribution in the evaluation of complex interventions. In relation to evaluation use, several authors point out that toc is particularly useful to guide the evaluation of complex programmes, as it provides a basis for gathering evidence, checking possible alternative explanations for change, and providing an outline for explaining and inferring the changes which have occurred (White and Phillips, 2012).

It was noted that toc has roots in the evaluation and community development traditions and, perhaps due in part to this, toc is used in ways which span conceptualisation and planning, reflection and learning and monitoring and evaluation. These varying uses are not always aligned and integrated. The way in which toc is used appears to depend on how it is developed, by whom and when.

JET's recent experiences in toc development

The descriptions below reflect on JET's recent experiences in toc development for the evaluation of three South African Education programmes. A brief description of each programme is followed by a discussion on how the toc was developed and how it was used for and, where relevant, beyond evaluation.

Funza Lushaka Bursary Programme (FLBP) implementation evaluation

JET was commissioned by the Department of Planning, Monitoring and Evaluation (DPME) and the Department of Basic Education (DBE), to undertake an implementation evaluation of the Funza Lushaka Bursary Programme (FLBP). The FLBP provides full bursaries to academically deserving students undertaking initial teacher education (ITE). The evaluation was included in the 2013-2014 national evaluation plan. Development of a toc to guide the evaluation - based on a programme goal and objectives outlined in the evaluation Terms of Reference (ToRs) - was one of the deliverables required.

The FLBP commenced in 2007 and the evaluation was commissioned in 2014, so the toc was developed retrospectively. The programme evolved and there were a number of changes to the goal, objectives and implementation business processes between 2007 and 2014, which were documented as part of the toc development process.

Prior to developing the toc, the evaluation team undertook a literature review - which included an international and national review of other bursary programmes - reviewed programme documents and conducted interviews with key personnel involved in conceptualising, designing and refining the programme. Following this, the evaluation team developed a draft toc and logframe which was discussed at a stakeholder workshop.

The toc workshop, held in August 2014, was a key event which involved 60+ key stakeholders including national and provincial government officials (including delegates from all nine provinces), Higher Education Institution staff including: Deans of Education, Academic Coordinators and Financial Administrators, and representatives of the National Student Financial Aid System (NSFAS) responsible for disbursing FLBP funds. The workshop was facilitated by JET and DPME staff. Participants worked in small groups which fed back to plenary, and

identified the goal and objectives of the FLBP, major events in the development and evolution of the programme, discussed strengths and weaknesses of the FLBP business processes and confirmed key logframe components (assumptions, outcomes, activities and outputs) relating to the programme objectives. The draft toc and logframe for the FLBP developed prior to the workshop was revised by the evaluation team after the workshop, then reviewed and approved by the evaluation steering committee.

A range of participatory methods were used at workshop, drawing on the traditions of Participatory Rural Appraisal (PRA), advanced by Robert Chambers, and utilised in evaluation by Irene Gujit and others. These included: participatory timeline mapping; participatory voting with stickers to identify successes and blockages in business processes; and "backwards mapping" to trace the linkages between activities, outputs, outcomes and objectives and identify assumptions at each stage. Participants worked in small groups – comprising a range of different "types" of delegates - supported by a facilitator from JET or the DPME. The activities and workshop processes were highly engaging, they tapped into participants "expert" institutional knowledge and participants left the workshop feeling that they had made valuable contributions. The hands-on, practical nature of the engagement created a non-hierarchical environment in which delegates of varying designations were able to contribute on an equal footing.

The workshop provided a very valuable source of data for the FLBP toc development, but ultimately, the official, "approved", toc was developed by the evaluation team and approved by the evaluation steering committee, in accordance with DPME processes outlined in the National Evaluation Framework. In this regard, the evaluation team experienced some tension between between developing a toc which was "technically sound" and in line with good practice guidelines and a toc which accurately captured participants' experiences and reported realities.

Further complicating factors were that the implementation evaluation which was comissioned covered the period 2007-2012, whereas the toc was developed for the programme in August 2014, at the time of the evaluation. Thus the toc was based on the present and was forward-looking, whereas the evaluation was retrospective. Additionally, the FLBP is a complex and complicated programme (see Funnell and Rogers, op cite). Some of the FLBP business processes are implemented differently in different provinces. The toc had to make accommodation for this.

The workshop stimulated rich discussion and debate about challenges with and proposed improvements to the programme design and considerations for the future direction of the programme. DBE delegates attending the workshop said these would be carried forward to future strategic planning meetings. Discussion and debate also took place between the evaluation team and the evaluation committee, regarding the scope of the programme (and thus the toc) and extent to which factors which the programme did not have direct control over should be included in the toc (e.g. how provincial education departments place FLBP teacher education graduates after they have completed their studies).

The FLBP evaluation is in the process of being finalised, and the evaluation team have not followed up to ascertain the extent to which the toc has/is being used beyond the evaluation. However, an "improvement" workshop" - based on the findings and recommendations of the

evaluation - took place in August, following the process as outlined in the National Evaluation Policy Framework.

BSSIP summative evaluation

The BSSIP was designed and implemented by JET's School and District Improvement Division (SDID) in line with JET's systemic school improvement model. The intervention – which commenced in 2009, was redesigned in 2012 and ran until August 2014 – was funded by the Michael and Susan Dell Foundation. The Monitoring and Evaluation Division at JET was commissioned to conduct a summative evaluation which would assess the extent to which the programme was implemented as intended and achieved its objectives.

JET's systemic school improvement model is centred on educational outcomes, effective school organisation, community involvement and district support. The underlying assumptions are that educational outcomes at school level will improve if teachers are effective, and teaching and learning environments are supported by effective school organisation and community involvement.

The school improvement model that guided implementation of the first phase of the programme (between 2009 and 2012) focused on seven components, as highlighted in Figure 2. The various components were geared towards improved support and monitoring of schools by districts, increased community involvement, improved functionality of schools, enhanced teacher competence and performance, and improved learning and educational outcomes.

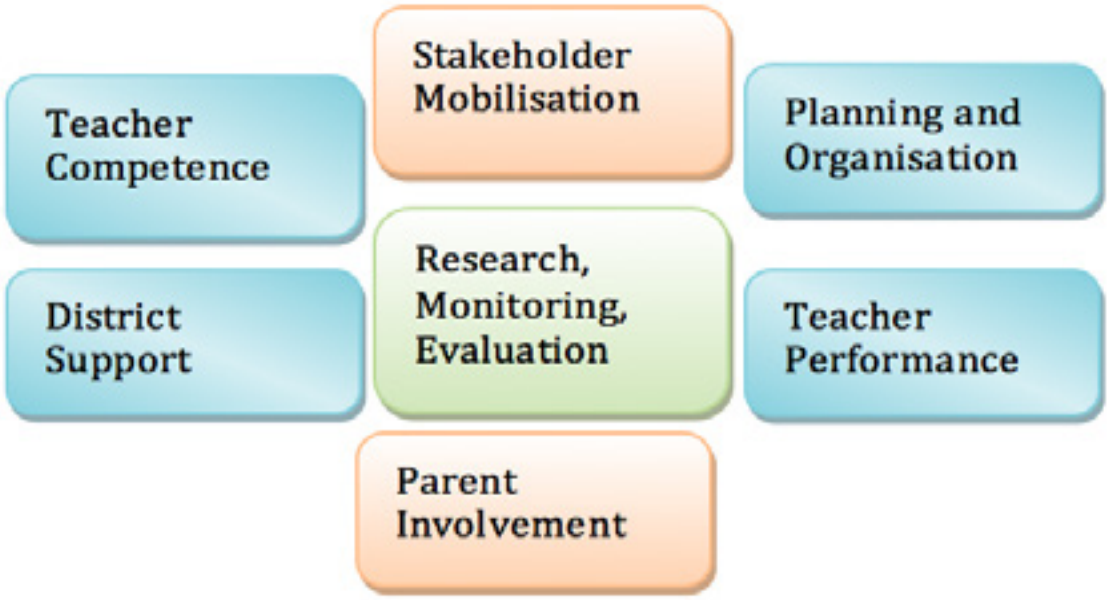


Figure 2: JET's Systemic School Improvement Model, Source: JET

In 2012 the intervention was redesigned, with the aim of increasing impact and enhancing the likelihood of overall improvements in teaching and learning. The new design focused more specifically on addressing needs within school and community settings. From 2012, the project focused on improving curriculum management at school level, through improved

planning and organisation and enhanced parent and community involvement.
The BSSIP was designed to improve learning outcomes in the Bojanala District, North West Province. The aim was to implement two key components: 1) planning and organisation and 2) parent and community involvement. The programme aimed to improve the quality of education through interventions in 25 schools in 15 villages (17 primary schools, three middle schools and five high schools).
On the one hand, the planning and organisation component of the programme aimed to improve school functionality, by improving curriculum management. This component focused primarily on imparting curriculum management skills and techniques to the main stakeholders responsible for curriculum delivery.

On the other hand, the parent and community component aimed, firstly, to make parents and community members aware of their roles in the education of their children. Secondly, this component aimed to establish a structure within each community to lead the process of identifying relevant projects and interventions that would address the needs of the schools and to plan, implement and monitor such projects. Thirdly, the parent and community component aimed to create opportunities for structures from different villages to meet frequently and share best practices and learn from each other.
By meeting these short term goals, the parent and community component aimed to address the long term goal of improving learner performance. Collectively, both programme components aimed to increase learner performance and overall school functionality.
At the start of the BSSIP, programme theory was depicted in the form of a logical framework which was used to guide implementation. This informed the toc which was developed for the evaluation. Though the BSSIP had an explicit programme theory that was used to guide implementation, it has often been found that interventions may have a toc, but that this toc is not always plausible (Rossi et al, 2004). Therefore, a plausibility assessment was conducted through an analysis of relevant literature. Additionally, through an analysis of programme documents and interviews with key stakeholders, a model describing the toc underpinning the intervention and the expected outcomes of the programme was developed.
Based on the literature review, it was evident that the BSSIP was a holistic school improvement intervention that aimed at addressing school functionally, by mediating the variables that affect curriculum management and parent and community involvement. As interventions by role players external to the school environment can play a crucial role in improving the teaching and learning environment in schools, the BSSIP was assumed to have fulfilled such a role. Literature confirmed that the BSSIP aimed to address the challenges faced in curriculum management and delivery and delivered curriculum improvement strategies which could contribute to the effective management of schools. It was therefore suggested that the BSSIP programme could create a platform where effective curriculum management could be explored within the intervention schools.

The literature review led to the conclusion that the BSSIP toc was plausible. Hence, a review of programme documentation and discussions with programme stakeholders led to development of a depiction of the programme toc to assist in the formulation of evaluation questions and guide and support the evaluation. This evaluator driven toc was formulated in the form of a variable oriented diagram which highlighted each programme component, the activities to be implemented under each component and the short, medium and long term outcomes

that were expected to occur as a result.
The toc diagram that was developed made it easier to identify what activities were crucial to implement to achieve the intended outcomes. Additionally, the toc assisted in formulating questions to guide the evaluation and in assessing the programme's results against the intended outcomes.
The toc developed by the evaluation team and the evaluation findings were presented at a workshop where lively debate took place regarding the success of the project, implementation challenges faced, and what it is possible and plausible to expect a project of this nature to achieve.
The BSSIP toc case study illustrates the supportive and instrumental role that a toc can play, not only in guiding an evaluation, but also in creating a platform for stakeholders to have interactive discussion on areas in which programme design can be improved upon in future.

Department of Higher Education and Training (DHET) College Improvement Project (CIP) formative and summative evaluation

The DHET funded CIP was implemented in all 15 Further Education and Training (FET) colleges – now referred to as Technical and Vocational Education and Training (TVET) colleges - in the Eastern Cape and Limpopo provinces between 2011 and 2014. The project's overarching goal was, to demonstrate an effective systemic model of FET College Improvement that can be replicated throughout the college sector and the success indicators for the project were: enrolment growth, certification rate, throughput rate and placement of students in Work Based Experience (WBE).

Given this goal, substantial effort was placed on learning, monitoring and evaluation (LME) to ensure that the toc could be tested. The evaluation therefore focused on understanding the extent to which the toc (the model) was sufficiently robust and could be used in other contexts. Thus the LME processes focused on understanding whether the intended activities were implemented as planned, and where not, to understand the rationale for the changes and explored whether this led to the intended goal and success indicators (and again if not, why not and the implications of this for the toc).

The project was implemented by JET and included a team at JET head office, which managed the project and provided support to two Provincial teams. The head office support and provincial teams were organised into function teams: these were the teams responsible for a cluster of outcomes such as teaching and learning, Education Management Information Systems (EMIS) etc. In addition, there was an internal and external LME team that supported the project and initiated monitoring and evaluation activities to ensure that the project was documented and that there was credible and valid evidence about the value of the model.

The project began by undertaking rapid assessments in both Provinces; the findings from the rapid assessment enabled the JET team to prioritise the interventions that would be implemented during the programme.

The LME work included: baseline studies, drawing on the rapid assessment results; development of a toc (discussed below); a survey conducted with students, lecturers and managers in all colleges; a formative evaluation involving an Outcome to Purpose Review (OPR); stake-

holder interviews and case studies; and a summative evaluative process which involved re-application of the survey to establish perceptions of change, coupled with more in-depth interviews to explore the factors that had enabled change and specifically to determine the extent to which the project had contributed to change.

Once the initial programme planning was undertaken, a workshop was held with the full project team to explore the toc; this focused on shaping the goal, objectives and success indicators of the programme (as outlined above). It was understood that these would be used to ascertain the success of the project. The workshop then considered the outcomes that would need to be achieved to realise the goal and objectives and explored the extent to which the planned activities would be sufficient to achieve the desired outcomes; this allowed for a reflective process on the planned activities and a review of whether the outcomes would be realistic, and if not, whether the goal/objectives could still be achieved.

After this workshop there was an iterative process with the functional teams (as explained above), which refined the activities and expected outcomes and also agreed on indicators for each of the outcomes. The outputs of these processes were then integrated and the revised toc was circulated for comment, and once agreed, became the reference document for the evaluative work that was undertaken.

In developing the toc it was understood that, given the complex nature of the project, certain activities may need to be added and others might become less important, given learning acquired during implementation and as a result of contextual changes within the TVET landscape. However, it was agreed that the goal and success indicators would remain constant and therefore the focus would be on understand what is required to achieve the goal and indicators.

During the various evaluative activities, it was found that many of the activities planned for the project had been changed during the course of implementation, based on revisions to the project scope, learning through implementation and contextual changes. There was however no on-going process to revise the toc to reflect these changes, and as a result, at the point of the summative evaluation, it was considerably harder to determine whether the achievement of the goal (to the degree that it was achieved) provided evidence that the toc was credible and that this programme had demonstrated a model that could be used elsewhere. This weakness appeared to relate to the reality that the toc process, rather than informing the project planning processes, ran in parallel to these. This meant that despite efforts to develop the toc from the outset, the reality was that the toc had to be developed – as of from new – at the conclusion of the project, so as to enable the project to retrospectively indicate what needed to be done to achieve the goal and objectives of the programme.

Despite this limitation, the toc developed at the end of the project provides a model for how such initiatives can be implemented, and as such, is a useful way of documenting the project so that it can be replicated. Further, whilst the parallel toc and planning processes were less than ideal, the work undertaken with project teams in both provinces created an opportunity for the teams to reflect on their initial assumptions and to explore where and why these were wrong. This process allowed for a far more robust toc to be developed at project close out,

with a clearer understanding of what needed to be in place for the activities to take place as planned and therefore for the objectives and ultimately the project goal to be achieved.

Discussion and conclusion

The case studies above provide some lessons and guidance regarding how toc development can improve planning processes, be used to find common ground and build consensus, stimulate dialogue – with a view to strengthening and improving programme design and implementation – and may increase the ownership and possible utilisation of evaluation findings. However, these are not straightforward nor guaranteed to achieve.

The FLBP evaluation illustrates how a participatory toc process can bring a diverse range of stakeholders together to discuss strengths and weaknesses and possible improvements and map the toc. The quality of the facilitation affects the success of the process and the end result. Power dynamics need to be managed, to ensure that people/voices are not marginalised. Important issues need to be raised and areas of disagreement discussed and debated – bearing in mind that it may not be possible to achieve consensus, and that different viewpoints should be documented.

It is important that there are mechanisms for key issues raised by the toc process to be taken forward as appropriate in other fora, i.e. at review and strategic planning meetings. This can be achieved by involving people driving these processes in the toc development and planning how the toc and the evaluation findings will be further used beyond the evaluation. But ultimately, there may be other factors (beyond the control of the evaluation) influencing this.

The FLBP toc process involved 60+ key stakeholders, notably government officials and various representatives of higher education institutions, but beneficiaries (i.e. students) were not involved (although they were extensively involved in other aspects of the evaluation). Who participated was determined by the evaluation steering committee. With complex and complicated programmes involving a range of role players, participation in toc development - which can be time and resource intensive - must be carefully planned. There may be, to a certain extent, a trade off between breadth and depth of participation as involving 60+ stakeholders is time consuming, and discussions/discussions may not be as in-depth as they would be with a smaller group, but a broader range of perspectives are incorporated. Participatory methods/ techniques suitable for use with large groups should be used and sessions facilitated by sufficient, adequately skilled facilitators, programme stakeholders can be supported to play a facilitation role.

Although the BSSIP toc process was largely evaluator driven, the toc used in the evaluation build on the initial logical framework used for programme implementation. The toc model made it easier for the evaluation team and programme stakeholders to clarify what short, medium and long term outcomes the programme had aimed to achieve and the means through which these were to be achieved. The toc that was developed assisted in identifying key evaluation questions and provided a basis against which to analyse the evaluation findings. The toc developed was used to discuss what was intended to be achieved and what was actually achieved. The toc model made it easier for the stakeholders involved to identify what gaps needed to be addressed in the possibility that a similar programme was to be replicated. How-

ever, issues raised during these discussions need to be taken forward, to influence review and strategic planning processes (which JET has embarked on in 2015).
The toc process followed for the DHET CIP also points to the value of a toc in assisting to articulate what a project intends to achieve and how it intends to realise this. This can facilitate testing of internal logic and reflection on whether the proposed activities can realistically achieve this goal, with the purpose of either renegotiating the goal from the outset - such that all players have a shared understanding of what can be achieved - or alternatively, reallocating resources to ensure that it is possible for the activities to be implemented to achieve the intended goal, taking cognisance of the assumptions that are made in this regard. This has the potential to create a basis for role players to understand what the project will do, what it is intended to achieve and what this assumes from all role players.

However, it may not be possible to develop a fully fledged toc at the outset of a complex programme (as in the case of the DHET CIP), as there are certain "unknowns" and the project may evolve as these become "known" and the landscape changes (as it did). Furthermore, the reality of this toc process is that it was only developed by the JET project team, there was no opportunity to engage with the DHET (the project funder and ultimate "owner") about the toc, nor to reach a shared understand of the assumptions underpinning the toc and test whether they were felt to be realistic. It was found that DHET did not have the dedicated capacity which the project logic anticipated, and could not deliver against all of the assumptions, and certain key activities could not be implemented by except by the DHET. Further, as the toc was not treated as a living document, there was no space created during the project for stakeholders to reflect on the toc and to consider the implications for the project and the roles that were expected and required by each role player. This may have been a lost opportunity, in terms of ensuring 'ownership' of the evaluation and of the project.

More positively, the ToC did ensure that learning was documented and, as a result, there is evidence about what is required to achieve the intended goal and objectives and related success indicators. This is crucial for determining how the project could be adapted and replicated and for informing learning about what it takes to achieve this aim.

The literature and case studies presented above indicate that a broad range of methods can be used to generate evidence and develop a toc, the particular methods used can be determined by specific needs/requirements and influenced by the innovation/creativity of the person/ team responsible for facilitating the process (often an evaluator).

There is value in involving a broad range of programme stakeholders in toc development in some way, as they are more like to feel "ownership" of the process. Following on from this, the toc which is developed may be more likely to be used beyond evaluation. However, the latter point is a conjecture which was not proved by the literature or the case studies. If certain key stakeholders/groups are left out of the process, there may be a "gap" in the logic of the toc, it may be "unimplementable" and/or role players may not be fully aware of or agree to their respective responsibilities.

All of the case studies point to the need for a toc to be developed at the outset, to guide implementation, but illustrate that the toc should be revisited, revised and updated. Changes to

the project/programme design and implementation should be documented for information/ communication purposes and because they will be required for evaluation. It can be useful to have a practical toc for implementation and a more detailed toc for evaluation purposes. Consensus on a toc is not always possible – or recommended - and it is useful to document where and how stakeholders disagree.

According to our experiences, toc development can be a catalyst for dialogue about successes, challenges and potential improvements, but who participates is critical (i.e. the need to have all key role player and stakeholder groups present) and there must be mechanisms for carrying issues from these discussions forward, into review and planning processes. This points to a need for project/programme planning and evaluation processes to be clearly aligned.

We have touched on ways in which we believe toc development (and specifically participatory processes for the development thereof) can increase ownership and we believe could increase the utilisation of evaluation findings in South Africa. However, the latter has not been empirically tested. This would be a fruitful area for further study and empirical research.

References

Anderson, A, 2005, 'The Community Builder's Approach to Theory of Change: A Practical Guide to Theory Development', New York: Aspen Institute Roundtable on Community Change: http://www.aspeninstitute.org/sites/default/files/content/docs/roundtable%20on%20community%20change/rcccommbuildersapproach.pdf

Coryn, C, Noakes, L, Westine, C and Schroter, D, 2011, "A systemic review of theory-driven evaluation practice from 1990 to 2009", p199-226 in American Journal of Evaluation, 32(2), Sage Publications.

Eyben, R, Kidder, T, Rowlands, J, and Bronstein, A, 2008, 'Thinking about change for development practice: a case study from Oxfam GB' Development in Practice 18(2), p201-212.

Forss, K, Rebien, C, and Carlsson, J, 2002, 'Process use of evaluations types of use that precede lessons learned and feedback', Evaluation, 8(1), 29-45.

Funnell, S and Rogers, P, 'Purposeful Programme Theory: Effective Uses of Theories of Change and Logic Models', San Francisco, CA, Jossey-Bass.

James, C, 2011, 'Theory of Change Review. A report commissioned by Comic Relief', Comic Relief:
http://mande.co.uk/2012/uncategorized/comic-relief-theory-of-change-review/

JET, 2015, http://jet.org.za/about-jet

Mackenzie, M, and Blamely, A, 2005, 'The Practice and the Theory: Lessons from the Application of a Theories of Change Approach, Evaluation, 11(2), 151-168.

Mason, P, and Barnes, M, 2007, 'Constructing Theories of Change: Methods and Sources', Evaluation, 13(2), 151-170.

Mouton, C, 2010, 'The History of Programme Evaluation in South Africa', Masters Thesis, University of Stellenbosch.

Robbins, 2001, 'Education Pathfinders: A short History of JET Education Services', JET, Johannesburg.

Rossi, H, Lipsey, W, and Freeman, H, 2004, 'Evaluation: a Systemic Approach, 7th edition, USA, Sage Publications.

Shapiro, I, 2006, 'Extending the Framework of Inquiry : Theories of Change in Conflict Interventions', Berghof Handbook, (5).

Stein D, and Valters, C, 2012, 'Understanding Theory of Change in International Development', Justice and Security Researcher Paper 1, London.

Sullivan, H and Stewart, M, 2007, 'Who owns the theory of change?' Evaluation, 12(2), 179-199.

Van Stolk, C, Ling, T, and Reding, A, 2001, 'Monitoring and Evaluation in Stabilisation Interventions: Reviewing the state of the art and suggesting ways forward', RAND Europe, prepared for DFID Stabilisation Unit.

Vogel, I, 2012, 'Review of the Use of Theory of Change in International Development', Review Report, DFID, UK.

Weiss, C, 1995, 'Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families, in Connell, J, Kubisch, A, Schoor, L and Weiss, C (eds), 'New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts, New York: Aspen Institute (65-92).

White, H, and Phillips, D, 2012, 'Addressing attribution of cause and effect in small impact evaluations: towards an integrated framework', Working Paper, London and Cairo: 3ie.

Wildschut, L, 2014, 'Theory-Based Evaluation, Logic Modelling and the Experience of SA Non Governmental Organisations', doctoral thesis, University of Stellenbosch, Stellenbosch.

Zauyamakando, T, 2012, 'Review of the Use of Theory of Change in JET Evaluations', unpublished Power Point Presentation.

Turning a Mess into a Message: Evaluating Complex Programmes Using the Keystone Node Approach

First author
Gordon Freer (University of the Witwatersrand)

Second author
Sebastian Lemire (UCLA) [1]

Abstract

A central component of theory-based evaluation is to understand the assumptions of a programme's theory of change (Chen and Rossi, 1992)[2]. However, given the complexities of many developmental programmes, limited evaluation resources are often stretched to provide adequate programme coverage. In developing an evaluation framework for a five year M4P (Making Markets Work for the Poor) DFID funded programme in Ghana, the authors sought to overcome this challenge, and in doing so developed the keystone node approach. This approach centres on the examination of keystone nodes within the theory of change (ToC). Like the keystone in a stone arch, if a theory of change keystone node fails, the theory of change crumbles at that point. As such, each node contains critical assumptions or steps of logic that are vital to the intervention theory progressing from one stage to the next. By focusing the evaluation on 'keystone nodes', evaluation efforts target the most learning-rich steps within the theory of change, allowing for a better understanding of complex, multi-year programmes.

[1] The authors are part of a wider evaluation team under WYG, contracted to conduct an independent evaluation of a DFID funded market development programme in Ghana. The authors write in their personal capacities and do not purport to represent the interests of any of the organisations mentioned.

[2] Terminology surrounding theory-based evaluation is rich and diverse. For the purpose of this paper we prefer the term "theory of change" to refer to the underlying theory of a programme. We understand that some would prefer the term "programme theory".

Introduction and context of theory based evaluation

Theory-based evaluations are here, there and everywhere. The vast and growing volume of journal publications, conference presentations, book titles, and conference strands dedicated to the topic speaks to the persistent interest in theory-based evaluation (Chen, 2005). The momentum is still building. As just one example, and of special relevance for the present paper, a recent working paper commissioned by the Department for International Development (DFID) called for broader use of theory-based evaluations in development aid (Stern et al., 2012). Theory-based evaluation continues to be one of the most developed and applied evaluation approaches.

Theory-based evaluations are great – in theory. In practice they often present evaluators with both promises and pitfalls. On a promising note, theory-based evaluations provide an opportunity to explore new territory and to examine how and in what way programmes work (or fail to work). This type of information is increasingly what evaluations are asked to provide because it allows for a deeper and more robust understanding of programme performance, which can in turn inform future design (Mayne, 2012; Dybdal, Nielsen, & Lemire, 2011). In this way, theory-based evaluations also present an opportunity to recognize errors, in either programme planning or implementation, or both, whereby learning from programme errors can take place. These are just some of the many promising features of theory-based evaluation.

Pitfalls exist as well – three of which are worth highlighting. The first relates to the design of the actual ToC. In many cases the ToCs employed in theory-based evaluations are often watered down or overly simplified, failing to capture and make explicit key assumptions, mechanisms and influencing factors (Coryn, Noakes, Westine, & Schröter, 2011). We suspect these bare-boned ToCs exist for many reasons. However, at least in part, they emerge from the common conflation of ToCs and logic models, both of which tend to be merged into a singular model, mingling different component parts to form a homogenous jumble. The image of melted Neapolitan ice-cream may come to mind. [3]

Another challenge revolves around the strong potential for confirmation bias when empirically testing the ToC. This bias may stem – at least in part from the programme implementer and funder bias towards reporting only on the positive, to show what worked, and how well it worked.

These pitfalls – fundamental to the application of theory-based evaluations – are what motivates this paper. In response, the paper promotes the keystone node approach – an analytical strategy4 to support theory-based evaluations.

---

[3] For a distinction between the two, readers are directed to Patton, 2015. "Logic models are first and foremost descriptive, that is describing the steps of a programme from intake through completion. A well conceptualised theory of change model is explanatory and predictive."

[4] We consider the approach an analytical strategy because it specifies (1) the focus and scope of the type of data to be collected as well as (2) how to connect these different types of data.

The presentation of the keystone node approach is structured around three chapters. In chapter one, we examine what we take to be three persistent and pesky pitfalls of theory- based evaluation. These pitfalls are what motivates the approach. We then consider the practical context within which the development of the approach is embedded – a theory-based evaluation of an M4P market development programme (Ghana MADE). Informed by this examination, we then outline the core steps of the keystone node approach, awarding particular attention to our experiential learning from each of the completed steps. A brief set of concluding thoughts bring the paper to a close. Please note that the developmental aspect of the design of the approach needs to be emphasised. The process described in this paper is very much a work in progress.

Three pesky pitfalls of theory-based evaluation

A central pitfall of theory-based evaluation relates to the design of the actual ToC. In many cases the theory of change which awaits an evaluator is a watered down or simplified ToC. Painting in broad strokes, a theory of change should outline the intended progression from programme or project activities and products to a desired set of outcomes and long term impact. However, in the authors' experience, the level of detail – or explanatory granularity – differs markedly across theories of change and is more often than not too low. As noted above, there are several reasons for this. From our perspective, one central reason stems from all too common conflation of ToCs and logic models. The end result, while displaying ToC thinking and complying with grant or funding requirements, allows for limited appreciation of underlying assumptions and mechanisms. As Stern et al. comments:

"There are weak and strong depictions of a programme's theory. In weak forms it can be little more than a logic model that expresses the intentions of policy makers ignoring the actions and intentions of other stakeholders; and not making clear the assumptions that are made about the conditions under which the programme success is assumed." (Stern, et al., 2012)

In some cases the presented ToC is also an amalgamation of who, what, where and when the programme aims to achieve its impact dusted with a sprinkling of ToC how and why. This combination and often confusion of a logic model and a ToC struggles to serve the two different purposes of these instruments. The logic model describes processes, often moving from one to another, that while logical, sometimes require a leap of faith, especially by those tempered in the reality of real world development. The ToC – at least the better ones - incorporate, explore and explain the assumptions behind these leaps of faith.

The logic model used to guide the development and implementation of programme activities often regard the projects as isolated and controlled systems (Valters, Theories of Change: Time for a radical approach to learning in development, 2015). The theories of change, on the other hand, identify the assumptions, taking into consideration a broader context, often beyond the control or even the influence of the project. "Assumptions represent the values, beliefs, norms and ideological perspectives, both personal and professional, that inform the interpretations that teams and stakeholders bring to bear on a programme." (Vogel, 2012). It is these ToCs that link the project's anticipated outcomes to the real world in which the outcomes will be achieved and encourage a flexible, adaptive response to unanticipated or unplanned diversions or interruptions, recognising that the development programme works in an environment which is in a constant state of change. It should be the ToC that drives a learning culture within the project, to document learning around the assump-

tions, the activities and the challenges of the project.
A consequence of working with a diluted or fused ToC is that the resultant evaluations have limited explanatory strength about how the programme works, which is the foundational rationale of the ToC – to better understand how and why change happens. The consequences of watered down ToCs are severe. As such, the issue has not gone unnoticed by evaluation practitioners. Informed by their extensive experience with ToCs, Delahais and Toulemonde call for a more narrow focus in theory-based evaluations, advising evaluators to focus their evaluative inquiry on select aspects of the ToC – as opposed to the ToC as a whole (Delahais & Toulemonde, 2012). In line with this thinking, the key node approach focuses on specific aspects in the ToC that holds the greatest promise for learning – a point we will elaborate on later in this paper.

A second pitfall of theory based evaluation is that of confirmation bias within the testing of the ToC. As Guijt reminds us, "While learning and accountability are not necessarily in tension, 'official policies that profess the importance of learning are often contradicted by bureaucratic protocols and accounting systems which demand proof of results against pre-set targets' (Guijt, 2010)." (Valters, Theories of Change: Time for a radical approach to learning in development, 2015) The imbalanced focus on "what works" (or even worse on "what works well" and the idea of "best practice") has become a mantra within the relationship between evaluators and programme and donor staff. As noted by Valters, "The tendency to view a Theory of Change as predominantly an upward accountability mechanism considerably constrains attempts to learn form the process." (2014). For better or for worse this mind-set has left the "what does not work" question out of focus, often wilfully ignoring it completely. This is unfortunate because potentially important learning is lost, forgotten lessons of what not to do, dumped in the waste paper basket, leading up to a mid-term review. We simply fail to learn from error.

More than that, even if learnings have taken place and have been recorded, (taking care of the second shortcoming outline above), the record of our productive errors might be buried deep within a report or an appendix accessible only to those brave enough to venture into the depths of the publication. What of the programme staff, the staff of organisations with which the programme partnered, the communities and individuals on the ground without whom the programme would have no purpose? To what extent does this lack of transparency build capacity of the staff for future programmes, to learn from their success, their challenges, their obstacles and their failures? How do partner organisations improve their own contribution to their communities? And surely shielding the individual and communities themselves from the results is a fundamental shortcoming of our attempts at "development"?

The keystone node approach that we are in the process of unpacking speaks to this issue by choosing to focus on areas where change is imperative for the project to work, and if change does not happen the programme either fails or is forced to adopt a "work-around" to which we as evaluators will be privy. In examining these focal points we rely on analytical techniques that promote a better understanding of how and in what ways programmes work or fail to work, how and in what way we might learn from our errors. We know they will emerge.
The third and final pitfall relating to theory-based evaluations is that of the infamous "analytical black box" (Lemire, Nielsen, & Dybdal, 2012). The lack of access to this sacred place,

open only to evaluators (and sometimes fiercely guarded by the same), often results in a lack of transparency on how causal statements are generated and justified in theory-based evaluations. This lack of transparency cuts against the grain of social science and leaves all but the evaluator in the dark as to how the causal conclusions were drawn. This is unsatisfactory. Towards opening the analytical black box of theory-based evaluation, the keystone node approach relies on a set of analytical techniques that are both operational and feasible, allowing for more transparency of how conclusions were drawn.

The context for the Keystone Node Approach – the evaluation of Ghana MADE
The development of the Keystone Node Approach emerges from the authors' experience of designing a theory-based evaluation for a five-year M4P (Making Markets Work for the Poor) programme in Ghana. Both the programme and its evaluation are funded by DFID. In specifying the terms of the evaluation, DFID called for a theory-based, mixed method evaluation of the programme which would contribute to the international debate on M4P programmes as well as to the validity of these programmes within "thin markets". Towards these aims, a central purpose of the evaluation is to promote robust knowledge about how and in what way M4P programmes reduce poverty among poor rural farmers. As such, the evaluation approach needed to be able to tell a story about how, for whom, and under what circumstances the programme brings about change (or fails to bring about change). Stated differently, the evaluation had to be able to establish and explain the causal mechanisms connecting programme activities and observed outcomes.

Three conditions inform the evaluation. First and foremost, the programme is expected to develop over the course of the implementation period, reflected by changing markets and resulting in modified programme activities. More than that, it is expected that the programme will adapt to local conditions and in effect violate the demand for uniform implementation. The evaluation approach therefore needed to reflect these adaptive changes.
Secondly, the dynamic nature of the market and the fact that the programme is designed to be responsive to market needs, forces the evaluation design to be flexible and responsive as well. At the root of the programme approach is a set of principles for pro-poor market development. The implementation of these principles varies dramatically from programme to programme and the purpose of the present evaluation is to determine how and why specific programme components generate (or fail to) systemic change. These components are represented as theories of change, depicting the intended way in which programme activities will lead to positive change for targeted beneficiaries.

Informed by these observations, a theory-based evaluation approach was developed with the core idea of building an evaluation around the development and evolution of a theory of change. More specifically, ToCs were developed for the programme overall (Ghana MADE) as well as the individual market sectors (e.g., rice, livestock, etc.). These ToC were developed at the programme design stage as part of sequence of workshops with the programme implementation team.

In developing the evaluation, and informed by the common pitfalls cited above, special attention was awarded the need to counter ToC dilution, confirmation bias, and analytical murkiness. The approach – which we term the Keystone Node approach – directs the evaluative

gaze on a select set of assumptions and contextual factors, or keystone nodes that are pivotal to the theory of change of the designed market intervention. Like the keystone in a stone arch, if a theory of change keystone node fails, the theory of change crumbles at that point. Let us consider the approach in more detail.

The Keystone Node Approach

Four steps comprise the Keystone Node Approach:

1. Identify the keystone nodes in the theory of change
2. Determine crossover significance for each keystone node
3. Collect data
4. Conduct cross-case analysis

These deceptively simple steps will be elaborated in what follows.

Step 1: Identifying the Keystone nodes.

The first step of the keystone node approach is to determine the nodes to be examined as part of the evaluation. In the present evaluation, the evaluation team adopted a "bottom up" approach in determining the keystone nodes, starting with the interrogation of market specific theories of change and understanding the logic and rationale for the proposed market interventions. This interrogation took place both as individuals, as an evaluation team and in a workshop setting with the full programme implementation team, many of whom are familiar with the daily life that the programme seeks to impact. This inclusive process (although it did not involve partner organisations or targeted communities) ensured that the discussion centring on the theory of change, the assumptions and the context remained grounded. The discussions flowing from these workshops offered the evaluation team a more fine-grained lens from which to learn about and assess specific ToC components of the programme.

The keystone nodes were chosen because they are likely to offer learning opportunities critical to the programme, but also learnings for the wider M4P audience. The primary consideration in choosing nodes was to maximise opportunities for learning. In each case, the sector theory of change was reviewed in detail, along with background diagnostics for that sector and detailed intervention designs. In selecting nodes, the following questions were asked:

Which points of the theory of change are highly reliant on one or more 'critical assumptions'? [5]

Does the node contain inherent learning potential, particularly within programme specifics such as thin markets (for example, instances of sustainability, systemic change and resilience)?

Across the programme as a whole, is there a good balance between early-stage theory of change nodes and nodes located further along the results chain into the future?

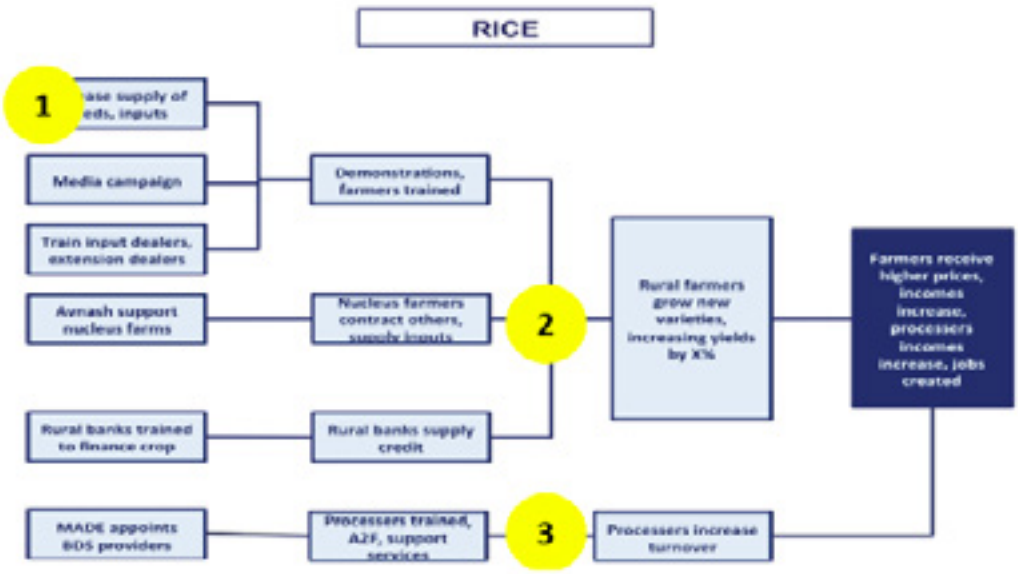
Does the node carry external relevance beyond the programme to other programmes on a regional or international basis?

The keystone nodes therefore focus on areas where motivating factors, causal mechanisms or incentives which are less well understood and therefore may hold the key to a proper understanding not only of that particular market but also, in combination with other evidence, for the programme as whole. As such, each keystone node is selected because it contains critical

[5] In contrast, 'general assumptions' are not necessarily programme-specific, sector-specific or node-specific. An example of general assumptions would be that peace continues in the Northern Savannah.

assumptions which are vital to the theory progressing from one stage to the next.
An example using the Rice ToC might serve to illustrate a keystone node. Using the process steps described earlier in the paper the evaluation team identified three keystone nodes within the rice sector. The programme's simplified version of the original theory of change is shown in Figure 1. For the purposes of this paper we elaborate on only one of the keystone nodes (marked 1 in Figure 1).

Figure 1: Original Rice ToC



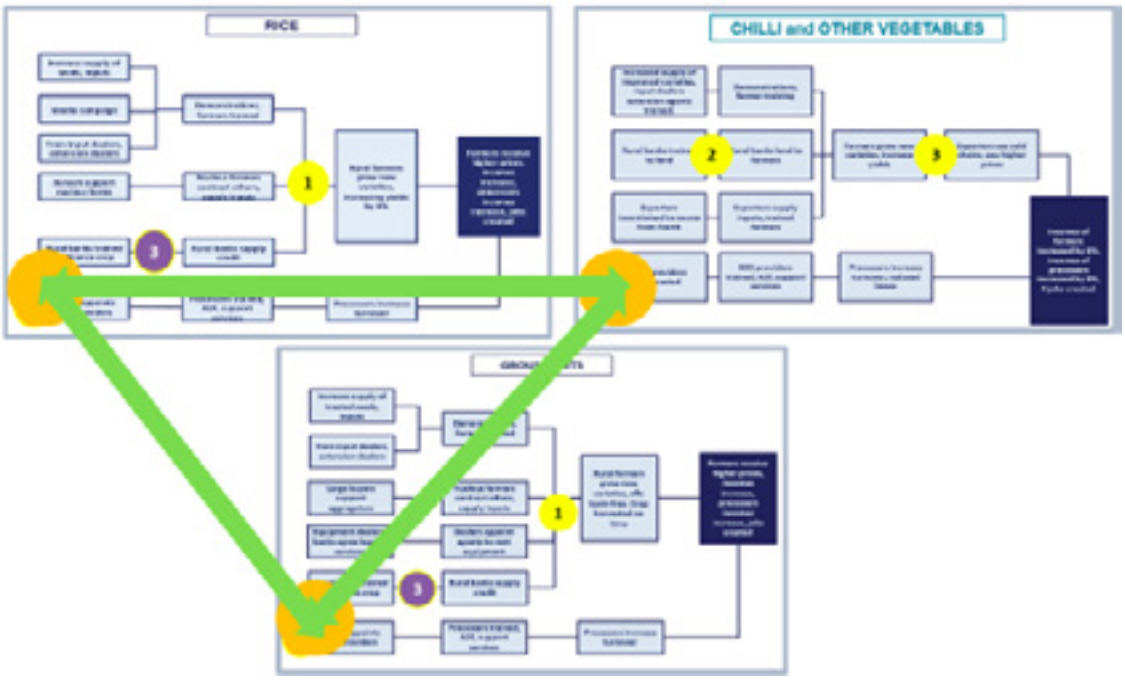
The programme's intervention in the rice sector seeks to address an unmet market demand for perfumed varieties of rice in Ghana, for which a premium market price is paid. The choice of nodes reflects the fundamental importance of ensuring adequate supplies of foundation seed for perfumed rice. Currently there is no reliable mechanism for introducing the perfumed variety. Currently only research institutes are authorised to certify that seed is 'true to type' and suitable for multiplication by the private sector. These institutions are unable to produce sufficient foundation seed and to supply sufficient breeder seed.

In the short term, the programme is intervening to encourage seed to be imported from Thailand. The need to ensure sufficient supplies of certifiable seed in reliable quantities is fundamental to the other programme interventions in this sector. The programme is confident that the market mechanism will take up new seed and distribute it (seed companies are keen to sell the new variety of seed) once the issue of supply in the early years has been met.
The choice of node R1 (increasing the supply of seed early in the theory of change), reflects its pivotal position in the overall intervention. If successful, seed will be multiplied in sufficient quantities, and institutional players will contract on an ongoing basis with commercial seed suppliers. If this node fails, then a large swathe of the ToC becomes irrelevant because of the dependence on the availability of new strain rice seed. Given these factors, this aspect of the ToC was regarded as pivotal to the success of general programme and was thus identified as a keystone node.

Step 2: Determining crosscut significance of keystone nodes.
Another central step in the keystone node approach is to determine the extent to which the selected nodes hold crosscutting significance. Stated differently: Do the nodes offer potential

learning for cross-cutting themes (such as gender or resilience)? And in relation to the programme as a whole, do the nodes selected offer opportunities to cross-refer between similar nodes in other sectors?
Guided by these questions, the evaluation team will seek to synthesise learnings across sectors to determine to what extent similar approaches within different market sectors have had an impact at programme level (See Figure 2).
12

Figure 2: Programme level evaluation of similar keystone nodes



This crosscutting significance is important for several reasons. First and foremost, it allows for the evaluation to understand the critical assumptions within each sector theory of change, and ultimately synthesise these findings at the programme-level theory of change. In this way, this "narrow" focus on specific nodes within market-level ToCs does not preclude – but rather supports programme-level evaluation. Secondly, the crosscutting significance supports the subsequent comparative analyses of the nodes, allowing for a better understanding of how these nodes function across different settings and contexts.

In the context of the present evaluation, the selected keystone nodes were scrutinised to determine the extent to which they addressed the cross cutting issues relevant to the programme. More specifically, each keystone node was assessed in terms of seven cross cutting issues: Thin markets, gender, sustainability, resilience, scale, systemic change and unintended consequences. While the assessment of crosscutting significance did not contribute to the initial choice or selection of the node, the assessment was undertaken to ensure that the selected keystone node reflected key programme issues to a greater or lesser extent. In determining the crosscutting significance, the evaluation team facilitated a workshop within which the evaluation team and the programme team collectively validated the selection of keystone nodes and the rationale surrounding the thinking behind each selection. During these workshops a number of keystone nodes were shifted, altered, or even deleted.

Step 3: Collecting Data on Keystone nodes.
The keystone node approach does give priority to specific data collections methods or sources. Qualitative and quantitative data are considered equally authoritative, albeit serving different purposes. As just one example, a breakdown in the programme at one of the keystone nodes may be readily apparent from monitoring data, but the reasons behind the failure or success may be less apparent. It is these reasons that attract the evaluation team to enquiry further, to rely on both quantitative and qualitative quality data.
The keystone node approach structures the data collection around a set of questions pertaining to salient aspect of the nodes. These include questions about whether:

1. The intervention was plausible? (Lemire, Nielsen, & Dybdal, 2012)
2. The intervention was been implemented as intended? (Weiss, 2000)
3. Unanticipated events caused changes in implementation and why?
4. Market players have been influenced to change the way in which they interact with their clients and suppliers?
5. Some market players responded more readily to the interventions than others?
6. The more unresponsive players will continue to play a valuable role in the ToC or if their involvement is superfluous?
7. And to what extent have market level TOCs been revised?

In the present evaluation, the data collection is just getting off the ground. However, recent reports from the programme show that interventions within the rice sector have not gone according to plan, with seed multiplication being reported as slow and dampened by unforeseen difficulties in brokering relationships between institutional research organisations and the private sector. As a result of both this change as well as other alterations in other intervention areas, alternative approaches are explored, proposed and new interventions designed. Accordingly, the theory of change for this sector may have to be refined.

Towards understanding the node and its refinement, we are currently in the process of conducting key informant interviews with individuals and institutional representatives who would have been involved in the discussions and planning for the replication and multiplication of the new seed varietals. The interviews have a specific focus on the perception of the institutional roles and the rationale for having to alter the approach of the intervention. Another central aim of the data collection is to assess programme fidelity (as reflected in question 1-3 above). An implicit assumption of the evaluation is that the interventions are firmly based on the sector-level theories of change. However, some of the specific interventions may fail in some way or be modified as the programme advances. This could be for a number of reasons: because circumstances change; a particular risk materialises; an uncertainty was misunderstood; or because a theory was flawed. The rice example above illustrates this nicely. The programme is designed to be tolerant of intervention change or even failure, within reason, while the evaluation needs to be similarly robust and flexible. If the programme implementation fails to progress through the theory of change as planned, the choice of examined node would shift accordingly to accommodate the change in programme implementation, with the intention to revert to the original choice of node later.

Documenting the fidelity of the interventions matters greatly for learning. The evaluation of the nodes aims to gather considerable information related to the node in relation to one of four scenarios;
1. Strong theory/weak implementation (implementation failure): The implementation has departed from the proposed theory. Feedback to the implementing partner would be to support a stronger alignment between the theory and implementation of the programme.
2. Strong theory/strong implementation: A learning moment. Something has been successfully tested and we will learn from the success or failure, allowing us to build on these results.
3. Weak theory/weak implementation: A worst case scenario because we don't know why the failure took place.
4. Weak theory/strong implementation (theory failure): Another learning moment. Something has been tested, but we are not sure why it works (or fails to work).

Step 4: Conducting cross-case analysis
Broadening the in-depth, node specific understanding developed within each sector ToC requires synthesis across similar nodes in other sectors to assess how the nodes perform across different settings, populations, and times.
Two strategies will be employed to conduct the programme level evaluation. The first strategy focusses on a comparison across nodes of cross cutting issues identified within the nodes detailed above. For example where nodes within different sectors have sought to address thin markets, the evaluation will be able to speak to impact of the programme to address this constraint. The second strategy relies on similarities of interventions within different market level ToCs, for example, where interventions have been designed to incentivise business development service providers (BDS) or micro finance institutions (MFI) to respond to anticipated market demand.

Final thoughts

The keystone node approach is a work in progress. Motivated by the common pitfalls of diluted theories of change, confirmation bias, and analytical murkiness restricting many theory-based evaluations, the approach is still in need of further development and practical application. Our modest hope is that the approach will advance theory-based evaluation towards tackling these persistent challenges. While it is currently too early to speculate on the findings, it is hoped that we will have sufficient evidence to critically reflect on the process and provide learnings to the programme to improve their future implementation.

References

Chen, H.-T. (2005). Practical programme evalaution: assessing and improving planning, implementaion and effectiveness. CA: Thousand Oaks.

Chen, H.-T., & Rossi, P. (1992). Using theory to improve programme and policy evaluations. New York: Greenwood.

Coryn, C., Noakes, L., Westine, C., & Schröter, D. (2011). A Systematic Review of Theory-Driven Evaluation Practice From 1990 to 2009. American Journal of Evaluation, 32(2), 199-226.

Delahais, T., & Toulemonde, J. (2012, July). Applying contribution analysis: Lessons from five years of practice. Evaluation, 18(3), 281-293.

Dybdal, L., Nielsen, S. B., & Lemire, S. (2011). Contribution analysis: applied reflections on scope and method. Canadian journal of Programme Evaluation, 29-57.

Green, D. (2012 ff.). From Poverty to Power. Retrieved from http://oxfamblogs.org/fp2p/.

Lemire, S., Nielsen, S. B., & Dybdal, L. (2012). Making contribution analysis work: a practical framework for handling influencing factors and alternative explanations. Evaluation, 18(3), 294-309.

Mayne, J. (2012). Contribution analysis: coming of age? Evaluation.

Patton, M. (2015). Qualitiative Research and Evaluation Methods. London: Sage.

Ruffer, T., & Wach, E. (2013). Review of M4P Evaluation Methods and Approaches. Hull: Itad.

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). Broadening the Range of Designs and Methods for Impact Evaluations. DFID.

Valters, C. (2014). Theories of Change in International Development: communication, learning or accountability? London: Justice and Security Research Programme.

Valters, C. (2015). Theories of Change: Time for a radical approach to learning in development. London: ODI.

Vogel, I. (2012). Review of the use of Theory of Change in international development. London: DFID.

Weiss, C. H. (2000). Which links in which theories shall we evaluate? New Directions in Evaluation, 87(Fall), 35-45.

Using Impact evaluation for education policy innovations: the case of early grade literacy in South Africa

First author
Nompumelelo Mohohlwane

Second author
Stephen Taylor

1.Background: Development through quality education

South African and international development planning is increasingly centered around education. Agreements such as the Millennium Development Goals (MDGs) and the Education For All (EFA) drive are testament to the international recognition of the pivotal role of education. In South Africa, the National Development Plan, which is arguably the country's most prominent planning document, affords a central role to education, while the Presidency officially regards improved quality of basic education as the country's number 1 priority (NPC, 2012).

Despite substantial progress in expanding access to schooling in developing countries over the past few decades, there is now a growing recognition that in many countries the learning outcomes achieved by those attending school are often dismally poor. Spaull and Taylor (2015), for example, demonstrate that despite improved access to schooling in many Southern and East African countries there are large proportions of children who reach grade 6 without having acquired basic literacy and numeracy skills. This is important since there is clear evidence that the quality of skills achieved (over and above the quantity of schooling attained) has a significant impact on economic growth and on the labour market prospects of individuals (Hanushek & Woessmann, 2007). While there are numerous social, psychological and other benefits of education, the quality of learning outcomes should be seriously considered when analyzing education system performance.

2.The education quality challenge in South Africa

Local and international assessments of learner performance consistently indicate poor performance by South African learners in Mathematics, Science, and Languages (Spaull, 2014). The low performance across the board is concerning but for the purposes of this paper the focus will be on literacy and reading.

The Pre-PIRLS 2011 results indicated that 29% of South African Grade 4 learners did not have the rudimentary reading skills required at a Grade 2 level. The situation was most severe for those learning in an African language. For example, 57% of learners that took tests in Sepedi or Tshivenda did not reach this level (University of Pretoria, 2012). The 2006 PIRLS study, which tested grade 5 South African learners on a somewhat more advanced reading test, showed that approximately 80% of children had not learned to read for meaning by grade 5, but at best could only extract basic factual details from a text. If children have not learned to read fluently by this time, it stands to common sense that they will not be able to cope with the requirements of the curriculum at higher grades. Weak reading foundations are therefore at the heart of the education quality challenge in South Africa.

South Africa's underperformance, even relative to many poorer countries, prompts an enquiry into the causes behind the limited ability of the education system to convert inputs into outcomes. South Africa's per pupil expenditure in schools in purchasing power parity (PPP) terms exceeded that of all the other thirteen countries in SACMEQ with the exception of Seychelles. An important part of the contextual background is the inequality resulting from the political history of South Africa. The changes in the education system following the end of Apartheid and the establishment of a new democratic state in 1994 have had limited success in changing the nature of schooling. Several authors, such as Fleisch (2008) and Spaull (2014), argue that South African schooling has still effectively got two systems in one. The first and largest part of the system comprises the historically disadvantaged schools and is characterized by inefficiency including poor school management, continuous underperformance, high and indiscriminate grade repetition and dropout. The second system refers to historically white and Indian schools where learner performance is at a higher standard, parents make substantial fee contributions, organizational and instructional processes are more efficient and schools are well endowed with infrastructure.

The South African government is well aware of these challenges and continues to allocate the largest share of government expenditure to education. Since the early 1990s education spending has become increasingly well targeted to poor schools (Gustafsson and Patel, 2008). Specific initiatives and policies implemented by the South African government to address equity challenges in education include the introduction of no-fee schooling which is implemented in about 77% of public schools, and the provision of daily meals through the National School Nutrition Programme to approximately 70% of schools focusing on the poorest schools (Department of Basic Education, 2014).

Despite these considerable efforts, however, learning outcomes remain low in South Africa and little is known about the effectiveness of particular policies and programmes designed to improve learning. Where evidence is available it is often self-reported, focused on inputs, anecdotal or part of a larger initiative where the effect of specific efforts is difficult to isolate. This motivates for an agenda of impact evaluation to inform policy-making going forward, as will be argued below.

3. The importance of early literacy learning in South Africa

Literature on the evidence of early learning emphasizes the importance of mastering certain learning foundations for the sake all further learning. The literature refers to 'self-productivity', explaining that skills acquired during one period generally persist into the next period and may make the acquisition of other skills in another dimension easier (Girdwood, 2013). In addition to the argument for the cognitive benefits of the development of good educational foundations and their lasting effects, James Heckman (2007), amongst others, contends that intervening earlier rather than later is more cost-effective. The costs of providing curriculum support for areas of learner deficits identified early, such as in the Foundation Phase are expectantly lower than mediating learning later in schooling where the gap between curriculum expectations and learner knowledge may be excessively large in a multitude of subjects, as Pritchett and Beatty (2015) have shown. The costs accrued at later stages include high rates of grade repetition and dropping out of the education system.

One critical learning foundation that needs to be acquired during the early grades of primary schooling is reading. A large theoretical literature points to the benefits of learning to read in the home (or first) language. One of the expected benefits is that second language acquisition should be easier once a firm grasp of the nature of reading and literacy has been attained in one language. A paper by Taylor and Coetzee (2013) provides some empirical evidence from South Africa that home language instruction in grades 1 to 3 caused improved English literacy in grades 4 to 6 compared with children who were taught in English as the language of instruction. This finding substantiates the argument that all learning builds on prior learning; as such mastery of a second language is enabled by the mastery of the first language. This points to the strategic value of finding ways to improve home language reading acquisition in the Foundation Phase. Yet, the reality is that the majority of children will experience a transition to English as the language of instruction in the fourth grade. Finding ways to strengthen English vocabulary and manage this transition most effectively will therefore also be important.

4.Education Policy Development and the Evaluation Process

A detailed review of South African education policy development is beyond the scope of this paper. What is clear, however, is that despite many policy changes and new programmes, little is known about the ultimate impact of these initiatives on learning outcomes. The lack of a focus on impact evaluation is not unique to South Africa, as the following quote illustrates: "Development programs and policies are typically designed to change outcomes, for example, to raise incomes, to improve learning, or to reduce illness. Whether or not these changes are actually achieved is a crucial public policy question but one that is not often examined. More commonly, program managers and policy makers focus on controlling and measuring the inputs and immediate outputs of a program—how much money is spent, how many textbooks are distributed—rather than on assessing whether programs have achieved their intended goals of improving well-being" (World Bank, 2010).

In cases where extensive research is done – at least in South Africa – it is typically focused on diagnosing areas requiring attention rather than evaluating possible solutions. Where interventions are evaluated it is often through conducting case studies or piloting in a small number of schools. The shortcoming of this approach is that the implementation model often used in case studies or small-scale pilots is often resource intensive and may be difficult to replicate at a larger scale.

A focus on evaluation is now emerging within the South African government through the introduction of the National Evaluation Policy Framework in 2011. This policy framework includes a National Evaluation Plan (NEP) which commissions independent evaluations of priority government programmes in a partnership between the custodian department and the Department of Planning, Monitoring and Evaluation (DPME, 2014). Several Department of Basic Education (DBE) programmes have been evaluated through the NEP, namely the Grade R programme, the Funza Lushaka Bursary Programme and the National School Nutrition Programme.

The evaluations referred to above are all retrospective evaluations, assessing how well programmes were implemented or if the intended programme goals were attained. Prospective

impact evaluations, where programmes are evaluated prior to being taken to scale, remain extremely rare. One exception to this is the impact evaluation of a new set of study guides introduced by the DBE in 2012 (Department of Basic Education, 2013).

5.Using Randomised Control Trials in Education

The major challenge in impact evaluation is the need to identify a counterfactual – what would have happened to programme recipients in the absence of the intervention? Since one can never actually observe a counterfactual to reality, one needs to use a "control group" or "comparison group" to provide a valid estimate of the counterfactual. Simply comparing recipients with non-recipients or pre- and post-outcomes amongst recipients is usually not likely to provide a valid estimate of the counterfactual since recipients are usually systematically different to non-recipients and outcomes would change over time in any event.
While various quantitative impact evaluation methods are available, the cleanest method for identifying an internally valid estimate of the counterfactual is obtained through conducting a Randomised Controlled Trial (RCT). Through using a lottery to allocate participants to intervention and control groups, an RCT constructs a credible "counterfactual" scenario – what would have happened to those who received an intervention had they not received that intervention.

Prospective impact evaluations also have the advantage for research of uncovering knowledge of the binding constraints in the school system. In complex environments, such as education, there are multiple factors influencing outcomes and it is not always clear which factors to address first. For example, high quality teaching requires both competent and motivated teachers, though it is not clear which of these is the more binding constraint in South Africa. Teacher knowledge in South Africa is weak: Carnoy et al (2011) found that grade 6 teachers recorded an average score of around 40% on a test designed to assess their mathematics knowledge for that grade. Furthermore, studies show low teacher motivation in poor schools, manifested in high absentee rates and low teaching activity (Reddy et al, 2010). Yet, it is unclear whether to address teacher capacity or teacher motivation first. The lack of rigorous evaluations to establish which of these challenges to address first is a shortcoming of conventional policy and programme development.

5.1 Practical considerations when implementing an RCT

Statistical expertise is required in the design of an RCT. This involves calculating the required sample sizes in each intervention and control group and conducting the random assignment. For practical reasons, when conducting an education RCT it is often necessary to assign schools as a whole to intervention or control groups, as opposed to assigning individuals to the different experimental groups. This leads to rather large required samples, which has cost implications. The need to raise funds takes time and requires significant stakeholder engagement and government support to convince donors to be involved.

There are two main components to an education RCT – there is the implementation of the new interventions and there is the evaluation of their impact. The evaluation side of the RCT involves the collection of outcomes data as well as contextual data for the sake of measuring changes in intermediate outcomes and identifying factors that mediate the impact of the intervention. Both the implementation and the evaluation components require financing and

should be conducted by separate organizations. In some cases, an NGO or a government department may fund and implement the interventions, thus reducing the need for additional fund raising. If reliable outcomes data already exist, through for instance a nationally standardized examination, then one might be able to significantly reduce costs associated with the evaluation side of the project.

Both academic researchers and implementing agencies face various perverse incentives when considering or conducting evaluations. A publication bias exists in academia where it is more likely to see studies with positive results published than studies showing no impact (Duflo, Glennerster and Kremer, 2006). RCTs are less prone to this bias, since the large investment of time and resources together with the high reliability of the results mean that even evaluations showing zero impact are likely to be published.
Government departments and NGOs may resist evaluations due to the risk of negative findings. Therefore, prospective evaluations of alternative programmes or variations of programmes under consideration may be more amenable to policy makers and programme managers who will then not feel that their entire work for several years is being judged.
For these reasons, the success of any RCT is dependent on extensive stakeholder consultation and support. This ensures that funds and other resources including personnel are availed; the integrity of the research design is upheld; the implementation of the interventions is conducted properly, and the findings are considered for programme or policy scale-up or redesign.

6.Experiences from two new reading evaluations in SA

Through a developing partnership between education researchers, government and donors, two studies are being undertaken to evaluate possible ways to improve reading acquisition in South African schools.

6.1 Evaluation of a remedial reading programme in Grade 4

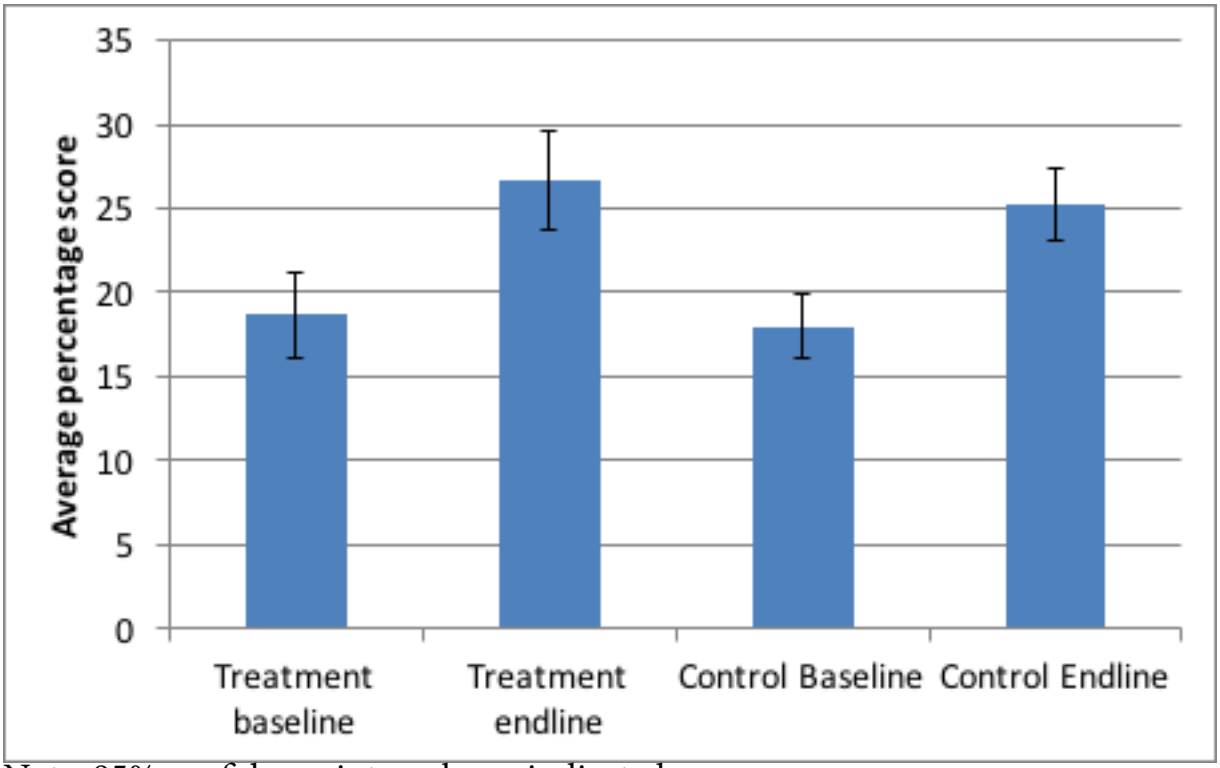
The Gauteng Primary Language and Mathematics Strategy (GPLMS) implemented between 2011 and 2014, included various new interventions focused on the early grades. As part of this, a Reading Catch-Up Programme (RCUP) was developed to strengthen the English skills of children in grades 4 and 5 whose first language was not English but who are required to learn using English as the language of instruction in those grades. A preliminary evaluation of the programme indicated large gains in the language skills of programme recipients over time (Hellman, 2012). However, there was no control group. Therefore, some strong assumptions had to be made about how much learning would have taken place over the period had there been no intervention.

These initially promising, though inconclusive, results prompted an RCT of the RCUP to be conducted in the district of Pinetown in the KwaZulu-Natal province in 2014 (Fleisch, Taylor, Schöer, and Mabogoane, 2015). The intervention lasted for 11 weeks and consisted of on-site teacher support by reading coaches and the provision of scripted lesson plans and additional graded reading books. The RCUP targeted Grade 4 learners in schools that transition to English as the language of instruction after using the home language in the Foundation Phase. The hypothesis underlying the programme was that the learning gaps in learner mastery of English at the end of the second term in Grade 4 may be caught-up through the

provision of a well-designed relatively short intervention. The programme was implemented in 40 intervention schools with a control group of 60 schools. Assignment to intervention and control group was done through a computerized lottery. Different organizations were contracted to conduct the intervention and the data collection for evaluation, and the evaluation agent was kept blind to which schools were in the intervention group versus the control group.

The most notable finding of the study was that although learners in intervention ("treatment") schools improved their test scores between the baseline and the endline assessment, the learners in control schools improved by a similar margin, as depicted in Figure 1. This illustrates the importance of obtaining an estimate of the counterfactual: in the absence of a randomly selected control group a false positive result would have been obtained. The main finding, then, is that the RCUP intervention had no statistically significant impact on the overall reading achievement of learners. However, treatment schools improved more than control schools in the spelling and grammar subcomponents of the test. The programme impact was larger for learners who initially had a basic minimum of English skills and for those whose teachers participated actively in the programme.

Figure 1: Average pre- and post-scores for intervention and control schools



Note: 95% confidence intervals are indicated
Source: Fleisch, Taylor, Schöer, and Mabogoane, 2015

The findings from the RCUP RCT yield several important policy lessons and raise several questions for further research. Firstly, it is evident that ostensibly well-designed programmes may not have as large an effect as one might expect in the absence of a rigorous evaluation. Apart from the design quality of a programme and the integrity of implementation, there may be contextual factors pertaining to learners, schools and communities which either preclude or are conducive to the effectiveness of an intervention.

Secondly, the findings indicate that the RCUP cannot yet be implemented on a wider scale with any confidence that it will have a significant impact on learning outcomes. This does not necessarily mean that reading remedial programmes in general, or even the RCUP specifically, should not be further explored. It does, however, mean that a revised version should first be experimented with and shown to work before government should consider a larger scale implementation.
Thirdly, the independently administered tests indicated that the learning deficits existing by the end of the Foundation Phase are apparently much larger than expected. The finding that initially better-performing learners gained more from the intervention may imply that the programme would have been more appropriate at the grade 5 level, even though it was covering topics that should have been covered in the Foundation Phase. Another possibility is that an 11-week intervention is simply too short a time to deeply influence classroom practice and learning.

6.2 The Early Grade Reading Study (EGRS)

The second RCT which is being conducted by the Department of Basic Education (DBE) is the Early Grade Reading Study (EGRS). This RCT has recently (February, 2015) commenced in 230 schools in the North West province and has been crafted to evaluate three competing interventions all aimed at improving home language (Setswana) reading acquisition in grades 1 and 2.

The sample of schools includes non-fee paying schools in the Dr Kenneth Kaunda and Ngaka Modiri Molema districts of the North West province. All schools selected use Setswana as the language of instruction in the Foundation Phase. All three interventions will occur over a two-year period working with the cohort of children entering grade 1 in 2015.
The RCT will evaluate the causal impacts of three interventions: (i) a teacher training course focused specifically on the teaching of Setswana reading and literacy, accompanied by scripted lesson plans and graded reading materials; (ii) an on-site support programme to teachers from reading coaches, accompanied by scripted lesson plans and graded reading materials; (iii) and a package designed to improve parent involvement in – and monitoring of – learning to read. Each intervention will be implemented in 50 schools within the sample. A further 80 schools have been selected as the comparison group. As was the case in the RCUP study, separate organizations have been contracted to undertake the implementation of interventions and the data collection for evaluation, with the evaluation agent blind to which group schools fit into.

This study is expected to shed light on several research and policy questions. Firstly, it will show which of three alternative interventions is most cost-effective. Although each intervention has a different unit cost, the improvement in test scores per Rand spent for each intervention will be calculated. The evaluation will also investigate whether the impacts of interventions are different for various sub-groups of learners or schools. This will inform the most appropriate targeting of interventions if scaled up. The study is also designed to look at long-term effects and spillover benefits of faster reading acquisition. Do the impacts of the interventions persist, dissipate or compound over time? If one succeeds in improving the acquisition of home language reading in the early grades, are there spillover benefits into other

learning areas such as Numeracy and First Additional Language? This will be measured using results of the Annual National Assessments (ANA) in subsequent years.

7.Conclusion

This paper has demonstrated that improving the acquisition of reading in the early grades is central to the education quality challenge in developing countries, and especially in South Africa. In the absence of solid evidence of effective policies and programmes to address this challenge, the use of prospective impact evaluations is recommended. The paper has pointed out the necessity for innovation and rigor to establish such evidence and understanding the binding constraints in the complex South African education system.

These points are substantiated through a description of two recent RCTs focusing on early grade reading in South Africa. The RCT of the Reading Catch-up Programme, implemented in Pinetown, Kwa-Zulu Natal has provided important lessons. The findings have highlighted the need for a valid counter-factual in measuring impact, which is a strength of the RCT methodology. The second RCT discussed, the Early Grade Reading Study (EGRS), which is being implemented in 230 schools in the North West province provides an exciting opportunity for further learning. It is anticipated that the findings will address some of the questions emerging from the RCUP study as well as provide substantive information on the binding constraints in the teaching of language in South African schools.

8.Bibliography

Boiseiere, M. (2004). "Determinants of Primary Education Outcomes in Developing Countries". Washington: The World Bank

Carnoy,M. Chisholm,L. Addy,N. Arends,F. Baloyi,H. Irving, M. Raab, E. et al. (2011). The process of learning in South Africa: the quality of mathematics teaching in North West Province. Technical Report. Stanford University School of Education and HSRC.

Department of Basic Education. (2014). Annual Report 2013/2014. Pretoria. Retrieved 15 May 2013 from www.education.gov.za

Department of Basic Education (2013). Technical Report. Developing and Evaluating the First Phase of the Grade 12 Mind the Gap Study Guide Series. Pretoria.

Department of Basic Education. (2012). Summary Report: Evidence Based Policy Review Workshop. Pretoria

Department of Basic Education. (2014). Terms of Reference. Implementation of Reading Interventions for the research project: "Early Grade Reading Study. Pretoria

Department of Planning, Monitoring and Evaluation. (2014). Concept for the National Evaluation Plan 2015/16. Pretoria.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. Center for Economic Policy Research.

Fleisch, B. (2008). Primary Education in Crisis: Why South African schoolchildren underachieve in reading and mathematics. Cape Town: Juta.

Fleisch, B.Taylor, S. Schöer, V. & Mabogoane, T. (2015). Assessing the impact of the RCUP: The findings of the impact evaluation of the Reading Catch-Up Programme. Paper Presented at the Wits School of Education Research Seminar.

Girdwood, E. 2013. Exploring the Evidence on Early Learning: A Literature Review. Impact Evaluation of the Introduction of Grade R on Learning Outcomes. Research on Socio-Economic Policy (RESEP). Department Of Economics, University Of Stellenbosch

Gustafsson, M. & Patel, F. (2008). Managing the teacher pay system. What the local and international data are telling us. Stellenbosch Economic Working Papers. 26(06)

Hanushek,E,A & Woessmann,L.(2007). Education quality and economic Growth. The World Bank: Washington

Heckman, J. (2007). The economics, technology, and neuroscience of human capability formation. NBER Working paper 13195.

Hellman, L. (2012). GPLMS Intersen Catch-up Programme: Analysis of Results. Memo.

National Planning Commission. (2012). National Development Plan: Vision for 2030. Retrieved January 18, 2013, from http://www.npconline.co.za/

Pritchett,L. & Beatty,A. (2015). Slow down, you're going too fast: Matching curricula to student skill levels. International Journal of Educational Development. 40 (1), 276–288.

Reddy, V., Prinsloo, C., Netshitingani, T., Moletsane, R., Juna, A., & Janse van Rensburg, D. (2010). An investigation into educator leave in the South African ordinary public schooling system.

Spaull, N. (2014) Education Quality In South Africa And Sub-saharan Africa: An Economic Approach (Unpublished doctoral thesis, University of Stellenbosch, Stellenbosch)

Spaull, N., & Taylor, S. (2015). Access to What? Creating a Composite Measure of Educational Quantity and Educational Quality for 11 African Countries. Comparative Education Review , 59 (1), 133-165.

Taylor, S. & Coetzee, M. (2013). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. Stellenbosch Economic Working Papers: 21/13

University of Pretoria. (2012). PIRLS 2011: Media Release. 11 December 2012. Retrieved 12 December 2012. from www.up.ac.za

World Bank. (2010) Impact Evaluations in Practice. Retrieved 10 October 2014 from http://www.worldbank.org/ieinpractice