Research Evaluation, 2020, 1–15
doi: 10.1093/reseval/rvaa024
Article

Understanding and evaluating the impact of integrated problem-oriented research programmes: Concepts and considerations

Brian M. Belcher [1,2,*] **and Karl Hughes** [3]

[1] Sustainability Research Effectiveness Program, College of Interdisciplinary Studies, Royal Roads University, Victoria V9B 5Y2, Canada, [2] Center for International Forestry Research, PO Box 0113 BOCBD, Bogor 16000, Indonesia and [3] World Agroforestry (ICRAF), United Nations Avenue, Gigiri, PO Box 30677, Nairobi 00100, Kenya

*Corresponding author. Email: brian.belcher@royalroads.ca.

Abstract

Researchers and research organizations are under increasing pressure to demonstrate that their work contributes to positive change and helps solve pressing societal challenges. There is a simultaneous trend towards more engaged transdisciplinary research that is complexity-aware and appreciates that change happens through systems transformation, not only through technological innovation. Appropriate evaluation approaches are needed to evidence research impact and generate learning for continual improvement. This is challenging in any research field, but especially for research that crosses disciplinary boundaries and intervenes in complex systems. Moreover, evaluation challenges at the project scale are compounded at the programme scale. The Forest, Trees and Agroforestry (FTA) research programme serves as an example of this evolution in research approach and the resulting evaluation challenges. FTA research is responding to the demand for greater impact with more engaged research following multiple pathways. However, research impact assessment in the CGIAR (Consultative Group on International Agricultural Research) was developed in a technology-centric context where counterfactual approaches of causal inference (experimental and quasi-experimental) predominate. Relying solely on such approaches is inappropriate for evaluating research contributions that target policy and institutional change and systems transformation. Instead, we propose a multifaceted, multiscale, theory-based evaluation approach. This includes nested project- and programme-scale theories of change (ToCs); research quality assessment; theory-based outcome evaluations to empirically test ToCs and assess policy, institutional, and practice influence; experimental and quasi-experimental impact of FTA-informed 'large n' innovations; *ex ante* impact assessment to estimate potential impacts at scale; and logically and plausibly linking programme-level outcomes to secondary data on development and conservation status.

Key words: research evaluation; impact assessment; evaluation tools; theory of change; transdisciplinary research; sustainability science

1. Introduction

Researchers and research organizations are under increasing pressure to demonstrate that their research contributes to positive change and helps to solve pressing societal challenges. Appropriate evaluation is therefore needed, not only to evidence research impact, but also to generate learning to improve research design and, ultimately, enhance impact. It is also critically important to demonstrate the contribution of research to solving development problems and leverage opportunities to attract, allocate, and optimize investments in research. This is challenging in any research field, but especially for integrated research programmes that cross disciplinary boundaries to intervene in complex systems.

The drive for increased research impact has led to a marked evolution in the way research-for-development (R4D) is understood, conceived, and implemented, with more inter-disciplinary research and transdisciplinary forms of collaboration between researchers, research users, and other stakeholders (Nowotny, Scott and Gibbons 2001; Kasemir, Jaeger and Jäger 2003; Hirsch Hadorn et al. 2006). This evolution reflects epistemological assumptions that are very different from those of traditional disciplinary approaches (Talwar, Wiek and Robinson 2011). There is greater appreciation of contingency and uncertainty in science (Gibbons et al. 1994; Nowotny, Scott and Gibbons 2001). There is also recognition that scientific knowledge alone is not sufficient for action, and rather that sustainable development entails many normative considerations that link knowledge and action (Functowicz and Ravetz 1993; van Kerkhoff and Lebel 2006). Furthermore, there is greater appreciation that the knowledge and values of stakeholders and intended users of research are relevant, valid, and important, and that each has their own motivations and biases that influence how they interact with and make use of new knowledge (Kasemir, Jaeger and Jäger 2003).

This changed understanding has led to fundamental changes in the way many researchers work. New problem-oriented research approaches have evolved to engage system actors in the research process as a way to increase research effectiveness. Variations on these approaches are known as Post-Normal Science (Functowicz and Ravetz 1993; Ravetz 1999), Mode 2 research (Functowicz and Ravetz 1993; Gibbons et al. 1994), Problem Driven Iterative Adaption (PDIA) (Andrews, Pritchett and Woolcock 2013), Transdisciplinary Research (TDR) (Klein 2006; Walter et al. 2007; Carew and Wickson, 2010; Pohl et al. 2010; Jahn, Bergmann and Keil 2012; Lang et al. 2012; Wolf et al. 2013), and Sustainability Science (Kates et al. 2001; Clark and Dickson 2003; Komiyama and Takeuchi 2006; Brandt et al. 2013; Kauffman and Arico 2014; Heinrichs et al. 2016; Kates 2017; Roux et al. 2017). There has also been a recent turn towards large, coordinated, multi-disciplinary research collaborations focused on major societal problems, such as the Grand Challenges in US universities (Popowitz and Dorgelo 2018) and the Global Grand Challenges (Bill and Melinda Gates Foundation n.d.).

The CGIAR (formerly known as the Consultative Group on International Agricultural Research), an international consortium on agriculture and natural resource management (NRM) research, provides a good example of this transition and the corresponding learning and impact assessment challenges. An organizational reform process, which began in 2008, increased accountability for realizing social, economic, and environmental outcomes and impacts, on top of the long-established commitment to producing high-quality science. This shift was made explicit as a commitment to 'shared responsibility' (ISPC 2015: 5) for impacts in terms of reduced poverty, improved food and nutrition security, and improved natural resources and ecosystem services (CGIAR 2016). A key aspect of this reform process was the creation of CGIAR Research Programs (CRPs) in 2011. These aimed, in part, to build broader and deeper partnerships, not only with other research organizations, but also with a range of policy and development actors at international and national levels, including conservation and development organizations, non-governmental organizations, policy actors, and other stakeholders. This emphasis on working through partnerships appreciates that high-quality scientific knowledge creation alone cannot address contemporary development and environmental challenges. The resulting research embodies many of the characteristics of problem-oriented TDR approaches. This is not to say that all CGIAR research is transdisciplinary; however, there is a growing proportion of CGIAR projects that apply TDR approaches, and the overall portfolio is increasingly integrated and focused on high-level challenges.

There has been good progress developing methods for assessing the societal impacts of research, especially in the health field (Greenhalgh et al. 2016). However, contemporary concepts and methods of research impact assessment are not well suited for complex integrated research programmes. Impact assessment in the CGIAR has developed in conjunction with a historically technology-centric research model, leading to a substantial mismatch between the prevailing approach to research impact assessment and current needs and opportunities. Simply stated, relying solely on counterfactual impact evaluation approaches is inadequate for evaluating the full range of CGIAR research or for engaged problem-oriented research more generally. There is a need for a more complete and comprehensive set of approaches for analysing and demonstrating the impacts of integrated inter- and transdisciplinary research programmes, such as those of the CGIAR, Challenge Programs, and sustainability science. We need a broader and more nuanced conceptual framework of how research contributes to change in complex systems and how we can evaluate those contributions for both learning and accountability.

This essay explores the needs and opportunities for improved evaluation and impact assessment in an international R4D context, with a focus on lessons from the Forests, Trees and Agroforestry (FTA) CRP. The authors have many years of experience with research evaluation and impact assessment in the CGIAR and other international research and development organizations. We begin with a review of the evolution of problem-oriented research, using examples from the FTA. We then discuss the inherent challenges in evaluating this kind of research, including the limitations of conventional impact assessment approaches. We recognize that designing and selecting appropriate evaluation methods depends on expectations, so we propose a set of principles to shape such expectations and guide the development of more appropriate and realistic evaluation frameworks for multifaceted research initiatives that aim to bring about transformational systems-level change. We conclude by proposing an integrated evaluation framework to operationalize these principles.

2. Evolving modes of research

To assess the impact of any intervention or innovation (including those informed by research), it is critically important to first understand the likely mechanism(s) through which it is expected to generate its intended effects, and to investigate the extent to which these expectations conform with reality (Pawson 2003; White 2009). Indeed, there is a pushback against evaluations that fail to interrogate why interventions, programmes, and projects succeed and/or fail (Harachi et al. 1999).

The standard and still prevailing mechanism through which R4D is expected to generate impact is the linear model of diffusion (Godin 2006), also known as the pipeline model (Hall et al. 2000). In this model, scientific discovery leads to technological innovation that is piloted, refined, and then disseminated to and adopted by intended users at scale, resulting in efficiencies, improvements, and benefits to society.

van Kerkhoff and Lebel (2006) suggest two versions of this conventional model that illustrate the link between knowledge and action. The trickle-down model holds that good research will be taken up by users based on its inherent value. The researcher's job is to do good quality, innovative science and share the resulting knowledge through normal scientific communications, such as peer-reviewed articles and conference presentations. The second version, the transfer and translate model, emerged out of research utilization studies in the 1970s which recognized that trickle-down approaches had largely failed to influence social policy. This version of the model acknowledges the need for additional effort to communicate science, but it is still framed as a one-way process of translating, transferring, and mobilizing science-based knowledge to users. Examples include agricultural extension services, where specialized intermediary agents transfer research findings to users, or evidence-based healthcare, which consolidates scientific knowledge through evaluations and syntheses of existing scientific literature and translates the knowledge, ostensibly for application in clinical practice and public policy. The model assumes that there is an objective truth that can be discovered by science and that the main barrier to improved outcomes is lack of access to that scientific knowledge by intended users.

There is a growing consensus among practitioners, policymakers, and the research community that the scaling of technological innovations alone cannot solve contemporary social and economic challenges (Howaldt 2019). van Kerkhoff and Lebel (2006) list the key critiques: science is socially and institutionally embedded and cannot be entirely objective in its definition and execution; scientific knowledge is socially constructed, in that observations are subject to interpretation, so knowledge is always uncertain; the boundary between science and the rest of society is artificial, created by social and political processes and therefore changeable and contestable; power and special interests shape the linkages between research-based knowledge and action; and science reflects cultural biases and inequalities.

Indeed, there are high and increasing expectations that science should serve society and the benefits should be demonstrable (Stokes 1997; Sarewitz 2016). Functowicz and Ravetz (1993) published their seminal work on 'Science for the Post-Normal Age', recognizing that in many contemporary societal problems, facts are uncertain and values play a major role in decision-making. As noted by Ravetz (1999), '[c]ontrary to the impression conveyed by textbooks, most problems in practice have more than one plausible answer, and many have no answer at all' (649). The ideas and concepts of post-normal science call for new problem-solving strategies in which the role of science is appreciated within complex and uncertain natural and social systems.

Increased focus on problem-solving and social engagement has led scholars and researchers to develop TDR approaches that integrate across disciplines and beyond expert knowledge to embrace non-expert and public knowledge and enable social learning (Randolph 2004; Hirsch Hadorn et al. 2008; Pahl-Wostl, Mostert and Tabara 2008; Robinson 2008; Lang et al. 2012). Lang et al. (2012) outline the key steps in TDR as: (1) joint framing of the problem and building a research team composed of different kinds of scientists and societal stakeholders; (2) co-producing solution-oriented and applicable knowledge through collaborative research; and (3) (re)integrating and applying the knowledge that has been produced in both scientific and societal practice. Thus, TDR promotes sustainable change with methods that give non-academic actors a role in the research process and which integrate and expand their knowledge and capabilities, leading to improved action.

Figure 1 illustrates, in a simple schematic diagram, some of the kinds of value that can be created at different stages in the research process through productive interactions.

Sustainability Science likewise takes a TDR approach. Kates et al. (2001) define the three core objectives of Sustainability Science as: (1) understanding the fundamental interactions between nature and society; (2) guiding these interactions along sustainable trajectories; and (3) promoting social learning necessary to navigate the transition to sustainability. A key characteristic of Sustainability Science is that research is defined by the problems it addresses rather than the discipline(s) it employs (Kates et al. 2001; Clark and Dickson 2003; Clark 2007; Bettencourt and Kaur 2011). Transitioning to sustainability requires socio-technical change in the rules, practices, and norms that guide the development and use of

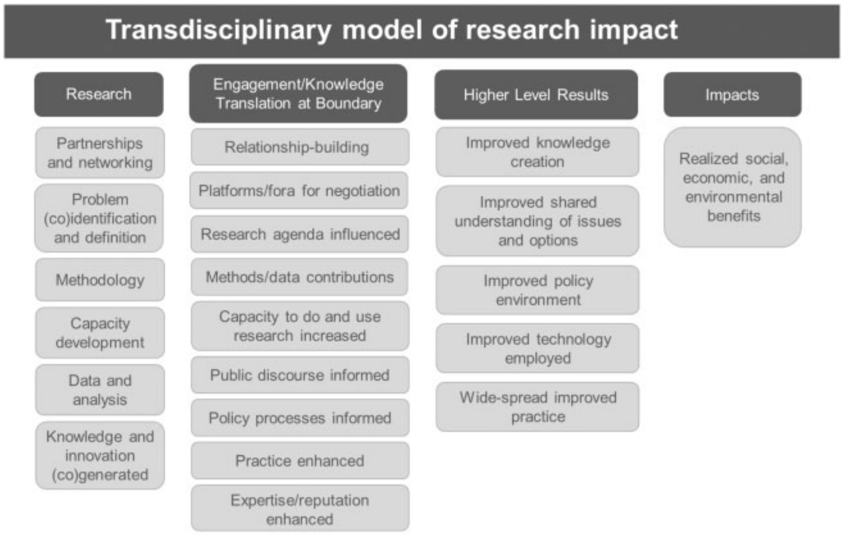
Transdisciplinary model of research impact
Research
Partnerships and networking
Problem (co)identification and definition
Methodology
Capacity development
Data and analysis
Knowledge and innovation (co)generated
Engagement/Knowledge Translation at Boundary
Relationship-building
Platforms/fora for negotiation
Research agenda influenced
Methods/data contributions
Capacity to do and use research increased
Public discourse informed
Policy processes informed
Practice enhanced
Expertise/reputation enhanced
Higher Level Results
Improved knowledge creation
Improved shared understanding of issues and options
Improved policy environment
Improved technology employed
Wide-spread improved practice
Impacts
Realized social, economic, and environmental benefits

Figure 1. TDR Contributions to Impact.

technologies, as well as the social and institutional structures for continual learning and adaptation (Smith, Stirling and Berkhout 2005; Miller et al. 2014).

This shift to more engaged, solution-oriented research, is evident at all levels, from heightened interest in TDR in graduate student research (Willetts and Mitchell 2016), university-led grand challenge programmes (Popowitz and Dorgelo 2018), and the United Nations Sustainable Development Goals (SDGs), which recognize the importance of synergies and trade-offs and the need to link physical, social, and natural capital. The key characteristics of use-inspired research are that research is problem-driven (as opposed to science-driven) and complexity-aware (as opposed to reductionist), and that it incorporates multiple sources of knowledge and knowledge co-production processes, supports decision-making (i.e. it does not just provide a technological solution), and pays attention to interactions between society and environment. In this mode, scientists not only generate new knowledge but also act as knowledge brokers and change agents (Miller, Muñoz-Erickson and Redman 2011). Interactive models acknowledge that societal actors other than scientists are important in creating science's societal impact and the mechanisms and pathways to impact increase manifold, changing the nature of the research questions, data, analyses, interpretations, and outputs. Spaapen and van Drooge (2011) use the concept of productive interactions as 'exchanges between researchers and stakeholders in which knowledge is produced and valued that is both scientifically robust and socially relevant' (212). In practical terms, productive interactions in research may mean capacity-building, co-learning, relationship-building, coalition-building, and multiplication of outreach beyond what would be possible in a classic disciplinary research project. In other words, the research *process* itself may generate benefits as much or even more than its final *products*. Moreover, the resulting impact on society may manifest at multiple levels simultaneously (e.g. from farmers' fields and local institutions through to sub-national and national policy formulation and implementation). The key point for the current discussion is that there are different modes of research, and in order for impact assessment to be reliable and useful, we need to be clear about the mode in which we are operating.

We now turn to the example of the FTA CRP for an illustration of how integrated, problem-oriented research has developed in practice. The example will also help consider specific needs, opportunities, challenges, and advances in research evaluation.

3. Evolution and evaluation of CGIAR research—Forest, trees and agroforestry case study

3.1 Technology roots

The CGIAR provides a good example of both an evolution (still incomplete) towards a more engaged, transdisciplinary sustainability-science approach, and the attendant evaluation challenges.

CGIAR research began in the early 1970s with a strong technology focus and an emphasis on crop-improvement (McCalla 2014). Major scientific advances in plant physiology, genetics, agronomy, and chemistry, as well as advances in plant breeding technologies, were packaged and delivered to users as improved seed and technology packages, resulting in remarkable increases in food production. However, it was quickly realized that technologies developed through research were often imperfectly suited to the priorities and circumstances of smallholder farmers. Despite overall increased yields, there were substantial gaps between research station potentials and realized yields in farmers' fields. Field practitioners recognized the need to better understand the constraints faced by smallholder farmers and their decision-making processes as a way to bridge the yield gap. 'Farming systems research' developed as a collection of methods for researchers to understand farm households and their decision-making beginning in the 1970s (Collinson 2000; Byerlee, Harrington and Winkelmann 2003).

However, even as CGIAR research expanded to a broader range of issues, the early successes of breeding programmes on crop production (especially semi-dwarf rice and wheat) and the impact of the green evolution on food security in India skewed donor interest towards crop breeding. As McCalla (2014) observes, promising systems research programmes were abandoned and converted to commodity-focused centres, and new ecology-oriented centres increasingly adopted commodity mandates as well.

After a relatively slow period of expansion, in the mid-90s, the CGIAR incorporated centres focused on NRM, including forestry [Center for International Forestry Research (CIFOR)], agroforestry [World Agroforestry (ICRAF)], water management (International Water Management Institute), and fisheries (World Fish), and a stronger eco-regional focus was implemented. Still, NRM research in the CGIAR has always had a strong emphasis on maintaining or increasing agricultural productivity and complementing CGIAR genetic improvement research. As Gregerson and Kelley (2007) observed, NRM research in the CGIAR 'is typically focused on producing knowledge that results in technology options, information and methods/processes that enhance [. . .] the productivity and stability of ecosystem resources' (13).

3.2 An evolving research model

External reviews of CGIAR social science (Barrett et al. 2009) and NRM research (ISPC 2012) advocated for more Sustainability Science and transdisciplinary approaches, and endorsed exploiting a wider range of interventions and impact pathways, with more emphasis on designing for impact with explicit theories of change (ToCs). Barrett et al. (2009) also explicitly recommended to '[f]ocus on impact but end the impact measurement obsession' (4). They recommended going beyond purely science-based partnerships to engage with government, civil society, stakeholders, and other relevant actors to help ensure that: research questions are relevant to development needs; values and concerns of intended users are represented in the research process; and pathways to impact are actively developed and supported.

An organizational reform process started in 2008 substantially accelerated the move towards engaged, solution-oriented research approaches. The shift was catalysed by an explicit commitment to 'shared responsibility' (ISPC 2015: 5) for impacts, defined as reduced poverty, improved food and nutrition security, and improved natural resources and ecosystem services (CGIAR 2016). This increased the accountability of research centres and individual researchers to realize social, economic, and environmental outcomes and impacts, on top of the long-established commitment to producing high-quality science.

A key aspect of the reform process was the creation of CRPs. These were intended to facilitate broader and deeper partnerships with other research organizations and with a range of policy and development actors. The focus on outcomes and emphasis on working through partnerships acknowledges that high-quality scientific knowledge creation alone cannot adequately address contemporary sustainable development challenges. The 2017 CGIAR Quality of Research for Development framework supports this approach by shifting from a traditional academic notion of science quality evaluation to a broader concept of research quality (discussed below) that is assessed on its potential and actual contributions to development processes (ISPC 2017).

3.3 Forests, trees and agroforestry consortium research programme

The FTA CRP is one of 15 CRPs within the CGIAR. It is led by CIFOR in partnership with ICRAF, Bioversity, and four non-CGIAR research organizations: Tropical Resources Institute (CATIE), the French Agricultural Research Centre for International Development (CIRAD), International Network for Bamboo and Rattan (INBAR), and Tropenbos (FTA n.d.). A large and increasing share of research performed by FTA (and within the CGIAR more generally) uses transdisciplinary approaches, engaging a range of scientists and societal stakeholders to frame the problem, co-produce applicable knowledge, and actively promote integration and application of research-based knowledge in complex systems.

FTA has developed a ToC for the entire programme, composed of five flagship 'projects' (FPs) on: (1) tree genetic resources; (2) forests and trees in livelihoods; (3) sustainable forest value chains; landscape dynamics, productivity, and resilience; and (5) climate change adaptation and mitigation (FTA n.d.). Each FP has its own ToC, and many individual projects within the FPs have explicit ToCs documented. At the programme level, key outputs are characterized as knowledge, tools, guidelines, models, and policy recommendations. Some FTA outputs are packaged as technological innovations, but few are truly discrete, stand-alone technologies analogous to an improved crop variety. Rather, most FPs model their work in a systems context and recognize that their efforts interact with multiple external actors and processes. The FPs are expected to work through targeted engagement with various actors in the system, contributing to capacity development of both researchers and research users. Many FTA research projects involve stakeholder engagement in one way or another. Co-generation of knowledge with various partners is intentional and considered a critical component of the research to impact pathway.

This integrated approach is exemplified in the R4D approach developed in FTA's FP2 on 'Enhancing how trees and forests contribute to smallholder livelihoods'. Coe, Sinclair and Barrios (2014) argue that sustainable increases in agricultural production and the maintenance of environmental services cannot be achieved simply by developing and promoting specific technologies. Rather, interventions need to adapt to fine-scale variation in social, economic, and ecological contexts. Furthermore, achieving benefits at the farm level will require appropriate service delivery mechanisms, markets, and appropriate institutions, along with technological innovation; all these aspects need to be addressed by research. Finally, the approach explicitly recognizes that scaling will require substantial effort beyond the capacity and reach of any research organization working alone. Coe, Sinclair and Barrios (2014) therefore recommend engaging actively and directly with development and private sector actors, what they call the 'development praxis' (73), to enable co-learning and scaling.

As an example, at the project scale, 'Support to the Development of Agroforestry Concessions in Peru' (SUCCESS) used stakeholder engagement and multiple impact pathways. SUCCESS aimed to support the implementation of a new tenure mechanism that offers agroforestry concessions (AFCs) to households as a way to realize positive ecological and socio-economic impacts. The project's multi-actor engagement approach aimed to develop smallholder knowledge and government capacities for AFCs, and build coalitions with key stakeholders to influence the political agenda. As illustrated in the SUCCESS ToC (Figure 2), the project worked to create space for dialogue, develop capacity and co-generate knowledge with smallholder farmers and government agents, and build coalitions among various actors in the system, in combination with more typical research collaborations and research-based knowledge outputs. The main institutional innovation was the AFC, but it was not developed by the research project. Rather, the project set out to help the AFC mechanism work more effectively.

The 'Global Comparative Study on Reducing Emissions from Deforestation and Forest Degradation' (GCS-REDD+) project focused research on identifying challenges and providing solutions to support the design and implementation of effective, efficient, and equitable policies and projects. The research involved more than 60 research partner organizations in 15 countries. Four main modules aimed to: (1) document and analyse relevant strategies, policies, and measures; (2) assess and learn lessons from sub-national REDD+ implementation (i.e. pilot projects); (3) analyse approaches to setting monitoring and reference levels as a contribution to the design of measurement, reporting, and verification standards; and (4) investigate potential synergies between REDD+ and climate change adaptation approaches. The programme was intended to contribute to improved policy and practice in sub-national REDD+ project implementation and at national and international policy levels. Each module involved a high degree of engagement with a range of partners. Contributions of the project were framed as changes in policy processes and practice, which resulted from both new knowledge and actions taken by various partners and stakeholders. The full ToC and a more elaborated explanation of the project outputs, engagement, and outcomes are presented in Young and Bird (2015). That evaluation recognized that, while overall REDD+ progress was limited by the international policy environment, there was evidence that the project positively influenced capacity and contributed to the discourse and development of improved systems for implementing REDD+ at international and national scales. Outcomes were achieved through: (1) the production and dissemination of high-quality independent research; (2) the development of approaches and tools that were applied by others; (3) provision of expert support at the international and national levels; (4) the hosting of international events and training; and (5) collaboration with and capacity development of national partners (Young and Bird 2015).

In projects of this kind, the research process itself generates value through partnerships and networking, identifying and defining the problem, methodological development, enhancing capacity, and otherwise influencing the research agenda. Each element can make valuable contributions independently or in combination with the data, analysis, and primary scientific knowledge generation process. As noted in a CGIAR review of NRM research (ISPC 2012), impact can result through: (1) new ways of thinking about land management and production systems change the paradigm of production; (2) decision-making and visualization platforms; (3) partnerships with innovators and entrepreneurs who are best placed to convert research outputs to practical application; and (4) coordinated and

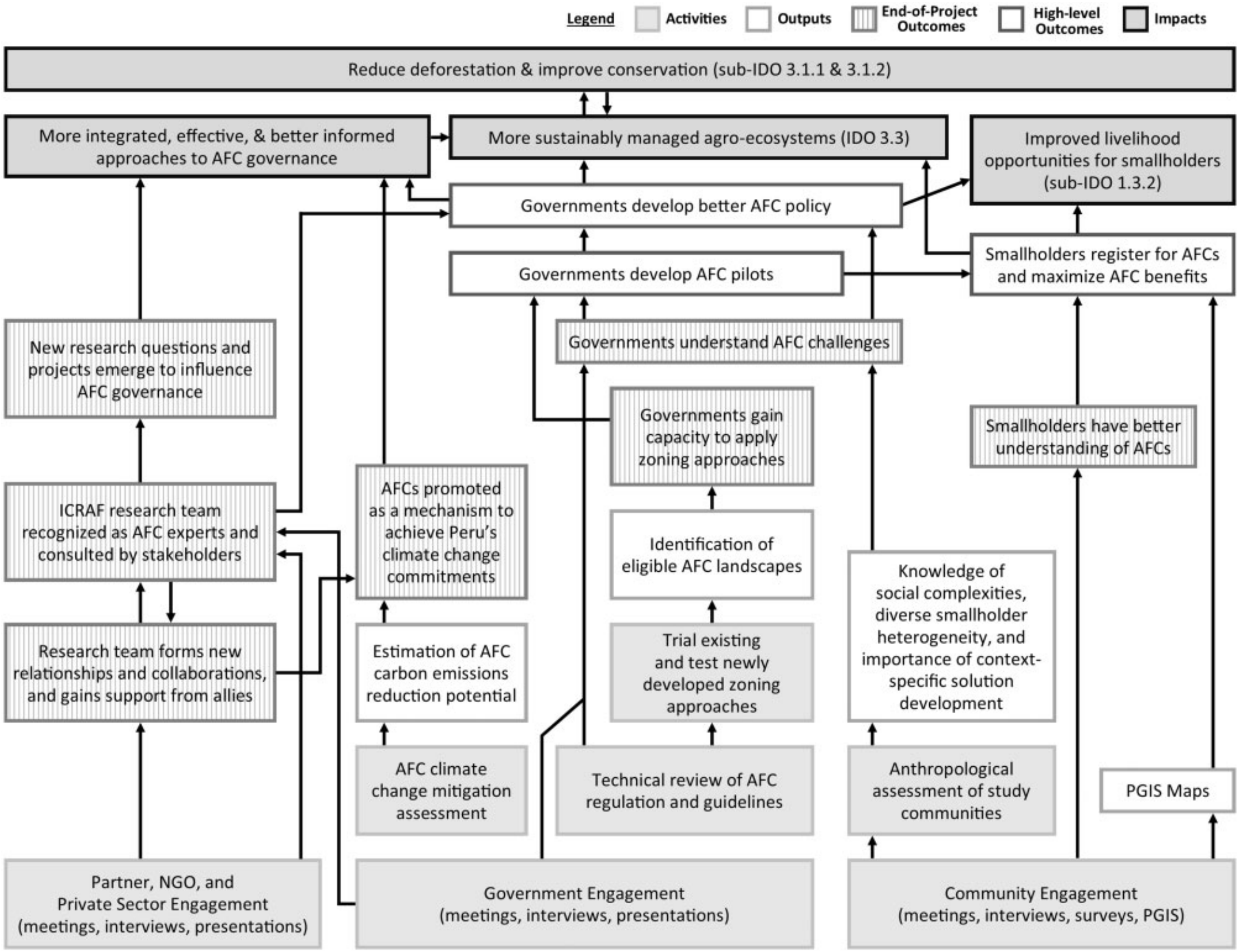
Legend
Activities
Outputs
End-of-Project Outcomes
High-level Outcomes
Impacts
Reduce deforestation & improve conservation (sub-IDO 3.1.1 & 3.1.2)
More integrated, effective, & better informed approaches to AFC governance
More sustainably managed agro-ecosystems (IDO 3.3)
Improved livelihood opportunities for smallholders (sub-IDO 1.3.2)
Governments develop better AFC policy
Smallholders register for AFCs and maximize AFC benefits
Governments develop AFC pilots
New research questions and projects emerge to influence AFC governance
Governments understand AFC challenges
Governments gain capacity to apply zoning approaches
Smallholders have better understanding of AFCs
ICRAF research team recognized as AFC experts and consulted by stakeholders
AFCs promoted as a mechanism to achieve Peru's climate change commitments
Identification of eligible AFC landscapes
Knowledge of social complexities, diverse smallholder heterogeneity, and importance of context-specific solution development
Research team forms new relationships and collaborations, and gains support from allies
Estimation of AFC carbon emissions reduction potential
Trial existing and test newly developed zoning approaches
AFC climate change mitigation assessment
Technical review of AFC regulation and guidelines
Anthropological assessment of study communities
PGIS Maps
Partner, NGO, and Private Sector Engagement (meetings, interviews, presentations)
Government Engagement (meetings, interviews, presentations)
Community Engagement (meetings, interviews, surveys, PGIS)

Figure 2. Simplified SUCCESS Project ToC.

widely accessible metadata. Impact pathways involve changing the context and informing and influencing transformative change processes.

3.4 Challenges for learning and impact assessment

This brief overview of the FTA research programme reveals several characteristics that confound learning and impact assessment in this and other integrated sustainability science and TDR programmes.

Spatial scale

The programme works at multiple spatial and temporal scales. Research is being conducted at the scale of genes, farm- and forest-scale management, and sub-national, national, and international policy and trade.

Time scale and time lags

Some work focuses on immediate problem-solving and some deals with long time scales. In many cases, processes initiated and/or supported by FTA research take time to mature and often require changes in the context in order to be fully realized.

Multiple interventions, multiple impact pathways

The research is being done within complex systems, with many other actors and processes operating simultaneously. As discussed above, the research aims to contribute through capacity-building and empowerment of various systems actors, relationship-building, methods development, problem definition, and agenda-setting, as well as science-based knowledge creation. It would be unrealistic to try to isolate pathways, which would miss important contributions.

Co-generated knowledge

Related to the multiple impact pathways, FTA research increasingly engages with, supports, and empowers other actors to do their work.

4. Evolving concepts and measures of impact

We have discussed how research has evolved to be more engaged, transdisciplinary, and change-oriented, and likewise how the CGIAR and its research have evolved to be more effective and impactful. But how do we assess its impact? Clearly, if the nature of research changes, so must the way we conceptualize and measure the impact of research.

The term ‘impact’ itself is poorly defined and ambiguously used (Belcher and Palenberg 2018). Most definitions of impact related to research implicitly assume a linear model of impact but focus on different parts of the model. ‘Academic impact’ or ‘research impact’ normally refers to the intellectual contribution made to a field of study within academia. There has also been a strong push to recognize and evidence the ‘societal impact’ of research. Many research funders ask for some indication of potential societal impact in grant applications. Societal impact typically refers to ‘an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia’ (REF 2011: 48). CGIAR donors tend to think of impact in terms of concrete realized benefits in improved human welfare or environmental conditions. When they ask for demonstrations of impact, they are often asking for tangible realized benefits: evidence that the world has been improved in some measurable way. They expect research impact assessment to demonstrate significant contributions to desired development goals and validate large-scale effects at the mission level in terms of reduced poverty, improved food security, and/or improved natural resource conditions (Raitzer and Winkel 2005).

Ex post impact assessment has been presented as critical to satisfy the accountability imperative for a publicly funded institution and essential for continued support by investors to the CGIAR (Kelley, Ryan and Gregersen 2008). There is a perennial promise that, if researchers can prove their impact, funders will provide more unrestricted support. As McCalla (2014) discusses, this promise has never been realized. However, as Raitzer and Winkel (2005) discuss, impact assessment of agricultural research has made little reference to actual demands from any audience group, and there has not been any subsequent systematic assessment of donor needs and expectations regarding impact assessment (Stevenson, personal communication).

Raitzer and Winkel’s (2005) survey reported that donors were primarily interested in demonstrating the contribution of research to development goals and validating large-scale effects at the mission level to justify and defend funding decisions to higher decision-making bodies. Somewhat surprisingly, respondents were not especially interested in trying to attribute credit among collaborating institutions and rather felt it appropriate that relevant contributions and investments should be considered in concert. Indeed, Raitzer and Winkel (2005) report that some respondents felt non-CGIAR influences were insufficiently credited in CGIAR impact studies. Respondents expressed interest both in large-scale estimates of adoption and productivity effects, as well as smaller-scale analysis of detailed effects at the household level.

As discussed above, since 2005, there have been many changes in the CGIAR and in the nature of the research performed by the CGIAR. There has also been a greater learning focus in the evaluation field (Patton 2008). Indeed, considering the complexity and multiple impact pathways of R4D and TDR more generally, it is clear that there is a pressing need to learn what works (and what does not work) and how to facilitate continuous improvement.

4.1 Impact assessment in the CGIAR

The current predominant approach to impact assessment in the CGIAR uses a counterfactual framework with experimental or quasi-experimental methods (Stevenson, Macours and Gollin 2018b). In the former, units (e.g. individuals, households, villages) are randomly assigned to different ‘treatment conditions’ and the resulting treatment groups are compared statistically, typically against pre-specified outcome or impact indicators. If random assignment is not feasible or appropriate, quasi-experimental approaches use various alternative strategies (e.g. difference-in-differences, regression discontinuity, and propensity score matching) to control for selection bias in the comparison of treated and untreated units (Khandker, Koolwal and Samad 2010). Both approaches aim to estimate statistically aggregated effects of a hypothesized cause (e.g. an intervention or improved crop variety). There has been a long and storied debate about the pros and cons of randomized control trials (RCTs), as well as quasi-experimental and other quantitative impact assessment approaches (Donaldson 2009; Frieden 2017; Deaton and Cartwright 2018). We do not intend to enter that debate. We fully appreciate that, used appropriately, experimental and quasi-experimental impact assessment methods can and have contributed substantially to empirical impact assessment work. However, these methods are insufficient on their own, and often inappropriate in the context of problem-oriented TDR.

First, as already discussed, TDR research is emergent and not discrete. Experimental and quasi-experimental approaches are only appropriate when there is a well-specified and discrete treatment. If the treatment itself is multi-pronged, evolving (e.g. if the research programme is co-generating knowledge, enabling and supporting various stakeholders and influencing the policy discourse), and/or under-specified (e.g. if the full range of influences is emergent and not fully known), there will be uncertainty as to which variation of the emergent research-informed innovation is responsible and for which specific effects. This limits scope for learning and for determining what should be scaled out further (Veerman and van Yperen 2007). Each TDR case is unique, and we need nuanced understanding of context to be able to understand processes of change.

Second, and more fundamental, experimental and quasi-experimental approaches are only appropriate for ‘large n’ interventions; that is, interventions that target sufficiently large numbers of units where it is possible—at least theoretically—to assign units (ideally at random) to varying treatment conditions (White 2010). Outside of laboratory or research station settings, the number of units that need to be assigned to such conditions is typically large. This is due to the inevitable increase in ‘statistical noise’ resulting from both the greater heterogeneity among such units and an inability to isolate them from extraneous factors that additionally influence the status of the outcome and impact indicators of interest. As discussed above, while policy-oriented research programmes may aim to generate benefits for many units (e.g. farming households), they typically work through a small number of units (e.g. a national climate policy; a regional land use planning framework; or an agri-food system) as a way of inducing such ‘large n’ change. They are ‘small n’ interventions, where generating statistically aggregated ‘treatment effect estimates’ is implausible and inappropriate. As we elaborate below, a more appropriate approach is needed to analyse and evidence the extent to which the ‘small n’ unit in question has changed and the likely factors responsible for that change.

Moreover, interactions are expected between targeted ‘small n’ outcomes (e.g. the mitigation of policy constraints) and ‘large n’ impacts (e.g. smallholder farmer income), thereby potentially violating what economists call the stable unit treatment value assumption (SUTVA) (Rubin 2005). SUTVA is violated and, by extension, the ability to precisely estimate the counterfactual when there are spillover effects between treated and untreated groups or equilibrium effects (e.g. water pollution) affecting both.

Finally, an inherent feature of TDR is the co-generation of new knowledge, learning, and/or improved capacities together with intended research users and other stakeholders. It follows that any resulting benefits for society and/or the environment can neither be solely attributed to the researcher, nor the research investment in question.

These same limitations may also apply to discrete research-informed innovations, such as improved crop varieties. Here too, greater emphasis is being placed on stakeholder engagement in research and technology development processes, such as farmer participatory varietal selection (see Joshi and Witcombe 1996). In this kind of research, technological innovation itself may be just one of many factors responsible for any realized benefits. This is nicely illustrated by Bulte et al. (2014) who conducted both traditional and double-blind RCTs of an intervention promoting improved cowpea varieties in Tanzania. The traditional RCT estimated a 27% average gain in yields over traditional varieties. However, the double-blind RCT estimated that two-thirds of the average increase was due to a placebo effect. That is, the farmers in the traditional RCT were told that they were receiving improved chickpea varieties and consequently invested relatively more in the management and care of the ‘improved’ crops.

Indeed, there is concern that the current bias towards quantitative approaches, most notably RCTs, is steering the evaluation agenda—and the research agenda—away from potentially impactful interventions that cannot easily be randomized (Deaton 2010). Again, the CGIAR provides a good example. In NRM research, there are difficult challenges in isolating lines of causality, attributing impacts to particular interventions, estimating meaningful counterfactuals, and establishing quantitative measures (Kelley, Ryan and Gregersen 2008). As a result, several observers have lamented the lack of evidence of NRM impact. Renkow and Byerlee (2010) found that NRM research had not shown the same returns on investment as crop genetic improvement research. They further suggested that investments in NRM research in the CGIAR should be reduced relative to crop genetic improvement. Renkow and Byerlee (2010) reviewed the impact of policy-oriented research in the CGIAR and found that analyses tended to be confined to documenting impact pathways as opposed to measuring specific impacts. They acknowledge ‘NRM work typically deals with systems, rather than components to a greater degree than other types of CGIAR research’ (397), which they suggest increases the location specificity of NRM research and probably limits the international public good dimensions. In fact, NRM research conducted to understand processes and interactions in systems and extrapolation domains has the same potential to yield global public goods as crop improvement research. When NRM processes are understood, extrapolation is possible.

A World Bank meta-evaluation of the CGIAR (Lesser 2003) also found that NRM research was under-evaluated and another review highlighted the lack of evidence of contributions of NRM research to impact at scale (ISPC 2012). The ISPC (2012) review recommended that it is ‘necessary and legitimate to pursue research [. . .] to develop new methods for impact assessment that recognize the contributions of NRM research’ (6). A series of studies on NRM research outcomes between 2013 and 2016 noted low rates of adoption of NRM technologies and practices and a general recognition of the lack of a clear and compelling vision for NRM research as a whole (Stevenson and Vlek 2018).

It is notable that many of these previous efforts followed a technology adoption model. Renkow (2010) posits a simplistic two-phase impact pathway. In the first phase, research outputs combine with political inputs to produce policy outcomes such as new laws, regulations, and institutions (with the implicit assumption that simply making new knowledge available will trigger policy change); in the second phase, those policy outcomes produce welfare changes (i.e. impacts). Moreover, the NRM research examples reviewed in the Standing Panel on Impact Assessment (SPIA) case studies were all technology-based interventions (Stevenson and Vlek 2018). Stevenson, Macours and Gollin’s (2018b) report, ‘The Rigor Revolution in Impact Assessment: Implications for CGIAR’, also focuses predominantly on assessing technological innovation. The report acknowledges that there are multiple impact pathways for CGIAR research, and it offers brief discussion of the need for methodological pluralism. However, the second main point in the conclusion identifies the need for work in this area:

> ‘impact evaluation and efficacy studies need to focus on causal relationships for which we have the greatest uncertainty and for which information would have the highest value. This suggests a greater focus on theory—away from searching for ‘what works’ in the abstract and towards finding out why certain things work and others do not in particular contexts [. . .] It is less obvious how to make methodological breakthroughs on tracing policy influence or measuring the outcomes from capacity-building efforts, though the principle of independent theory-based evaluation should be prominent’ (Stevenson, Macours and Gollin 2018b: 29).

In another discussion on estimating impacts of R4D, the authors state: ‘[e]ven if one starts from the viewpoint that contributions to science and to capacity from the research process itself will be excluded, the challenges associated with estimating benefits from a research-based technology or other innovation are enormous’ (Stevenson, Macours and Gollin 2018a: 3). They go on to point out that it is important to recognize these other kinds of contributions, but do not explore this important question further.

5. Towards an integrated, mixed methods approach for evidencing R4D impact

5.1 Principles for evaluating multi-faceted TDR initiatives intervening in complex systems

How then do we conceptualize and assess the impact of emergent, stakeholder-informed, and systems-focused research? We assume that few would disagree that problem-oriented TDR has potential to significantly contribute to solving some of the world’s most pressing challenges, and, in turn, positively impact society and the environment. We also assume general agreement that research funders and the public more generally have a right to reliable feedback on what this contribution is and how it can be strengthened. Therefore, unless we want to restrict research to narrowly focus on developing and improving ‘large n’ technologies and move away from seeking to influence change in complex systems, expectations for what counts as credible evidence needs to be considerably broadened. The inevitable complexity and limitations also need to be explicitly recognized. We propose a set of evaluation principles to guide the evaluation of TDR programmes.

1. *Use a portfolio of methods.* A multi-scale, transdisciplinary research programme will generate multiple kinds of outputs and pursue multiple impact pathways. It is important to assess key elements using the most suitable methods. A portfolio of multi-method and multi-level evaluative work is needed for most TDR programmes. Theory-based qualitative approaches (discussed below) can be used to assess outcomes of interactive work with stakeholders seeking to improve policy and/or practice. Some research-based innovations will be suitable for testing using conventional 'large n' impact assessment approaches. A policy implementation assessment may be useful to understand the extent to which the co-generated policy recommendations were implemented. *Ex ante* modelling approaches can further help estimate the range of likely impacts associated with such implementation (Kelley, Ryan and Gregersen 2008). A programme should aim to build a broad portfolio of evaluative evidence over time.
2. *Focus on central aspects (key 'nodes') of the overarching theory of change.* The portfolio of evaluative inquiry should cover the primary components or 'nodes' of the overall (project or programme) ToC. The ToC nodes may relate to specific kinds of outcomes (e.g. changes in stakeholder capacity or aspirations) and actual impacts on the ground (e.g. reduced deforestation, improved household income). They also may encompass change at different scales and/or be structured around defined research areas. Given that these different components of the research programme are expected to work together to contribute to systems-level change, evaluative efforts should include a focus on interactions.
3. *Aim for representativeness.* The overall set of assessments should aim to identify and select sets of representative cases and contexts as a basis for learning from the range of experience and for extrapolation of results.
4. *Prioritize quality over quantity.* It is better to have fewer high-quality evaluations than numerous evaluations of dubious quality. As argued by White and Phillips (2012), qualitative 'small n' evaluations need to be implemented following rigorous qualitative research protocols to be able to successfully address issues of cause and effect. While there is debate on the relevance of independence in evaluation (Picciotto 2013), it is critical that the evaluation team be as impartial as possible, mitigate potential sources of bias in data collection, and ensure the veracity of key findings (e.g. via triangulation and/or the identification of 'signatures'). Attention will also be needed to improve the application and specification of theoretical assumptions in research project design and within evaluations. This will help frame sets of evaluations (i.e. series of case studies) to improve their explanatory power. Careful and transparent evaluation design, data collection, analysis, and reporting, along with improved peer-review and other quality control efforts, can help bolster credibility.
5. *Build evaluability and evaluation into research projects where possible.* Research projects can be designed to facilitate evaluation and generate evidence for impact assessment. For example, explicitly documenting a project/programme ToC at inception supports strategic planning and, not incidentally, provides a framework for monitoring, data collection, and outcome evaluation. Experimentally testing the impact of policy or technology options at the project scale, as part of the research process, can generate evidence to estimate impact when such options are implemented at scale, bearing in mind the inherent external validity considerations. Participatory policy option development with stakeholders can help explicate underlying assumptions, which can be used in *ex ante* impact assessment. The more creatively research can fulfil its own objectives, while generating evidence of its own impact, the better.
6. *Look for unintended consequences.* Intervening in complex systems is likely to generate unforeseen outcomes and impacts, both positive and negative. Support provided to a regional government, for example, to strengthen its land use planning and management system may negatively impact a particular user group in ways unforeseen in the research engagement's earlier phases. Evaluations of specific aspects of the research programme and those examining interconnections should deliberately seek to identify unintended consequences and investigate them when they are found.
7. *Facilitate and document learning for enhanced research effectiveness.* As indicated above, evaluating a complex TDR programme is unlikely to be a one-off exercise. Rather, several sources of evaluative evidence, together with relevant impact-related research evidence, will be combined to paint an overall picture. Each of these evaluation and relevant research pieces should include (ideally co-generated) recommendations for strengthening research impact. These recommendations should then be taken on board, formally documented, and ideally highlighted in one or more overarching evaluative pieces. Research programmes that learn from failure and continuously seek to strengthen the modalities through which they generate societal impact are intrinsically valuable and should be attractive from an accountability perspective as well.
8. *Acknowledge (and embrace) the inherent limitations.* Outcomes and impacts may be realized at multiple levels, likely with varying time-lags and uncertain impact trajectories. Many potential benefits may be difficult to anticipate *a priori*, and some may not be traceable at all. High-level outcomes and impacts in complex systems are beyond the control or even influence of most research programmes. Change processes can be modelled and monitored, with evidence of contributions to the process, but high-level impacts can rarely be attributed directly to the research. Therefore, expectations for precise and comprehensive impact measurement should be tempered.

6. Impact assessment strategy framework for an integrated TDR programme

Figure 3 illustrates a stylized set of TDR impact pathways. A research project/programme aims to create new knowledge and/or innovations to address development problems, but the process can (and in TDR, is designed to) also make substantial contributions in problem framing, methodology, collection and sharing of data, etc. This work, up to and including outputs, is within the sphere control of the project (Hearn 2010). Each of these processes and outputs is expected to support, encourage, or otherwise influence other actors by changing knowledge, attitudes, skills, and/or relationships, resulting in changes in policy and/or practice (sphere of influence) that ultimately contribute to impacts in terms of human and/or environmental social condition. Here we set out the elements of an impact assessment framework for a TDR programme.

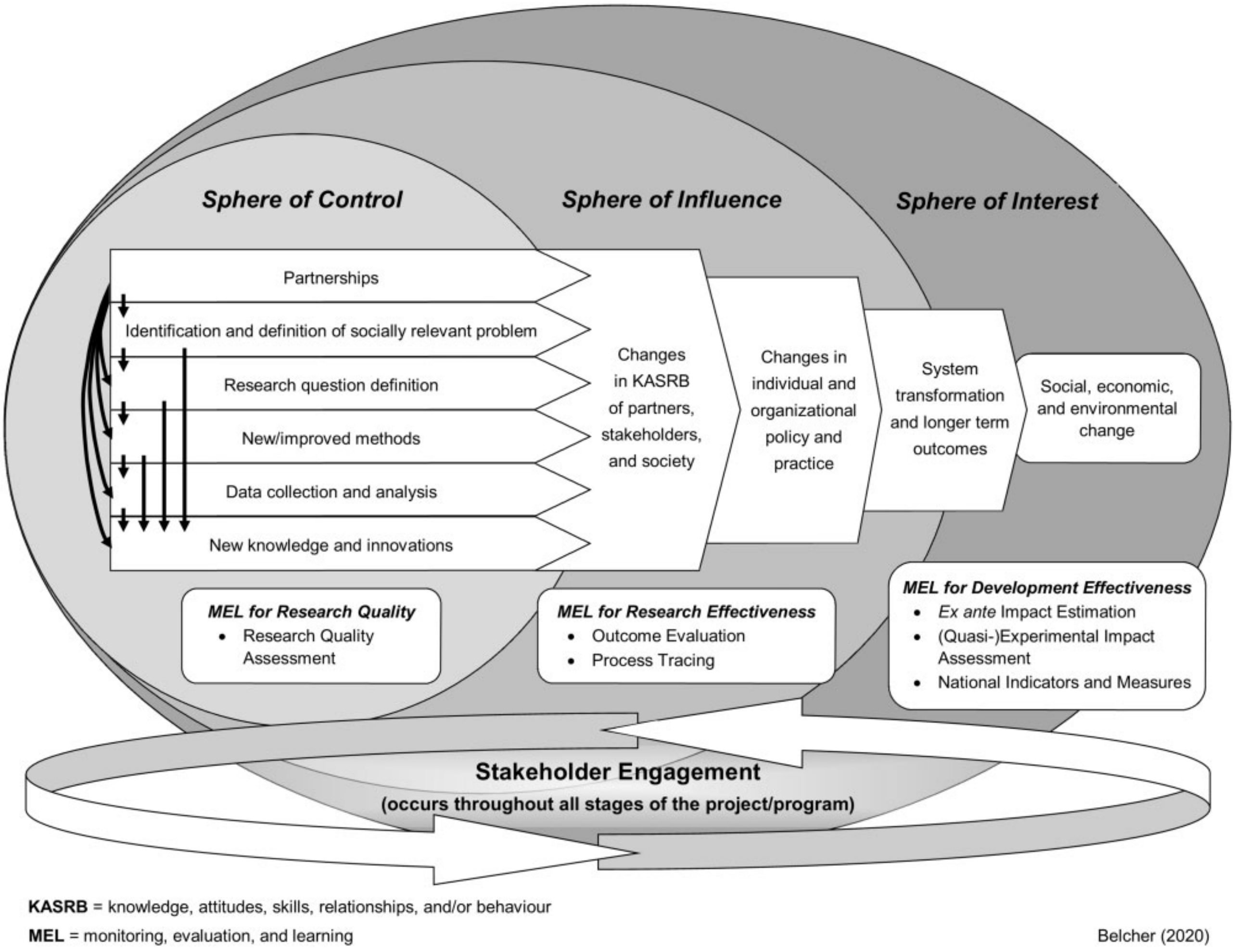
Sphere of Control
Sphere of Influence
Sphere of Interest
Partnerships
Identification and definition of socially relevant problem
Research question definition
New/improved methods
Data collection and analysis
New knowledge and innovations
Changes in KASRB of partners, stakeholders, and society
Changes in individual and organizational policy and practice
System transformation and longer term outcomes
Social, economic, and environmental change
MEL for Research Quality
• Research Quality Assessment
MEL for Research Effectiveness
• Outcome Evaluation
• Process Tracing
MEL for Development Effectiveness
• Ex ante Impact Estimation
• (Quasi-)Experimental Impact Assessment
• National Indicators and Measures
Stakeholder Engagement
(occurs throughout all stages of the project/program)
KASRB = knowledge, attitudes, skills, relationships, and/or behaviour
MEL = monitoring, evaluation, and learning
Belcher (2020)

Figure 3. Research-to-Impact Pathways.

6.1 Challenge framing

Integrated TDR programmes, like FTA, are particularly challenging to evaluate. Not only does the programme embody the characteristics of problem-oriented TDR described above, it comprises five distinct research areas through its FPs. Each FP includes multiple projects, most of which are funded bilaterally, with donor priorities inevitably constraining strategic alignment. FTA is effectively an umbrella for several relatively distinct pathways, rather than a single initiative. The same is true of Grand Challenge Programs and other large inter- and transdisciplinary programmes. Nevertheless, these programmes are expected to ultimately contribute to mission-level development impacts. In FTA's case, these are specified as global targets in terms of reduced deforestation and forest degradation; reduced rural poverty and inequality; reduced loss of biodiversity and ecosystem services; and improved land-use governance and management. Evidence is needed to show that the programme has contributed to these targets.

6.2 Nested ToCs

A fundamental assumption of an integrated programme is that its individual components are strategic and coordinated in a way that will contribute collectively and significantly to high-level outcomes and impacts. Planning and evaluation can be greatly facilitated by developing and documenting the ToC for the overall programme to model the key actors, activities, outputs, outcomes and intended impacts. A ToC at the programme scale needs to encompass the full range of these different elements, so detail is necessarily limited. It is, therefore, useful to have more detailed ToCs at sub-programme scale (e.g. sets of closely-linked projects) and project scale. At the project scale, it is possible to precisely specify outcomes to guide planning and evaluation (discussed below). The aim is to develop a systematic, integrated, and nested ToC and evaluation framework as a base for other evaluative work.

6.3 Research quality appraisals

Research quality can be assessed within the sphere of control. Quality is broadly defined to include characteristics of design and implementation that will achieve outputs that are relevant, credible, legitimate, and effective (ISPC 2017). *Relevance* refers to the appropriateness of the problem positioning, objectives, and approach to the research for intended users. *Credibility* pertains to rigour of the design and research process to produce dependable and defensible conclusions. *Legitimacy* refers to the perceived fairness and representativeness of the research process. *Effectiveness* refers to the utility and actionability of the research's knowledge and social process contributions. Belcher et al. (2016) provide a set of criteria for TDR evaluation, organized within these four principles.

6.4 *Ex ante* impact estimation

A ToC provides a good base for *ex ante* impact assessment to estimate potential impacts. *Ex ante* impact assessment can be done at any scale, with trade-offs between precision and scope. Ultimately, information about mission-level impact for each of the main challenges is needed. Such estimates will use evidence and information from a combination of methods (discussed below) to estimate plausible ranges of FTA's impact vis-à-vis the above targets, as well as other potential impacts, including possible negative impacts. *Ex ante* assessment will necessarily require application of theory and assumptions. Making the theory and assumptions transparent and therefore open to challenge and empirical testing is indeed one of the main benefits of the process. Ideally, sensitivity analysis should be done to assess the influence of key assumptions on impact estimates.

6.5 Theory-based outcome evaluation

Much of the FTA research portfolio aims to exploit multiple processes in complex systems. Here we need project-scale theory-based approaches to test whether and how research has contributed to a change process. Theory-based approaches are well suited to 'within-case' research and involve analysing 'the causal links that connect independent variables and outcomes, by identifying the intervening causal processes, that is, the causal chain and causal mechanisms linking them' (Reilly 2010: 734). A key aspect of this approach involves ruling out alternative explanations for an observed event or change and identifying indicators or 'signatures' (Mohr 1999: 72) that, if they occur, provide support for a hypothesized cause. Such approaches work through affirming explanations that are consistent with the facts and rejecting those that are not.

The theory can be derived inductively, working backwards from observed changes and developing explanations. This is the approach used in methods such as Realistic Evaluation (Pawson and Tilley 1997; Maxwell 2004; George and Bennett 2007) and Process Tracing (Collier 2011). In these approaches, the evaluator develops theories about how the project works to generate particular effects. Emphasis is placed on understanding the nature of the project and its operation. Theories about the mechanisms and circumstances by which the project or programme has contributed to effects for specific subgroups in particular contexts are then iteratively developed and tested.

The ToC can also be developed *ex ante*, as part of a project or programme's design, as a set of hypotheses about what outcomes and impacts are intended or expected to result in a particular case. The hypotheses can then be tested deductively using empirical evidence from the completed project. This approach is used in methods such as Outcome Mapping (Earl, Carden and Smutylo 2001), Payback Framework (Buxton and Hanney 1996), and Contribution Analysis (Mayne 2001, 2012).

FTA has adopted and refined a theory-based case-study approach specifically for assessing the outcomes of research at the project scale (see Belcher, Davel and Claus 2020). The method uses a detailed project ToC as the analytical framework (Weiss 1997; Coryn et al. 2011; Vogel 2012; Belcher, Davel and Claus 2020). The ToC models the change process, providing description and explanation of both how and why the project is expected to cause or contribute to a result or a set of results (i.e. outputs, outcomes, and impacts). It details the primary actors, steps, and pathways in the change process and specifies the theoretical reasons for the changes. A well-specified ToC is essentially a set of hypotheses about each step in the change process that can be tested empirically using data from document review, surveys, and interviews with key informants to assess actual outcomes against expected outcomes at each stage in the ToC. In lieu of a reliable counterfactual, it is important to consider and test competing hypotheses for how a change may have happened, and leave room to include additional elements and alternative explanations for how the project or programme may or may not have affected the outcome in question based on *ex post* analysis (Rossi and Freeman 1989; Donaldson 2009; White 2009; Hitchcock 2018).

Outcome evaluations provide evidence of the scope and scale of qualitative changes and change processes in the overall effort to address programme challenges. They answer the question: who is doing what differently as a result of the research?

6.6 Experimental/quasi-experimental IA

As explained above, experimental and quasi-experimental approaches are useful for assessing the effectiveness of 'large n' innovations. Some research outputs will be delivered as distinct technologies or institutional innovations applied over a large number of units (e.g. smallholder farming households). In these cases, experimental and quasi-experimental impact assessment can be used.

As much as possible and where relevant and feasible, such impact assessment work should be incorporated as part of the research process itself. That is, if the research involves the use of a discrete intervention in multiple iterations, it may be possible to randomly assign the treatment and compare both before and after and with and without the treatment (experimental). If true random assignment is not feasible, quasi-experimental approaches, such as those mentioned above, can potentially be used to control for selection bias in the comparison of treated and untreated units.

For example, a research project is currently underway to experimentally compare alternative extension approaches as part of an effort to enhance the implementation of Peru's AFC policy instrument (discussed above). While this pilot will focus on evaluating the effects of the extension approaches on the uptake of sustainable land management practices, efforts will also be made to estimate the impacts on farming household income and deforestation. In other cases, as necessary and appropriate, stand-alone quasi-experimental IA will be needed as part of the overall impact evidencing strategy. The resulting evidence can then be used to estimate impact when the innovation in question is promoted at scale, following relevant changes in policy and practice.

6.7 High-level indicators and measures

If we can demonstrate that the research has been successful at stimulating or contributing to change within the sphere of influence, it is reasonable to expect further knock-on changes. That is, if key actors act differently as a result or partially as a result of the research project, that may contribute to further changes that will help transform systems and ultimately lead to social, economic, and environmental benefits.

A key element of the Outcome Evaluation approach is the explicit identification of end-of-project outcomes, defined as outcomes that would be ambitious but reasonable to expect and to observe within the time-frame and resources of the project being evaluated (Belcher, Davel and Claus 2020). End-of-project outcomes can be assessed empirically.

Higher-level (i.e. beyond end-of-project) outcomes and impacts are also modelled in the ToCs, and in *ex ante* impact assessments, to illustrate and explain the causal logic to mission-level impact. However, these kinds of changes are typically well outside the sphere of control and sphere of influence of a research project or programme. Changes in the sphere of interest may be indicated by United Nations SDG indicators and other secondary data on development and conservation status, but it is not possible to attribute those changes to research because there are so many other factors that affect their status.

6.8 Aggregating up

The overall strategy needs to identify and focus on the key nodes (Principle 2) and use appropriate, rigorous methods (Principles 1, 4, and 6) to evaluate representative sets of projects/contexts (Principle 3). Within the sphere of control, the focus needs to be on research quality, with quality broadly defined to include characteristics of design and implementation that will achieve outputs that are relevant, credible, legitimate, and effective (ISPC 2017). Within the sphere of influence, we need to know whether and how a research programme (including knowledge creation and supporting activities) is actually encouraging, supporting, or otherwise influencing key actors in the system to make positive changes in policy and practice. In the sphere of interest, we can model impact pathways and estimate potential reach and significance, and monitor secondary data on social, economic, and environmental benefits. However, in complex systems, with multiple actors, processes, and time-lags, it is theoretically impossible to make definitive attribution claims (Principle 8). Instead, a plausible case can be made that research has contributed if we can demonstrate that: (1) there is a strong theoretical basis to expect that the research programme will contribute to high-level outcomes and impacts; (2) the research has produced good quality (relevant, credible, legitimate, and effective) outputs; and (3) expected intermediate and end-of-project outcomes (e.g. changes in policy and practice) have been realized.

7. Conclusion

We have explored the evolution of engaged TDR approaches and the inherent challenges of measuring (or estimating) the impact of TDR and R4D more broadly. Our central argument is that multifaceted research initiatives seeking to intervene and bring about positive change in complex systems cannot be treated like discrete 'large n' interventions. TDR programmes are dynamic and have high potential to target multiple aspects of a problem or issue simultaneously. Relying on conventional quantitative impact assessment approaches alone, or even as the primary mode of evaluative inquiry, is therefore inappropriate. There are other methods that are suitable for interrogating cause and effect relationships in key areas of such research, most notably mechanism or explanatory approaches.

Inevitably, however, there are inherent limitations in measuring TDR impact; this needs to be recognized and accepted. Above all, we need to avoid a potentially perverse outcome where the ambitions of researchers, development practitioners, policy-makers, and others aiming to achieve transformational change are stifled by the results agenda. Contemporary and urgent global challenges are too vast and complex for piecemeal solutions.

However, we cannot use the inherent measurement challenges as an excuse to not evaluate, learn, or be held accountable. We therefore advocate for a holistic, multi-method, and integrated approach; one that is appropriate for the nature of a TDR programme, and one that does not rely on any single evaluation method or static framework. To this end, we offer eight intuitive principles to guide the development of evaluation strategies for larger multi-faceted TDR programmes and present a framework for an integrated TDR programme. We hope that these prove useful to not only researchers, research managers, and evaluators but also research investors.

Theory-based evaluation approaches are still being developed and, especially in the R4D context, there is considerable scope for further methodological improvement. As we gain experience with the methods, it will be possible to streamline and improve efficiency and lower costs. As noted above, increased emphasis on explicit ToC development and documentation as part of research project and programme design, along with improved monitoring, will reduce the costs and increase the rigour of *ex post* evaluations. Improved application and specification of theoretical assumptions in project design and evaluations will help frame sets of evaluations (i.e. series of case studies) to improve explanatory power. Careful and transparent evaluation design, data collection, analysis, and reporting, along with improved peer-review and other quality control efforts can help address credibility concerns. Integration of theory-based evaluation within a portfolio of methods, taking advantage of the strengths and compensating for the weaknesses of each approach, can help strengthen both accountability and learning in the research programme. This will involve: nested ToCs (project and programme scales); research quality appraisals to check that the research is addressing the 'right' issues in the 'right' ways; experimental and quasi-experimental impact assessments to both support and evidence the effectiveness of 'large n' innovations addressing specific dimensions of the challenge in question; project and programme level theory-based outcome evaluations to test ToCs and assess contributions to proximate outcomes; and modelling (including *ex ante* impact assessment) and extrapolation to estimate mission-level impacts.

Acknowledgements

This article was developed from a paper presented at a meeting of the Independent Steering Committee of the Forests, Trees and Agroforestry (FTA) CGIAR Research Program. Anne-Marie Izac, Vincent Gitz, and Florencia Montagnini provided helpful feedback on that paper. The current version was improved based on insightful comments and suggestions from Keith Child and Nancy Johnson. Rachel Davel provided editing support to the paper and manuscript. We thank the anonymous reviewers for their helpful feedback to improve the manuscript.

Funding

This work was supported by funding from the Canada Research Chairs programme and the Canadian Social Sciences and Humanities Research Council (SSHRC).

Conflict of interest statement. None declared.

References

Andrews, M., Pritchett, L., and Woolcock, M. (2013) 'Escaping Capability Traps through Problem Driven Iterative Adaptation (PDIA)', *World Development*, 51, 234–44.

Barrett, C. B. et al. (2009) *Stripe Review of Social Sciences in the CGIAR*. Rome: CGIAR Science Council Secretariat. <http://barrett.dyson.cornell.edu/files/papers/SocialScienceStripeReview_Aug2009.pdf>

Belcher, B. M. (2020). *Research for Changemaking: Concepts and Lessons for Research Effectiveness. Keynote Presentation to Canadian Changemaker Education Research Forum*. Toronto, Canada. <https://researcheffectiveness.ca/wp-content/uploads/sites/7/2020/01/Belcher-C-CERF-Keynote-Jan-15-2020.pptx> accessed 15 Jan 2020.

Belcher, B. M. et al. (2016) 'Defining and Assessing Research Quality in a Transdisciplinary Context', *Research Evaluation*, 25, 1–17.

Belcher, B. M., Davel, R., and Claus, R. (2020) 'A Refined Method for Theory-Based Evaluation of the Social Impacts of Research', *MethodsX*, 7, 100788.

Belcher, B. and Palenberg, M. (2018) 'Outcomes and Impacts of Development Interventions: Toward Conceptual Clarity', *American Journal of Evaluation*, 39, 478–95.

Bettencourt, L. M. A. and Kaur, J. (2011) 'Evolution and Structure of Sustainability Science', *PNAS*, 108, 19540–5.

Bill and Melinda Gates Foundation (n.d.) 'About Grand Challenges', <https://gcgh.grandchallenges.org/about> accessed March 2020.

Brandt, P. et al. (2013) 'A Review of Transdisciplinary Research in Sustainability Science', *Ecological Economics*, 92, 1–15.

Bulte, E. et al. (2014) 'Behavioral Responses and the Impact of New Agricultural Technologies: Evidence from a Double-Blind Field Experiment in Tanzania', *American Journal of Agricultural Economics*, 96, 813–30.

Buxton, M. and Hanney, S. (1996) 'How Can Payback from Health Services Research Be Assessed?', *Journal of Health Service Research and Policy*, 1, 35–43.

Byerlee, D., Harrington, L., and Winkelmann, D. L. (2003) 'Farming Systems Research: Issues in Research Strategy and Technology Design', *American Journal of Agricultural Economics*, 64, 897–904.

Carew, A. L. and Wickson, F. (2010) 'The TD Wheel: A Heuristic to Shape, Support and Evaluate Transdisciplinary Research', *Futures*, 42, 1146–55.

CGIAR (2016) 'CGIAR Strategy and Results Framework 2016-2030: Overview', <https://cgspace.cgiar.org/handle/10947/3865> accessed March 2020.

Clark, W. C. (2007) 'Sustainability Science: A Room of Its Own', *PNAS*, 104, 1737–8.

Clark, W. C. and Dickson, N. M. (2003) 'Sustainability Science: The Emerging Research Program', *PNAS*, 100, 8059–61.

Coe, R., Sinclair, F., and Barrios, E. (2014) 'Scaling up Agroforestry Requires Research 'In' Rather than 'For' Development', *Current Opinion in Environmental Sustainability*, 6, 73–7.

Collier, D. (2011) 'Understanding Process Tracing', *Political Science & Politics*, 44, 823–30.

Collinson, M. P. (ed.) (2000) *A History of Farming Systems Research*. Oxford: CABI Publishing.

Coryn, C. L. S. et al. (2011) 'A Systematic Review of Theory-Driven Evaluation Practice from 1990 to 2009', *American Journal of Evaluation*, 32, 199–226.

Deaton, A. (2010) 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, 48, 424–55.

Deaton, A. and Cartwright, N. (2018) 'Understanding and Misunderstanding Randomized Controlled Trials', *Social Science & Medicine*, 210, 2–21.

Donaldson, S. I. (2009) 'In Search of the Blueprint for an Evidence-Based Global Society', in Donaldson, S. I., Christie, C. A., and Mark, M. M. (eds.) *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* pp. 1–18. London: SAGE Publications.

Earl, S., Carden, F., and Smutylo, T. (2001) *Outcome Mapping: Building Learning and Reflection into Development Programs*. Ottawa: International Development Research Centre.

Frieden, T. R. (2017) 'Evidence for Health Decision Making—beyond Randomized, Controlled Trials', *New England Journal of Medicine*, 377, 465–75.

FTA (n.d.) 'What is FTA?', <http://www.foreststreesagroforestry.org/what-is-fta/>

Functowicz, S. O. and Ravetz, J. R. (1993) 'Science for the Post-Normal Age', *Futures*, 25, 739–55.

George, A. L. and Bennett, A. (2007) *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.

Gibbons, M. et al. (1994) *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. London: SAGE Publications.

Godin, B. (2006) 'The Linear Model of Innovation: The Historical Construction of an Analytical Framework', *Science, Technology, & Human Values*, 31, 639–67.

Greenhalgh, T. et al. (2016) 'Research Impact: A Narrative Review', *BMC Medicine*, 14, 78.

Gregerson, H. and Kelley, T. (2007) 'The History of Natural Resource Management Research in the Cgiar', in Waibel, H. and Zilberman, D. (eds.) *International Research on Natural Resource Management: Advances in Impact Assessment*, pp. 12–20. Oxford: CABI Publishing.

Hall, A. et al. (2000) 'New Agendas for Agricultural Research in Developing Countries: Policy Analysis and Institutional Implications', *Knowledge, Technology & Policy*, 13, 70–91.

Harachi, T. W. et al. (1999) 'Opening the Black Box: Using Process Evaluation Measures to Assess Implementation and Theory Building', *American Journal of Community Psychology*, 27, 711–31.

Hearn, S. (2010) 'Outcome Mapping: Planning, Monitoring and Evaluation. Outcome Mapping Learning Community', <https://www.slideshare.net/sihearn/introduction-to-outcome-mapping> accessed March 2020.

Heinrichs, H. et al. (eds.) (2016) *Sustainability Science: An Introduction*. Dordrecht: Springer.

Hirsch Hadorn, G. et al. (2006) 'Implications of Transdisciplinarity for Sustainability Research', *Ecological Economics*, 60, 119–28.

Hirsch Hadorn, G., Hoffmann-Riem, H., Biber-Klemm, S., Grossenbacher-Mansuy, W., Joye, D., Pohl, C., Wiesmann, U., and Zemp, E. (eds.) (2008) *Handbook of Transdisciplinary Research*. New York: Springer.

Hitchcock, C. (2018) 'Probabilistic Causation', in Zalta, E. (ed.) *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/causation-probabilistic/>

Howaldt, J. (2019) 'New Pathways to Social Change–Creating Impact through Social Innovation Research', *Journal for Research and Technology Policy Evaluation*, 48, 37–48.

Independent Science and Partnership Council (ISPC) (2012) *A Stripe Review of Natural Resources Management Research in the CGIAR*. Rome: ISPC Secretariat. <https://ispc.cgiar.org/sites/default/files/ISPC_StrategyTrends_NRM_StripeReview_0.pdf>

ISPC (2015) *Strategic Study of Good Practice in AR4D Partnership*. Rome: ISPC. <https://ispc.cgiar.org/sites/default/files/ISPC_StrategicStudy_Partnerships.pdf>

ISPC (2017) *Quality of Research for Development in the CGIAR Context*, Brief N.62. Rome: ISPC. <https://ispc.cgiar.org/sites/default/files/pdf/ispc_brief_62_qord.pdf>

Jahn, T., Bergmann, M., and Keil, F. (2012) 'Transdisciplinarity: Between Mainstreaming and Marginalization', *Ecological Economics*, 79, 1–10.

Joshi, A. and Witcombe, J. R. (1996) 'Farmer Participatory Crop Improvement, II Participatory Varietal Selection, a Case Study in India', *Experimental Agricultura*, 32, 461–77.

Kasemir, B., Jaeger, C. C., and Jäger, J. (2003). 'Citizen Participation in Sustainability Assessments', in Clark, W. C., and Wokaun, A. (eds.) *Public Participation in Sustainability Science: A Handbook*, pp. 3–36. Cambridge, UK: Cambridge University Press.

Kates, R. (2017) 'Sustainability Science', in Richardson, D., Castree, N., Goodchild, M. F., Kobayashi, A., Liu, W., and Marston, R. A. (eds.) *The International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pp. 1–4. Boston, MA: Wiley-Blackwell.

Kates, R. et al. (2001) 'Sustainability Science', *Science*, 292, 641–2.

Kauffman, J. and Arico, S. (2014) 'New Directions in Sustainability Science: Promoting Integration and Cooperation', *Sustainability Science*, 9, 413–8.

Kelley, T., Ryan, J., and Gregersen, H. (2008) 'Enhancing Ex Post Impact Assessment of Agricultural Research: The CGIAR Experience', *Research Evaluation*, 17, 201–12.

Khandker, S. R., Koolwal, G. B., and Samad, H. A. (2010) *Handbook on Impact Evaluation: Qualitative Methods and Practices*. Washington, DC: The World Bank.

Klein, J. T. (2006) 'Afterward: The Emergent Literature on Interdisciplinary and Transdisciplinary Research Evaluation', *Research Evaluation*, 15, 75–80.

Komiyama, H. and Takeuchi, K. (2006) 'Sustainability Science: Building a New Discipline', *Sustainability Science*, 1, 1–6.

Lang, D. J. et al. (2012) 'Transdisciplinary Research in Sustainability Science: Practice, Principles, and Challenges', *Sustainability Science*, 7, 25–43.

Lesser, W. H. (2003) 'The CGIAR at 31: An Independent Meta-Evaluation of the Consultative Group on International Agricultural Research', in *Thematic Working Paper: Review of Biotechnology, Genetic Resource, and Intellectual Property Rights Programs*. Washington, DC: The World Bank.

Maxwell, J. A. (2004) 'Using Qualitative Methods for Causal Explanation', *Field Methods*, 16, 243–64.

Mayne, J. (2001) 'Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly', *Canadian Journal of Program Evaluation*, 16, 1–24.

Mayne, J. (2012) 'Contribution Analysis: Coming of Age?', *Evaluation*, 18, 270–80.

McCalla, A. F. (2014) 'CGIAR Reform – Why so Difficult? Review, Renewal, Restructuring, Reform Again and Then 'the New CGIAR'—so Much Talk and so Little Basic Structural Change—Why?', Agriculture and Resource Economics Working Paper N.14-001. Los Angeles, CA: The University of California, Davis. <https://escholarship.org/content/qt7h04960c/qt7h04960c.pdf>

Miller, T. R. et al. (2014) 'The Future of Sustainability Science: A Solutions-Oriented Research Agenda', *Sustainability Science*, 9, 239–46.

Miller, T. R., Muñoz-Erickson, T. A., and Redman, C. L. (2011) 'Transforming Knowledge for Sustainability: Fostering Adaptability in Academic Institutions', *International Journal of Sustainability in Higher Education*, 12, 177–92.

Mohr, L. B. (1999) 'The Qualitative Methods of Impact Analysis', *American Journal of Evaluation*, 20, 69–84.

Nowotny, H., Scott, P., and Gibbons, M. (2001) *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. Cambridge, UK: Polity Press.

Pahl-Wostl, C., Mostert, E., and Tabara, D. (2008) 'The Growing Importance of Social Learning in Water Resources Management and Sustainability Science', *Ecology and Society*, 13, 24–7.

Patton, M. Q. (2008) *Utilization-Focused Evaluation*. Thousand Oaks, CA: SAGE Publications.

Pawson, R. (2003) 'Nothing as Practical as a Good Theory', *Evaluation*, 9, 471–90.

Pawson, R. and Tilley, N. (1997) *Realistic Evaluation*. London: SAGE Publications.

Picciotto, R. (2013) 'Evaluation Independence in Organizations', *Journal of MultiDisciplinary Evaluation*, 9, 18–32.

Pohl, C. et al. (2010) 'Researchers' Roles in Knowledge Co-Production: Experience from Sustainability Research in Kenya, Switzerland, Bolivia and Nepal', *Science and Public Policy*, 37, 267–81.

Popowitz, M. and Dorgelo, C. (2018) 'Report on University-Led Grand Challenges', <https://escholarship.org/uc/item/46f121cr> accessed March 2020.

Raitzer, D. and Winkel, K. (2005) *Donor Demands and Uses for Evidence of Research Impact—the Case of the Consultative Group on International Agricultural Research (CGIAR)*. Cali, Colombia: CGIAR.<https://cgspace.cgiar.org/handle/10568/76143>

Randolph, J. (2004) *Environmental Land Use Planning and Management*. Washington, DC: Island Press.

Ravetz, I. R. (1999) 'What is Post-Normal Science', *Futures*, 31, 647–54.

Reilly, R. C. (2010) 'Process Tracing', in Mills, A. J., Eurepos, G., and Wiebc, E. (eds.) *Encyclopedia of Case Study Research*, pp. 734–6. London: SAGE Publications.

Renkow, M. (2010) 'Assessing the Environmental Impacts of CGIAR Research: Toward an Analytical Framework', *SPIA Working Paper*. Rome: ISPC Secretariat. <https://ispc.cgiar.org/sites/default/files/docs/Renkow2010_0.pdf>

Renkow, M. and Byerlee, D. (2010) 'The Impacts of CGIAR Research: A Review of Recent Evidence', *Food Policy*, 35, 391–402.

Research Excellence Framework (REF) (2011) *Assessment Framework and Guidance on Submissions*. Bristol, UK: REF. <https://www.ref.ac.uk/2014/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf>

Robinson, J. (2008) 'Being Undisciplined: Transgressions and Intersections in Academia and Beyond', *Futures*, 40, 70–86.

Rossi, P. H., & Freeman, H. E. (eds.) (1989) *Evaluation: A Systematic Approach*. Newbury Park, CA: SAGE Publications.

Roux, D. J. et al. (2017) 'Transdisciplinary Research for Systemic Change: Who to Learn with, What to Learn about and How to Learn', *Sustainability Science*, 12, 711–26.

Rubin, D. B. (2005) 'Causal Inference Using Potential Outcomes: Design, Modeling, Decisions', *Journal of the American Statistical Association*, 100, 322–31.

Sarewitz, D. (2016) 'Saving Science', *The New Atlantis*, 49, 4–40.

Smith, A., Stirling, A., and Berkhout, F. (2005) 'The Governance of Sustainable Socio-Technical Transitions', *Research Policy*, 34, 1491–510.

Spaapen, J. and van Drooge, L. (2011) 'Introducing 'Productive Interactions' in Social Impact Assessment', *Research Evaluation*, 20, 211–8.

Stevenson, J., Macours, K., and Gollin, D. (2018a) 'Estimating Ex Post Impacts and Rates of Return to International Agricultural Research for Development', *SPIA Technical Note N.6*. Rome: ISPC Secretariat. <https://ispc.cgiar.org/sites/default/files/pdf/ispc_technicalnote_expost_impacts_ar4d_0.pdf>

Stevenson, J., Macours, K., and Gollin, D. (2018b) *The Rigor Revolution in Impact Assessment: Implications for Cgiar*. Rome: ISPC SPIA. <https://ispc.cgiar.org/sites/default/files/pdf/ispc_synthesis_study_rigor_revolution_cgiar.pdf>

Stevenson, J. R. and Vlek, P. (2018) *Assessing the Adoption and Diffusion of Natural Resource Management Practices: Synthesis of a New Set of Empirical Studies*. Rome: Independent Science and Partnership Council (ISPC).

Stokes, D. E. (1997) *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, DC: Brookings Institution Press.

Talwar, S., Wiek, A., and Robinson, J. (2011) 'User Engagement in Sustainability Research', *Science and Public Policy*, 38, 379–90.

van Kerkhoff, L. and Lebel, L. (2006) 'Linking Knowledge and Action for Sustainable Development', *Annual Review of Environment and Resources*, 31, 445–77.

Veerman, J. W. and van Yperen, T. A. (2007) 'Degrees of Freedom and Degrees of Certainty: A Developmental Model for the Establishment of Evidence-Based Youth Care', *Evaluation and Program Planning*, 30, 212–21.

Vogel, I. (2012) *Review of the Use of 'Theory of Change' in International Development, Review Report*. London: DFID. <http://www.dfid.gov.uk/r4d/pdf/outputs/mis_spc/DFID_ToC_Review_VogelV7.pdf>

Walter, A. I. et al. (2007) 'Measuring the Societal Effects of Transdisciplinary Research Projects: Design and Application of an Evaluation Method', *Evaluation and Program Planning*, 30, 325–38.

Weiss, C. H. (1997) 'Theory-Based Evaluation: Past, Present and Future', *New Directions for Evaluation*, 76, 68–81.

White, H. (2009) 'Theory-Based Impact Evaluation: Principles and Practice', *Journal of Development Effectiveness*, 1, 271–84.

White, H. (2010) 'A Contribution to Current Debates in Impact Evaluation', *Evaluation*, 16, 153–64.

White, H. and Phillips, D. (2012) 'Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework', *3ie Working Paper 15*. <https://www.3ieimpact.org/evidence-hub/publications/working-papers/addressing-attribution-cause-and-effect-small-n-impact>

Willetts, J. and Mitchell, C. (2016) 'Assessing Transdisciplinary Doctoral Research: Quality Criteria and Implications for the Examination Process', in Fam, D., Palmer, J., Riedy, C., and Mitchell, C. (eds.) *Transdisciplinary Research and Practice for Sustainability Outcomes*, pp. 122–36. Oxford: Routledge.

Wolf, B. et al. (2013) 'Evaluating Research beyond Scientific Impact: How to Include Criteria for Productive Interactions and Impact on Practice and Society', *Gaia*, 22, 104–14.

Young, J. and Bird, N. (2015) *Informing REDD+ Policy: An Assessment of CIFOR's Global Comparative Study*. London: ODI. <https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/10024.pdf>