

4-1-2006

Afterword: the emergent literature on interdisciplinary and transdisciplinary research evaluation

Julie Thompson Klein

Wayne State University, julietklein@comcast.net

Recommended Citation

Klein, J T 2006. Afterword: the emergent literature on interdisciplinary and transdisciplinary research evaluation. *Research Evaluation*, **15**(1), 75-80. doi:[10.3152/147154406781776011](https://doi.org/10.3152/147154406781776011).

Available at: <http://digitalcommons.wayne.edu/englishfrp/4>

This Article is brought to you for free and open access by the English at DigitalCommons@WayneState. It has been accepted for inclusion in English Faculty Research Publications by an authorized administrator of DigitalCommons@WayneState.

Afterword

Afterword: the emergent literature on interdisciplinary and transdisciplinary research evaluation

Julie Thompson Klein

The complexity of evaluating interdisciplinary (ID) and transdisciplinary (TD) research defies a single standard. Yet, common elements appear in the emergent literature. Five overriding themes stand out. (1) Quality is a relative concept, driven by variability of goals and criteria. (2) A coaching model of evaluation nurtures the research process. (3) Integration is central to the process. (4) Social and cognitive factors interact, requiring management of information and decision-making. (5) The need for change in peer review has led to a variety of strategies. ID and TD evaluation is a generative activity that entails acts of “capitalizing” and “harvesting” expertise while “calibrating” standards to produce new “cultures of evidence”.

Julie Thompson Klein is Professor of Humanities in the Department of Interdisciplinary Studies, Wayne State University, 111 Linden Court, Ypsilanti, MI 48197, USA; email ad5820@wayne.edu and julietklein@comcast.net. She is a member of the Academy of Scholars.

THIS SPECIAL ISSUE appears at a time when evaluating interdisciplinary (ID) and transdisciplinary (TD) research is a subject of widening discussion. In the past, published accounts did not constitute a ‘literature’. They were scattered across multiple forums. They were also longer on anecdotal, intuitive, and normative views than empirical, longitudinal, and large-scale studies. In the absence of clear guidelines, editors Laudel and Origgi recounted in their introduction, faculty and administrators had to “muddle through”. The complexity of the task is daunting. More than one discipline or field is involved, with sometimes conflicting notions of quality and appropriate indicators. Criteria are not identical at all levels, ranging from individual projects to national research systems. They vary across stages, from reviewing *ex-ante* funding proposals to assessing *ex-post* research performance. The context of research also differs by knowledge domain, institutional location, and the purpose and forms of integration.

The complexity of the task defies a single standard or set of metrics. Yet, Max Krott wrote in a chapter on evaluation of TD research in the *Encyclopedia of Life Support Systems* (Krott, 2003) even in the absence of a single ‘best’ procedure, certain elements are key to any act of evaluation. Those elements are revealed in an emergent literature that may be grouped into three clusters. The locus of investigation ranges widely, from small-scale studies of centers and programs to large-scale studies of national initiatives. Some concentrate on academic work and others on trans-sector TD collaborations

with stakeholders in society. Their methodology and conceptual frameworks differ as well, drawing on both qualitative and quantitative approaches. Yet, the variety of studies adds to the robustness of the literature while capturing the bibliographical trace of insights in earlier writings. This Afterword defines the clusters and identifies a number of cross-cutting and overriding themes.

The emergent literature

Cluster 1. Evaluation of ID research

- The contributions in this special issue;
- Archived responses to Boix Mansilla and Gardner (2004);
- Chapter on evaluation in National Academy of Sciences report, *Facilitating Interdisciplinary Research* (2004);
- Vickers (1997);
- Bruun *et al* (2005).

The reality of ID evaluation, Irwin Feller emphasized in this issue, is shaped by ‘multiples’: multiple actors who make multiple decisions in multiple organizational settings that have multiple context-dependent measures of quality. The principle of ‘multiples’ is affirmed in Veronica Boix Mansilla’s exploratory study of five exemplary organizations. The key insight is that variability of goals drives variability of criteria. Generally speaking, researchers engaged in pragmatic problem-solving and product development placed a higher premium on viability, workability, and impact. Projects seeking algorithmic models of complex phenomena were associated with simplicity, predictive power, and parsimony. Efforts aimed at a more grounded understanding of multidimensional phenomena favored work reaching new levels of comprehensiveness, careful description, and empirical grounding.

Boix Mansilla also identified a larger number of indicators of quality than previous studies. Researchers relied systematically on conventional indirect indicators, such as numbers of publications, citations, and grants and rankings of prestige. They did so reluctantly, however. Such measures sidestep the question of what constitutes “warranted interdisciplinary knowledge”, revealed by more primary or epistemic indicators that address the substance and constitution of the work. “Proxy” criteria, informants felt, reduced assessment of their ID work to the criteria of particular disciplines.

The emergence of new and unexpected impacts must be considered as well. “ID impacts,” Boix Mansilla cautioned, “are often diffused, delayed in time, and dispersed across diverse areas of study and patterns of citation practice.” The 2004 report of the National Academies of Science, *Facilitating Interdisciplinary Research*, cited numerous examples. Research on nitrate and sulphate cycles, for instance,

proved to be relevant not only for agricultural production but also research on global climate change and the greenhouse effect. In addition, ID research has had measurable outcomes in multiple fields or disciplines and produced results that feed back into disciplinary research. It has fostered expanded research vocabularies, tool sets, and the ability to work in more than one discipline. Some individuals have also gone on to participate in new subfields, multidisciplinary advisory or review groups, and external public-policy initiatives (pp. 149–154, 159). Different target groups, Krott adds, also make use of knowledge in ways that are unknown at the start of a project and, Defila and DiGiulio (1999) caution, short-term effects should not be overestimated at the expense of recognizing long-term effects.

The publications in Cluster 1 reveal another cross-cutting theme and a point of debate in the literature — the role of disciplines. The first epistemic criterion that emerged from Boix Mansilla’s study is consistency with multiple antecedent disciplinary knowledge. The majority of her informants felt their work was “reasonably consistent” with antecedent disciplinary knowledge, though being “consistent” does not mean demonstrating expert knowledge of the entirety of the disciplines being used. Researchers use the theories, traditions, methods, or schools of thought they borrow in a competent manner. When work “violated” fundamental tenets of a discipline or revealed limitations, Boix Mansilla noted, additional justification was required.

Yet, Dan Sperber pointed out in the CNRS virtual seminar on interdisciplinarity, novelty of the boundary-crossing kind often has the intent of “advancing understanding” by undermining current understanding, calling the *raison d’être* of disciplinarity into question. To be “consistent” with antecedents can betray the very character of interdisciplinary work. Comparably, in a discussion of assessment in Canadian studies, Jill Vickers emphasized that some new ID and TD fields reject disciplinarity in whole or in part. Some disciplines, moreover, have undergone so much change that characterizing them as “stable” matrices is problematic, and in some cases evidentiary protocols are in dispute (Vickers, 1997: 13, 22, 33).

Cluster 2. Evaluation of TD science in the USA

- Evaluations of Transdisciplinary Tobacco Use Research Centers; see especially Stokols *et al* (2003); on concept mapping methodology, see website of William Trochim: <<http://www.socialresearchmethods.net/mapping/mapping.htm>>.
- Evaluation criteria for TTURCs grant proposals (published in revised form in Klein, 2003, 2004).

Inclusion of transdisciplinarity is important because ID research is often a component of TD projects, and studies of two forms — TD science and trans-sector

TD problem solving — have yielded valuable insights and four evaluation models that have generically applicable features. The original meaning of ‘transdisciplinarity’ introduced in the early 1970s connoted common axioms that transcend the narrow scope of disciplinary worldviews through an overarching synthesis. Leading examples over time include synthetic theories such as general systems, structuralism, Marxism, policy sciences, feminism, and ecology.

The emergence of the first new form was signaled in 1992 when Patricia Rosenfield called for “trans-disciplinary science”. TD science fosters systematic theoretical frameworks for defining and analyzing social, economic, political, environmental, and institutional factors in human health and well-being. It goes beyond ID combinations of existing disciplinary approaches to generate new topic-based domains. The intellectual outcomes include new hypotheses for research, integrative theoretical frameworks for analysis of particular problems, novel methodological and empirical analysis of those problems, evidence-based recommendations for public policy, and changes in trainees’ career development outcomes (Stokols *et al*, 2003; Rosenfield, 1992).

The concept of TD team science is being developed in a group of centers affiliated with the US National Cancer Institute in the National Institutes of Health. Evaluators of one program in particular, the Transdisciplinary Tobacco Use Research Centers (TTURCs), used concept-mapping methodology to gain an overview of outcome domains in large-scale collaboration on complex health problems. Brainstorming generated 262 potential outcomes that were ultimately reduced to five general regions or clusters of evaluation: scientific integration, collaboration, professional validation, communication, and health impacts. Temporality was an added consideration, recognizing differences across research phases.

The map of outcomes was then translated into a logic model that depicts the sequence and causal relationships of outcome constructs. Together, the map and the model guided development of approaches to measurement based on hierarchical thematic analysis of qualitative data. Familiar indicators are used, including publications in recognized journals. However, added weight is placed on the quality of the collaborative TD process; the production of new or improved methods, models, and theories; and interventions in health practices, policy, and outcomes. Furthermore, indicators are not restricted to one phase. They have a feedback relationship that a strictly linear model of evaluation cannot capture.

The model begins with the basic activities of the research centers (training, collaboration, and TD integration) and the earliest expected outcomes. Basic activities lead to development of new and improved methods, science, and models. Improved interventions are tested and lead to publications, which also result from and describe intermediate

products of improved methods, science, and models. Publications lead to recognition and institutionalization of TD research, feeding back into the overall infrastructure and capacity of centers and increasing support for basic activities. In addition, publications provide a content base for communicating scientific results to a broader community. Recognition provides a secondary impetus for communications and publications, policy implications also result from communications and publications, and translation to practice is influenced by improved interventions. There is a dynamic relationship, though, between translation to practice and policy implications. Health outcomes are influenced by treatments and new health practices and by related policy changes. Positive or negative health outcomes, in turn, feed back into new policies and practices.

Cluster 3. European TD movement

- Chapters on expectations and assessment in Tress *et al* (2003);
- Defila and DiGiulio (1999);
- Krott (2003);
- Spaapen and Dijkstra (2005);
- <td-net> network for transdisciplinarity in sciences and humanities <<http://www.transdisciplinarity.ch/>>.

The second new form of transdisciplinarity was signaled by a new discourse of trans-sector TD problem-solving that arose in Europe during the late 1980s and early 1990s. The distinguishing feature of this discourse is the externality of complex problems and the participation of a wider range of stakeholders. Problem domains vary widely, though. Some collaborations focus on innovative technology and product development. Others focus on controversial social issues that have an impact on community stakeholders. The international network <td-net> is an electronic forum for work in both areas as well as ongoing interest in unity of knowledge. The network’s website and the forthcoming *Handbook of Transdisciplinary Research* sponsored by <td-net> capture work based primarily in Europe and in North–South partnerships, including the DACH group’s comparative studies of projects and programs in Switzerland, Austria, and Germany. The most comprehensive report on TD evaluation was commissioned by the Swiss National Science Foundation. After reviewing the literature and experiences, Defila and DiGiulio built an evaluation model that illustrates another cross-cutting theme in the literature. Flexibility and sensitivity to context, Liv Langfeldt exhorted in this issue, are key parameters of evaluation.

The heart of Defila and DiGiulio’s method is a modular questionnaire called the Catalogue of Criteria. The Catalogue provides the largest possible number of building blocks to “construct” a meaningful self-evaluation or external evaluation of a

research program. It takes a heuristic, generative “pool” approach sensitive to the particulars of overarching and subproject levels as well as separate phases. All categories may not apply at all phases, for example scientific quality versus integration/synthesis or project organization/management. The time and number of evaluations can be adjusted, the question of who performs the evaluation and how criteria are weighted are left open, not all aspects of a project need to be assessed in each round of evaluation, and not every program needs to take all questions into account. Context-related adaptations, deletions, and additions are expected.

The last of the new evaluation models in the TD literature is the Research Embedment and Performance Profile (REPP), developed by Sci_Quest in the Netherlands. Sci_Quest is an independent research agency with a long-standing interest in assessing scientific research in a policy context or broader societal context. In their final report based on two methodological studies in the areas of agriculture and pharmaceuticals, Spaapen and Dijstelbloem describe the model. It draws on the field of science and technology studies, furnishing a grounded theory for evaluation that incorporates ideas from Gibbons *et al.*'s theory of Mode 2 knowledge production, Nowotny *et al.*'s criterion of “socially robust knowledge”, and innovation studies, especially Callon and Larédo's Compass Card for research labs.

The central insight is that the mobility of participants and interaction and communication patterns furnish a heuristic for identifying differences in social domains or contexts for knowledge production. In each context, different expectations exist, with attendant norms, values, and priorities. The REPP facilitates reconstruction of the relevant environment and the performance of a group within it. It seeks patterns and profiles, comparing results with a group's self-proclaimed mission. A quantifiable benchmark is set for each indicator in consultation with researchers and policy-makers. Scores are then plotted on a radar-like graph that represents varied activities in a quantifiably balanced way.

Overriding themes

Five overriding themes emerge when situating this issue within the growing literature on evaluation.

The expanded meaning of quality

Quality, Spaapen and Dijstelbloem concluded, is a relative concept determined by relations within the environment of a group and their goals. Ultimately, “good” research can have many profiles. In the context of innovation and creativity, a strict set of criteria or “uniform yardstick” may be counterproductive. A standard assessment procedure can help in charting a program's interactions with a broader environment and insuring that work is

The central insight is that the mobility of participants and interaction and communication patterns furnish a heuristic for identifying differences in contexts for knowledge production

scientifically sound and credible. Yet, research must “attune a pluralism of interests and values” within a dynamic set of programs and contexts where new opportunities may appear. A thorough and stringent review process, Langfeldt adds, may bias against controversial, risk-taking, non-conventional research and “radical” interdisciplinarity that is not as well established as “regular” interdisciplinarity or lacks recognized grounds for assessing quality.

The value of coaching the process

Spaapen and Dijstelbloem described the REPP as a coaching model rather than a jury model. It facilitates self-reflection about what members are supposed to be doing and how well they are doing it. Others endorse coaching as well. Klein advocates using the “Guiding Questions for Integration” designed originally for judging TTURC grant proposals as both an evaluation checklist and a sequence for nurturing the integrative process. Defila and DiGiulio intend the Catalogue of Criteria to be used as guidelines for coaching TD work and, in this issue, Laudel presented an actual model of coaching interdisciplinarity in the evaluation scheme of German collaborative research networks. It is needed at the *ex-ante* stage of funding proposals as well. The Academy of Finland Integrative Research (AFIR) team also recommends that national funding agencies coach the ID and TD research process in collaboration with research coalitions and external reviewers (Bruun *et al.*, 2005: 171–172).

The centrality of integration

Integration, Krott maintains, is the critical point of focus for evaluating TD research. Acknowledging its importance, the second epistemic criteria in Boix Mansilla's study is “balance in weaving perspectives together” into a generative and coherent whole. Integration was also one of the four “hot spots” identified in the American Association for the Advancement of Science (AAAS) Symposium on quality assessment, in the form of “reaching effective syntheses”. Klein, and Defila and DiGiulio, admonish that integration must be engaged from the very beginning in the work process, and the AFIR team recommends that funding agencies pay explicit

attention to it in writing requirements for grants. Applicants should be asked to explain why an integrative approach is necessary, what kind of integration is proposed, how it will be carried out from both intellectual and organizational standpoints, and what level of preparedness participants have (Bruun *et al.*, 2005: 172–173, 195–196).

The interaction of social and cognitive factors

It is important, Rainer Kamber cautioned in the CNRS seminar, to recognize that certain forms of integration do not presuppose certain forms of social cooperation, and vice versa. That said, the TTURCs model does not sharply separate cognitive-epistemic and social factors. Krott describes a TD project as “a social interaction” that requires careful attention to the information and decision-making process. In the context of a heterogeneous mix of disciplines, compromises need to be made and the “best” option available may be a partial consensus. “Competence”, Defila and DiGiulio add, is defined partly in terms of how well management of the overarching project implements intended methods for consensus building, integration, and networking across subprojects. Klein agrees, emphasizing the importance of allowing time for mutual learning. Communication and negotiation of difference are the linchpins of collaborative ID and TD research. In the collection of essays, *Interdisciplinary and Transdisciplinary Landscape Studies*, Aenis and Nagel (2003) highlight two axiomatic considerations for TD evaluation: the metalevel of interdisciplinarity (communication among researchers) and participation (communication between researchers and regional actors as well).

The need for change in peer review

The final overriding theme — peer review — loomed large in this issue, and so merits extended final comment. The underlying assumption is that qualified experts certify proposed or completed work using what Laudel called the “yardstick” of rational and “objective” decision. Recent findings suggest this regime of control may not always work against ID and TD research. The AFIR team found that they are being facilitated in regular categories of the Academy of Finland’s research funding system, and Langfeldt cited a study of *ex-post* evaluations of Dutch physics groups that revealed no significant correlation between peer ratings and the degree of interdisciplinarity. This research question merits more investigation. Nonetheless, finding adequate reviewers remains a widespread need at present.

Given that ID research is a new synthesis of expertise, Laudel questions whether “peers” in the strong sense of the word exist. A “commonly agreed yardstick” must be developed to avoid reinforcing “cognitive particularism” and, Krott adds, guidelines for peer review must be formulated to minimize, not maximize, disciplinary standards. A variety of

strategies have been proposed or implemented. The AFIR team cited use of joint panels, and the US National Academies of Science report highlighted “matrix evaluation” combining separate discipline-based reviews with a full panel review involving both disciplinary and interdisciplinary members (166). The report on the AAAS Symposium cited Nina Fedoroff’s proposal for more “fit and agile” review groups constituted by smaller “on-the-fly” review teams brought together electronically and, in NIH panels, “interpreters” who bridge the epistemic gap among content experts by performing an intermediary role. Langfeldt also called attention to special funding programs that “bypass” conventional social-cognitive control mechanisms of the research community, though Feller cautioned that special programs may “ghettoize” ID research.

Boix Mansilla’s notion of “expert communities” is a bridge concept between discipline-based and ID- and TD-based evaluation. Identifying experts who fit the “problem space” is crucial. They are, in the words of the AAAS Symposium report, “close to the substance of the work”. In established interdisciplinary fields, such as biochemistry, scholars have negotiated standards of quality over time, yielding more or less common beliefs about what counts as quality work. The task is more difficult in ‘incipient’ and ‘emerging’ fields. Standards of excellence and communities of scholars have not been established yet. In the case of review for tenure and promotion at Duke University, Peter Lange reported, candidates can contribute names of suitable peers. Langfeldt’s comparative study of regular grant schemes in 12 European and US research councils also identified cases in which applicants could suggest names of reviewers.

ID panels, Lamont, Mallard, and Guetzkow suggested in this issue, are “sites where new rules of fairness are redefined, reinvented and slowly recognized”. In their study of multidisciplinary fellowship competitions in humanities and social sciences, they found that traditional rules of deference to expertise and respect of disciplinary sovereignty were the most frequent source of conflict. In the process of ID evaluation, participants “define, construct and enact” rules in an “intersubjective production of the belief in fairness”. The distance of non-expertise, the study also showed, plays a positive role in lowering disciplinary bias. Broadly speaking, a review group must strike a balance between theoretical breadth and empirical accuracy, interdisciplinary appeal, and disciplinary mastery of methods. The role of methodological pluralism is essential to insuring that panels achieve a more consensual quality assessment.

Researchers differ on whether special criteria are needed. Some questions in the German evaluation scheme may be directly related to interdisciplinary collaboration, but Laudel contends that ID evaluation is possible without special criteria. The critical elements are relative empowerment of applicants and enforced “interdisciplinary learning” of reviewers

with careful monitoring of the institutional rules of assessment. In the past, Dan Sperber observed in the CNRS seminar, people seeking legitimization of interdisciplinary initiatives had to be both parties and judges, educating their judges in the process of doing and presenting their work. Similarly, Boix Mansilla commented earlier, in highly innovative work where novel territories are being charted, developing validation criteria to gauge their progress becomes part of the inquiry process. Direct and longitudinal involvement of reviewers of the kind Laudel reported is rare, but it represents an important closing lesson. The German model built on an “interdisciplinary culture” that emerged after decades of successful ID research. Manifested in federal agencies and universities, Laudel explained, this culture fostered a “‘moderate’ interdisciplinarity” that enables individuals to communicate with others, ask the right questions, and present research results in a simplified way that is accessible across fields. The acts of “capitalizing” and “harvesting” multiple expertise and “calibrating” review standards reported in this issue and the rest of the emergent literature beckon the long-term development of more widespread and sustainable interdisciplinary and transdisciplinary cultures of evidence.

References

- Aenis, T and U J Nagel 2003. Impact indicator definition within a transdisciplinary research group. In *Interdisciplinary and Transdisciplinary Landscape Studies: Potential and Limitations*, eds. B Tress, G Tress, A Van der Valk and G Fry. Delta Series 2, pp. 160–169. Wageningen.
- Boix Mansilla, V and H Gardner 2004. Assessing interdisciplinary work at the frontier: an empirical exploration of ‘symptoms of quality’. Virtual seminar on *Rethinking Interdisciplinarity*. Centre National de la Recherche Scientifique (CNRS). <<http://www.interdisciplines.org/interdisciplinarity/papers/6>>, last accessed 31 July 2006.
- Bruun, H, J Hukkinen, K Huutoniemi and J T Klein 2005. *Promoting Interdisciplinary Research: The Case of the Academy of Finland*. Publications of the Academy of Finland, Series 8/05. Helsinki: Academy of Finland.
- Defila, R and A DiGiulio 1999. Evaluating transdisciplinary research. *PANORAMA* (Newsletter of the Swiss Priority Program Environment, Swiss National Science Foundation), 1/99. <<http://www.ikaoe.unibe.ch/forschung/ip/Specialissue.Pano.1.99.pdf>>, last accessed 31 July 2006.
- Klein, J T 2003. Thinking about interdisciplinarity. *Colorado School of Mines Quarterly*, 103(1), 101–114.
- Klein, J T 2004. *Guiding Questions for Integration. Integration Symposium 2004 – Proceedings*, CD-ROM, 5–8. Canberra: Land and Water Australia.
- Krott, M 2003. Evaluation of transdisciplinary research. *Encyclopedia of Life Support Systems*. Oxford: EOLSS Publishers. <<http://www.eolss.net>>, last accessed 31 July 2006.
- National Academy of Sciences, 2004. Evaluating outcomes of interdisciplinary research and teaching, *Facilitating Interdisciplinary Research*, ch. 8, pp. 149–170. Washington, DC: National Academies Press.
- Rosenfield, P, 1992. The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science and Medicine*, 35(11), 1343–1357.
- Spaapen, J and H Dijstelbloem 2005. *Evaluating Research in Context: A Method for Assessment*. The Netherlands: Consultative Committee of Sector Councils for Research and Development (COS).
- Stokols, D, et al 2003. Evaluating transdisciplinary science. *Nicotine and Tobacco Research*, 5, 1–19.
- Tress, B, G Tress, A Van der Valk and G Fry eds. 2003. *Interdisciplinary and Transdisciplinary Landscape Studies: Potential and Limitations*, Delta Series 2. Wageningen.
- Vickers, J 1997. “[U]nframed in open, unmapped fields”: teaching and the practice of interdisciplinarity. *Arachne: An Interdisciplinary Journal of the Humanities*, 4(2), 11–42.