8th International Conference on Evaluation for Practice

Improvement by Evaluation

Peer Reviewed Full Papers of the 8th International Conference on Evaluation for Practice: "Evaluation as a Tool for Research, Learning and Making Things Better"

A Conference for Experts of Education, Human Services and Policy
18–20 June 2012, Pori, Finland

Edited by:
Satu Kalliola, Pekka Kettunen, Ossi Eskelinen, Kati-Jasmin Kosonen, Ilmari Rostila and Anu Leander

University of Tampere
School of Social Sciences and Humanities
Unit at University Consortium of Pori

Improvement by Evaluation

Peer Reviewed Full Papers of the 8th International Conference on Evaluation for Practice
"Evaluation as a Tool for Research, Learning and Making Things Better"
A Conference for Experts of Education, Human Services and Policy
18 – 20 June 2012, Pori, Finland

Edited by Satu Kalliola, Pekka Kettunen, Ossi Eskelinen, Kati-Jasmin Kosonen, Ilmari Rostila and Anu Leander

ISBN: 978-951-44-8859-7

Notes:
Authors are responsible for the accuracy of all submitted materials including references, quotations and spellings. If there is significant use of copyrighted materials in the article, the author is responsible for securing permission to use the material.

University of Tampere
School of Humanities and Sciences
Unit at University Consortium of Pori

http://tampub.uta.fi/english/

Table of Contents:

IV Evaluation of Education Outcomes and Learning

V Evaluation of Academic Staff Performance

VI Different Aspects of Evaluating Social Care and Health Services

Preface

It is easy to reach an agreement about the basic objectives of scientific conferences: conferences are about presenting, sharing and reflecting upon new research methods and results. However, opinions differ as to the best way to communicate the new findings of the conference to the conference delegates and especially to the wider audience – there are differences in traditions. Some scientists have always written a full paper to be published in the conference proceedings or, more recently, on the conference website. Others have simply given an oral presentation with the assistance of a variety of tools – from overhead projector slides to videos –, and their core new findings have been saved to be published on prestigious academic forums.

Depending on the viewpoint, the 8th International Conference on Evaluation for Practice, (June 18–20, 2012, Pori, Finland), may be seen as an established conference with a history of over 20 years in sharing the latest knowledge on the evaluation of human services, or as a promising new conference, due to its scope being expanded to cover also the evaluation of working life and regional development. Under these circumstances, the scientific committee of Eval 2012 decided to compile an electronic publication of all those conference papers that would pass a peer review. Writing full papers was only an option: it was left for the presenters to decide whether they would like to develop their presentations to full papers to be reviewed or whether they simply wanted to give their presentation in the conference sessions.

The conference theme coordinators organized the peer review process. An editorial group was formed by the local organizers, Ossi Eskelinen, Kati-Jasmin Kosonen, Anu Leander, Ilmari Rostila and Satu Kalliola, complemented by Pekka Kettunen from the scientific committee.

It is time to thank all those nearly 30 anonymous reviewers, mainly from outside the scientific committee, who helped the authors of the submitted full papers to revise them. The revision was not an easy task. Some of the authors gave up in an early phase and not all of the papers could be accepted. Ultimately, a collection of 18 papers was compiled from a diverse body of Eval2012 participants. Some of the authors already have solid careers as evaluation researchers while some are taking their first steps in this challenging research area.

I am happy to introduce these Eval 2012 conference papers in a book that focuses on the fundamental aim of evaluation research: making things better.

Satu Kalliola
Eval 2012 Chair

The Challenging Art of Evaluation

Pekka Kettunen, University of Jyväskylä
Satu Kalliola, University of Tampere

The many dimensions of evaluation as research and practice

Since the first International Conference on Evaluation for Practice, held in Huddersfield, United Kingdom in 1995, this conference series has emphasized the pursuit of using the results of evaluation research in the improvement of human services and other organizational life. This as such is not unusual, since most evaluation researchers see their field as a significant and valuable genre of applied social research. Although there exist many modern definitions of evaluation research, they all contain some aspects of the classic, and very practical, definition of Rossi & Freeman (1991, 13) who see evaluation research as a robust arena of activity directed at collecting, analyzing and interpreting information on the need for, implementation of, and effectiveness and efficiency of intervention efforts to better the lot of humankind. However, the practice orientation of evaluation researchers may force them to face some conflicts during their careers as evaluation as a tool of improvement is faced with a number of challenges.

Not surprisingly, evaluators and evaluation theorists do not agree on everything and a number of theoretical and methodological debates go on. On the other hand, beyond the theoretical premises the sheer purpose of using evaluation as a tool of improvement faces particular problems. Evaluators may not have the sufficient time and resources to conduct a proper evaluation. In addition, finding empirical information may be difficult. Pollitt (2003) argues that if a program does not routinely monitor its own outputs and outcomes in a systematic and reliable way, it can be difficult for evaluators to measure them.

When it comes to effects, it is crucial to proceed carefully. Chen (2003) warns about excessively hasty explications of causal relationships and encourages reflection upon the actual benefits of statistical testing. The evaluation of effects begins with formulating, in an explicit way, the methods applied and the desired results as well as the way these results are being measured. Only then is it possible to start generalizing the results.

Should evaluation be critical? Certainly, but that does not encourage decision-makers to undertake evaluations. Vedung (2009) argues that evaluation results will be used when there is a consensus between the initiator and evaluator as to what the aims of the evaluation will be. Managers do not necessarily want to reveal how badly their organizations are performing. Moreover, the people whose life situation the evaluators would like to improve, may not want to be helped.

Another viewpoint to the usage of evaluation is the perspective of evidence. If we reach the point where we can "speak truth to decision-makers" (Wildavsky 1993), there will be no limits to the usage of evaluation information. In real terms, however, evaluation is but an ingredient in political decisions-making (Pollitt 2003).

Based on the various problems mentioned above, new methodological approaches to overcoming these problems are emerging almost constantly. For example, various participatory approaches have been introduced. Many of the Eval2012 articles discuss these grievances and suggest steps forward.

The many dimensions of Eval 2012 conference papers

This volume has been compiled as a freely evolved combination of the themes that the participants of the 8th International Conference on Evaluation for Practice have found

important in the framework of conducting and applying evaluation research under the umbrella of "Making things better". The first articles, authored by *Alexander Fink* & *Allison Link, James Herbert* and *Anne Kallio, Anne Pässilä* & *Tuija Oikarainen* all take a stance on the role of knowledge gained as a result of evaluation process.

Fink and *Link* are in the core of learning from evaluation when they claim that what is found and learned should be put into action. In other spheres of life it is often valuable to just learn, without tying learning to practical changes. In the realm of evaluating social interventions the whole point is to conduct the next intervention in an improved way. They argue that the leaders and managers of organizations are the actual gatekeepers in translating and using evaluation to "make things better". They review numerous leadership theories in order to find practically sustainable solutions to the dilemma mentioned by *Herbert*: what are the prerequisites for applying evaluation results in the future activities of the organization in question.

Herbert approaches knowledge yielded by evaluations from the viewpoint of creating opportunities for learning in Australian nongovernmental organizations. His two cases shed light on the various challenges that must be overcome in order to actually understand and reflect upon evaluation information. His article revolves around the conceptualizing practitioner learning.

Kallio, Pässilä and *Oikarainen* reflect upon the various types of knowledge that the employees of the workplaces possess before the workplace evaluation processes and suggest employee-driven evaluations. They find it essential to take this knowledge into account while planning evaluation, development and training interventions in workplaces. Their context is the Finnish working life. One of their key concepts, employee-driven evaluation, seems to combine the classic ideas of employee participation and human resources as key resources of all work organizations, to the search for various ways of knowing.

Joe Dee Keller, *Janice Laakso*, *Christine Stevens* and *Cathy Tashiro* go even further in their article by presenting that it is precisely the evaluation researcher who has to consider all of the abovementioned dimensions: listening to the people whose life is being studied and conveying their voice to the policy makers. Their evaluation research case study is most challenging, involving immigrants and refugees with diverse cultural backgrounds and language skills, living in the north-western parts of the United States. Among their methodological choices is photovoice: the participants document their lives by using a camera, and later the photographs are discussed to produce a mutual understanding of the context. However, the main emphasis of the article is on the overall ethical concerns of evaluation researchers who should simultaneously cope with political pressures and methodological challenges to serve the people in need. This demands a self-reflexive attitude and continuous questioning of the method-driven approaches of classic evaluation research.

Esa Jokinen seeks to apply this kind of reasoning in a less fragile and culturally diverse environment, namely in the evaluation of the Finnish local government reform. He argues that self-reflexivity should be an essential part of evaluation competence, combined with the skills to initiate and use dialogue with the stakeholders in the evaluation process.

This far, we have presented the main points of evaluation research conducted within the boundaries of one nation. An evaluative approach is also a useful tool in international comparisons, if there exists reliable and valid data from culturally and historically diverse organizations, or the researchers find a way to collect it. *G. Anthony Giannoumis* and *Rune Halvorsen* have chosen to analyze policy documents form the U.S., UK and Norway in order to investigate how the national, and also supranational, policies balance and mediate regional and international economic and social needs in the context of E-Accessibility, regulated by the United Nations Convention on the Rights of Persons with Disabilities. The results show that the different national policy traditions truly matter. The authors conclude by emphasizing

the utility of judicial enforcement, the flexibility of providing a low threshold administrative complaint mechanism, and the importance of monitoring. *Jaroslav Dvorak* examines and assesses institutional and policy arrangements for evaluation in three new EU member states, Bulgaria, Poland and Lithuania, in the context of managing European Structural Funds support and fulfilling their evaluation requirements. The results of the analysis of diverse data consisting of relevant documents, expert interviews and questionnaires for public officials show that there are differences between countries, but they are all on their way to building evaluation capacity.

Nowadays, there seems to be at least some agreement that learning, and also other outcomes of learning, may be evaluated in much more sophisticated ways than by monitoring the exam results of pupils, students or adult learners. *Blair Stevenson* and *Nancy Doddridge* write about comparisons between two different approaches to assess Indigenous education in Quebec, Canada. Their results cover the idea of culturally competent evaluation, and their conclusions point out the need to use assessment tools in supporting the teachers and schools, instead of the federal government.

Australian researchers *Shelley Kinash, Trishita Mathew, Romy Lawson, James Herbert, Erica French, Tracy Taylor, Cathy Hall, Eveline Fallshaw* and *Jane Summers* present the results of their collaborative research project among a large number of business schools. Their objective is to shed light on the philosophy and motivation for Assurance for Learning, which is presented in the larger framework of higher education evaluation and which is also connected to other reforms in higher education. In the article, evaluation is seen as a process and quality assurance as an intended outcome.

The focus of *Heikki Hannula, Elena Ruskovaara, Jaana Seikkula-Leino* and *Anne Tiikkala* is not on learning as such, but rather on the teaching. They address the issue of entrepreneurship education practices, which are seen as societally very important in the present era of welfare state transitions. They present a tool developed for measuring these activities among teacher educators.

In the framework of the evaluation of higher education, the evaluation of university staff performance has emerged as a current, and also partly controversial, topic. *Mikael Collan, Jan Stoklasa* and *Jana Talasova* present three different examples of academic faculty evaluation systems from the Czech Republic and Finland. In addition, *Jan Stoklasa* and *Jana Talasova* continue the discussion with *Pavel Holecek* and present the details of a model used in the Czech Republic, called linguistic fuzzy modeling.

The most familiar area of evaluation research is the improvement of public interventions, especially with regard to social care and health services. *Susan Fletcher* shows how an evaluation of a cardiac rehabilitation program led to a review of the way chronic disease programs were offered in a rural agency. She anticipates that the outcomes of the project will allow local agencies to provide cardiac rehabilitation programs that better meet the needs of participants. Nothing stops the evaluators from going further and generalizing the results on the basis of the qualitative analysis.

The article by *Jodi Constantine Brown* aims at describing and evaluating an existing yoga program for vulnerable adults with cancer. Patients who chose to attend a yoga class were compared to a group of patients waiting for their treatment appointment who opted not to participate in the yoga class. The results of the study show that diverse populations appreciate the experience of the yoga classes, but the results do not conclusively indicate that yoga positively affects quality of life.

Riitta Meretoja and *Mikko Saarikoski* ask how nursing students experience their clinical environment, the supervision provided by their personal nurse supervisors, and the level of intervention with their nurse teacher. They conclude that the majority of students (respondents) were very satisfied with the achievement of their own learning goals and felt

that supervision supported their professional development. However, the respondents were quite critical of how their earlier theoretical nursing studies supported their learning during their clinical placements. These kinds of findings are valuable in improving the learning results and in that way the quality of health-care.

Janissa Miettinen discusses the research methodology in practice evaluation. Miettinen's starting point is to better understand how to improve child welfare. She then discusses realistic evaluation and critical realism and finds abduction, the creation of hypotheses, the best way of proceeding when seeking to determine the outcomes of child open-care interventions.

Vuokko Niiranen and *Alisa Puustinen* focus on the welfare reform the purpose of which is to increase the size of the municipalities in Finland. The authors further point out that causal relationships are extremely difficult to identify in this context. They suggest that evaluation of the reform should try to include several potential factors affecting the results in the evaluation model. Working with this kind of complex model, they come to tentative conclusions of the outcome of the reform.

As the only representatives of economics in this volume, *Ari Karppinen, Saku Vähäsantanen, Teemu Haukioja* and *Arja Lemmetyinen*, show how it is possible to construct an easily reliazable yet reliable quantitative procedure for evaluating the economic effects of tourism at a regional level. After a theoretical framework and practical definitions, the authors proceed to describe their procedure that both specifies regional level effects of tourism and also makes forecasts for the future. The main emphasis is on the assessment of an annual tourism income and employment in the case region of Satakunta.

The dynamic evaluation research

The diverse compilation of Eval 2012 conference papers shows that evaluation research has an inherent potential to improve policies, services and education. Decision-makers and other bodies involved do not necessarily find evaluation agreeable as it may also lead to cutting existing activities. In that way we can see the critical, democratic character of evaluation. It is a viewpoint which argues that all public services must serve the needs of the citizens and all people in need. The articles in this volume speak for the notion of Rossi & Freeman (1991, 14), who see that in all cases the aim of evaluation research is to provide the most valid and reliable findings possible within political and ethical constraints and the limitations imposed by time, money and human resources.

It is difficult to succeed as an evaluation researcher unless one is willing to strive for methodological rigor and a balance between conflicting values of the various stakeholders of programs and their evaluation. Overall development of methodologies seems to be constantly ongoing, and it is not at all out of place to see evaluation research as a very dynamic way to respond to the many calls of our global world.

References

Chen, H.-T. (2005) *Practical Evaluation. Assessing and Improving, Implementation and Effectiveness.* Thousands Oaks: Sage.

Pollitt, C. (2003) *The Essential Public Manager.* Open University Press: Maidenhead.

Rossi, P. H. & Freeman, H. E. (1991) *Evaluation. A Systematic Approach.* Fourth Edition. Sage: Newbury Park, London, New Delhi.

Vedung, E. (2009) *Utvärdering i politik och förvaltning.* Lund: Studentlitteratur.

Wildavsky, A. (1993) *Speaking Truth to Power. The Art and Craft of Policy Analysis.* 4th Edition with a new introduction by the author. New Brunswick (NJ): Transaction.

I

Learning and Knowledge in Evaluation Research

"Making things better": the role of leadership in evaluation use

Alexander Fink, University of Minnesota, School of Social Work
Alison Link, University of Minnesota, Department of Postsecondary Teaching and Learning

Abstract

This study suggests that to "make things better" using evaluation is about "leadership", in part. Translating evaluation findings into actual activities that change social groups, organizations, programs, or policies requires an agent, who can be thought of as a type of leader. A study of leader(ship) from the point of view of using evaluation to "make things better" is the focus of this paper. In order to translate and use evaluation findings, leadership requires many necessary and subtle discernments, discriminations, and decisions, such as analysis of what needs/wants to "be better"; the sources, issues, problem to be changed; the analysis of political and group dynamics involved in change; assessments of alternative change strategies, and the like. These themes are basic to leadership scholarship and exploration in that context can lead to better and more effective practice in translating and using evaluation to "make things better".

Keywords: leadership, use, utilization, improvement

Introduction to Leadership Theories

Some evaluators believe that it is their role to help ensure that evaluation is actually used for program and policy improvement. These actions can be seen as 'leadership'. The scholarly and practical leadership literature--and the everyday practice of leadership--have long focused on the idea of "making things better" for individuals, organizations, governments, and society. The theory and practice of leadership can become a helpful lens for evaluators concerned with encouraging effective evaluation use.

'Leader(ship)' is a slippery term because its definition and meaning shift dramatically by context, worldview, place, and history. There are many definitions of what it means to lead, of who is a leader, and of what differentiates leadership and management. In this review, we focus on leadership literature as a body distinct from management literature. Michael Quinn Patton's distinction between *complicated* and *complex* situations may help illustrate this distinction, and why we have chosen to focus on leadership literature as a tool for improving evaluation use. Management, on the one hand, frequently addresses problems that are *complicated*--problems where "what needs to be done is challenging and difficult, but *knowable*" (Patton, 2011, p. 87). Management is traditionally linked relatively bounded problem areas, including "planning, organizing, staffing, and controlling" (Northouse, 2010, p. 9). Leadership, on the other hand, consistently focuses on *complex* situations, where there is much higher uncertainty and higher potential for disagreement about what course of action should be taken (Patton, 2011, p. 90). Although there is some crossover in the fields of leadership and management, the leadership literature may be particularly relevant and of more immediate interest to evaluators who are concerned with evaluation use in highly complex situations, where there are many 'unknowns' and 'unknowables'.

Turning, then, to leadership: we notice that traditional definitions of 'a leader' focus on a person with positional authority (e.g. the CEO of a company). A second perspective challenges this. It considers the ways in which a person with or without the title or status of 'leader' might choose to act "to mobilize people to tackle tough problems" (Heifetz, Grashow & Linsky, 2009, p. 1). Defined in this way, leadership is "a process whereby an individual

influences a group of individuals to achieve a common goal" (Northouse, 2010, p. 3). Note the emphasis here on intentional action. Note too that titular authority is not necessary within this conceptualization. Viewing leadership in this second way--as the action of an agent influencing others to achieve a common goal--will help the evaluator to position herself within the leadership literature and consider what the field of leadership might offer her as a guide for supporting evaluation use.

We have conducted a review of the leadership literature, starting with Northouse's classic survey *Leadership: Theory and Practice* (2010) and encompassing additional standard literature from the field. The criteria used to select leadership models for this review can be found in Appendix A. We will report and discuss three models of leadership: Adaptive Leadership, the Style Approach, and Primal Leadership. We start by providing an evaluation example to read using these different leadership lenses. Then, we will examine each of these models in detail, attending to specific distinctions in each relevant to evaluation use. After examining each model, we propose questions an evaluator can consider to improve evaluation use. A table summarizing these theories and their corresponding questions can be found in Appendix B.

A Case Example: Evaluating Technology in Schools

Schools in the Kyrene School District of Chandler, Arizona, USA are rapidly adopting technology-enhanced classrooms, spending $33 million on laptops, interactive screens, software, and support since 2005. The district has been praised for its integration of technology, though standardized test scores in reading and math have remained stagnant, despite these investments. Simultaneously, the district has seen massive budget cuts and has been forced to increase class sizes and decrease course offerings in the arts and physical education. Although data showing growth in test scores is weak, administrators and teachers believe that some positive changes are in fact occurring, and that evaluators should be able to find valid measures of change in student performance. Broader trends in use of technology nationwide do not seem to convincingly show that the integration of technology does anything more than amplify already existing successes or failures (Richtel, 2011).

This story of K-12 technology integration is repeated thousands of times in school districts across the United States. An evaluator asked to explore the issue of technology integration in schools faces a number of important considerations in the U.S. educational context:

- There is a huge array of stakeholders in school technology debates, including students, teachers, administrators, and local and national politicians. On the national level, the discussion on K-12 technology integration is still somewhat decentralized and nebulous, with a vague consensus that "more technology is better".
- "Success" in educational initiatives in the U.S. context has traditionally been measured through standardized test scores. Now, teachers, parents, and administrators are questioning whether standardized tests capture the full extent of student learning.
- School budgets are shrinking rapidly, class sizes are growing, and every dollar spent on technology must be weighed against other priorities.

This is a good example of a "complex" evaluation situation where leadership models can add to the evaluator's toolbelt. Let us now read the case using the three theoretical pairs of eyeglasses. The experienced evaluator may already be using one or more of these lenses of leadership, possibly without naming them as such, or even knowing that they are recognized and named. In that case, what follows can have the advantage of making certain aspects of

evaluation practice both named (and thus teachable as such), and potentially more self-aware and self-reflective.

Adaptive Leadership

The Adaptive Leadership approach introduces a useful distinction between technical and adaptive challenges. *Technical challenges* are those which can be addressed by current knowledge and ways of doing things. For example, an evaluator working with the Kyrene District's technology initiative might consider examining the effectiveness of this technology program using already developed instruments, like standardized tests. The methods, processes, and instruments for testing the effectiveness of this technology have already been developed. In contrast, *adaptive challenges* require "changes in people's priorities, beliefs, habits, and loyalties" (Heifetz, et al., 2009, p. 20). For example, this might involve assessing and changing the school's organizational structure, culture, and practices as it loses funding and staff, while still being accountable for improving student outcomes. The Adaptive Leadership lens shows that "making things better" through leadership involves mobilizing others to recognize and address adaptive challenges, not simply the technical issues.

Adaptive Leadership theorists have proposed several methods of addressing adaptive challenges--some of which might be useful to evaluators. A first example says that the practice of leadership includes three stages: the first stage involves "Diagnosing the System" through practices such as making explicit cultural norms, understanding the organization's political landscape, and recognizing "default behaviors" (e.g. conflict avoidance, competitiveness rather than cooperation, etc.) (Heifetz, et al., 2009, p. 49). Second, leadership requires "Mobilizing the System" by slowly moving people toward useful changes in their attitudes, behaviors, and own practices (p. 109). Last, leadership requires conscious self-reflection and thoughtfulness about one's role as a leader and change agent, recognizing that there are risks and dangers to encouraging staff toward creating real, adaptive change rather than using a technical "quick fix". Note that someone acting as a leader might also simply be wrong in her diagnosis, poor in her strategy, and incompetent in making all of this happen. This is a model, not a normative practice.

The distinction between types of tasks offered by the Adaptive Leadership perspective is helpful for evaluators wanting to make evaluation results more useful. By recognizing the program or organization as a "system"--a set of structures, culture, and habituated ways of responding to problems--the evaluator can recognize qualities of the organization that prevent the changes from occurring and she can begin to address these through the evaluation process itself (Heifetz, et al., 2009, p. 49). For example, meetings with stakeholders might be used to move a program into a place where it moves beyond primarily technical considerations and begins to recognize ongoing adaptive challenges that should be addressed. This can launch a discussion on how staff, constituents, and, possibly, evaluators can contribute to making these changes.

Evaluators can utilize this framework to understand and react appropriately and in a timely manner to whatever keeps the organization from changing to address challenges they are presented with. Programs and organizations face a variety of technical challenges and adaptive challenges always in ever-changing and turbulent environments. Evaluators might find opportunities to point out and then work to improve the processes by which programs address technical challenges, thereby helping programs to operate more efficiently and effectively. Additionally, in seeking out and identifying adaptive challenges, or providing necessary input to adaptive problems for programs, evaluators may help programs address deeper and more difficult challenges.

Guiding Questions

When considering the Adaptive Leadership lens, an evaluator concerned with effective evaluation use can ask herself:

- What kind of challenges does this evaluation address: technical or adaptive? What are the consequences of each type of challenge for the type of leadership one should or might choose?
- Can evaluation findings be made relevant and drive change for the type of problem at hand? That is, the results do not provide technical suggestions to solve adaptive challenges, or vice-versa. (For example, the evaluator does not attempt to address the problem of measuring the overall quality of education in the Kyrene School district by suggesting standardized tests as the sole mode of measurement.)
- Can the evaluator open the space to address adaptive challenges as a part of the evaluation process itself? By this we mean, can the evaluator help to "Mobilize the System" of the organization toward adaptive change through how the evaluation is conducted? For example, by including an advisory group or intended users in the evaluation process?

Style Approach

The Style Approach to leadership helps diagnose additional layers of complexity in a leadership/evaluation situation. Like Adaptive Leadership, the Style Approach is not concerned with innate leadership skills or charismatic predispositions. Rather, the Style Approach pinpoints concrete *behaviors* that leaders can engage in to drive change. This approach introduces a distinction between *task behaviors* and *relationship behaviors*, and suggests that "making things better" will involve a mix of activities--some of which are aimed at improving productivity and outcomes, and some of which are aimed at improving relationships between people (Northouse, 2010, p. 69). Some leadership theorists have also described this as a dual continuum of *concern for production/results* and *concern for people* (Blake and McCanse, 1991, p. 26).

The distinction between *task-orientation* and *relationship-orientation* can be a useful heuristic for evaluators when gauging a new organizational context. There is a huge range of attitudes and motivations with which a stakeholder can approach program evaluation, and the Styles Approach to leadership can help illuminate this. For example, in the case of the Kyrene District, a school principal may be faced with improvement mandates or testing parameters set by the district. The principal may choose to invest in and assess the use of classroom technology under the belief that it will raise test scores to the degree required by the district. This is a good example of the "Authority-Compliance" management style, focusing primarily on outcomes measures and external accountability. A principal in a different district may have less external pressure from above, and may be more keen to "integrate the demands from the external pull into the efforts of internal push in an attempt to manage both" (Alaimo, 2008, p. 76). This might involve offering hands-on technology training workshops for faculty, instituting learning communities for staff, and encouraging teachers and students to share best practices in technology use. This corresponds with the "Team Management" style that balances concern for both relationships and results. Yet another principal may be driven less by outcomes and more by image: she may feel pressure from parents, students, or administrative peers to keep up her and her district's image by staying on top of the latest technology trends, but have very little scrutiny surrounding test scores. This corresponds closely to a "Country-Club Management" style (Blake and

McCanse, 1991, p. 29).

Not only can the Styles Approach to leadership help diagnose the organizational context for an evaluation, but it can also inform the types of questions evaluators and stakeholders choose to ask. The Styles Approach suggests that evaluators and stakeholders will want to expand their inquiry to include both task-oriented and relationship-oriented evaluation questions. Turning to our case example, we remember that school technology programs in the U.S. have frequently focused on *task* questions. Concern for results--which, in the U.S. context is defined primarily as improvement in standardized test scores--has been central to the discussion on technology in schools. And yet, though test scores have remained stagnant in the Kyrene District, some teachers and administrators still feel a 'gut instinct' that some kind of growth is occurring--something which evaluators will want to sit up and take note of (Richtel, 2011). This tension between numbers and gut can be instructive.

The Style Approach suggests that evaluators and stakeholders will want to look beyond outcomes measures and examine changes in *relationship behaviors* to help understand the full picture. For instance, evaluators in the Kyrene District may find evidence that technology has fundamentally changed student-teacher or teacher-teacher relationships in the Kyrene District's classrooms--making students more excited to participate, or teachers more willing to collaborate--but that it has had no impact on test scores. It would be useful to explore why this might be the case. An eye to relationship behaviors can help evaluators explain outcomes measures with greater nuance, and even pinpoint what might be missing from traditional metrics such as test scores.

Guiding Questions

Through the lens of the Style Approach to leadership, an evaluator concerned with effective evaluation use can ask herself:

- What kind of concerns does this evaluation address: task-oriented, relationship-oriented, or a mix of both? Are there additional questions the evaluation should address, in order to get a more rounded picture of task and relationship behaviors in an organization?
- Where do the evaluator's stakeholders fit along a continuum from *concern for results* to *concern for people*? Are they predisposed to focus more on *results* or more on *people*? How can an evaluator be mindful of this when presenting evaluation results to stakeholders?
- From the perspective of "making things better", leadership literature tells us that both task and relationship behaviors are necessary for effective leadership and change. How can an evaluator concerned with "making things better" address both *concern for results* and *concern for people* in an evaluation design?

Primal Leadership

A third leadership model that helps diagnose and deal with complex situations is Primal Leadership. This perspective suggests that the emotional and relational work of a leader is the most important act of leadership. Thus, the primary job of leadership is "driving the collective emotions in a positive direction and clearing the smog created by toxic emotions" (Goleman, Boyatzis, & McKee, 2002, p. 5). Basic to this approach is the belief that solving problems requires developing "resonance" with others and using that resonance through appropriate leadership styles to move people to address problems at hand. *Resonance*--the ability to perceive and influence the movement of emotions through followers--is developed

through “emotional intelligence”. As many evaluators know, emotional states of stakeholders and other clients provides important clues to designing, implementing, reporting, and using an evaluation.

The Primal Leadership perspective suggests that evaluators can adopt a series of six distinct “resonant” leadership styles depending on the situation at hand. The *visionary* style moves others toward a shared vision of the future with clear direction. The *coaching* style works with individuals and connects their interests and desires to programmatic or organizational goals. The *affiliative* style connects people to each other and creates harmony amongst groups. The *democratic* style places value on input from others and builds ownership through participation in decision-making. The *pacesetting* style moves people to meet goals. Finally, the *commanding* style gives clear and authoritative direction, but should be applied sparingly and reserved for times of crisis (Goleman, et al., 2002, p. 55).

Turning to our case example in the Kyrene District, a democratic style for conducting the evaluation may most effectively engage parents, students, teachers, and district stakeholders in everything from study design through use of findings. A visionary style, on the other hand, might become a useful rhetorical approach to presenting findings, because it can encourage stakeholders to buy in to a common vision for how the program can improve, how to best implement findings, and how to heed the drive for change.

Guiding Questions

Examining the Primal Leadership model invites the following questions for utilization-focused evaluators:

- What “resonant” styles should be used to build relationships with stakeholders and intended users at each stage of the evaluation so as to increase the likelihood of evaluation findings? For example, the democratic style may be appropriate for developing stakeholder buy-in at the beginning of the evaluation, but might be replaced by a coaching style toward the middle, when the evaluator has a better sense of individual stakeholders and programmatic goals.
- What does the evaluator need to know about the culture and structure of the organization or program in order to present findings in an appropriately “resonant” style?

Conclusion

If evaluation treats each of these leadership perspectives as a set of eyeglasses, putting on each in turn can steer the evaluator toward important considerations to enhance evaluation use. If evaluation has trifocals and puts on all three leadership lenses at once, the result can be incisive and nuanced.

The Adaptive Leadership lens invites us to locate improvement tasks in evaluation use along a scale of increasing complexity, from technical to adaptive challenges. We must resist the urge to over- or under-complicate: an evaluator’s suggestions and solutions should fit the level of complexity at hand. We cannot meet an adaptive challenge with a tool that is strictly technical. Similarly, we cannot become mired down in complexity when a more simple, technical solution is available to us.

The Style Approach to leadership introduces an added dimension to the evaluation equation: *relationships*. The Style Approach suggests that in order to “make things better”, we must apply our competence not only to improving outcomes measures, but also to relationship-building among intended users and other stakeholders. Using evaluation

findings to promote positive change will come as much through people as it will through tools and procedures.

The Primal Leadership lens picks up where the Style Approach leaves off by helping us think about how to engage people and build strong interpersonal relationships. Equipped with an array of leadership styles--from *visionary* to *coaching* to *democratic* and beyond--an evaluator can tailor her style of engagement at each stage of the evaluation process. In order to be used, evaluation must seek strong "resonance" with stakeholders, and Primal Leadership gives us a framework for understanding what it takes to make evaluation resonate.

The field of leadership is broad, and we have tried to distill several key paradigms that will be of interest to the evaluator who wants to (or is called to) see her findings put to use for policy and program improvement. Here, we have merely brushed the surface of the leadership literature and approaches to its practice. By introducing three perspectives on leadership, we made the case that the field of leadership studies can contribute useful distinctions and questions for evaluators who want to enhance the likelihood of their findings being used for positive social betterment in policies and programs. We encourage the curious evaluator to explore this literature more deeply and continue to consider how leadership and evaluation can come into fruitful dialogue about "making things better".

Acknowledgments
Thank you to Michael Baizerman for the review and editing of this paper.

References

Alaimo, S. P. (2008). Nonprofits and evaluation: Managing expectations from the leader's perspective. In J. G. Carman & K. A. Fredericks (Eds.), *Nonprofits and evaluation. New Directions for Evaluation,* 119, 73–92.

Blake, R. R. & McCanse, A. A. (1991). *Leadership dilemmas: grid solutions.* Houston: Gulf Publishing Company.

Day, D. V., Gronn, P., Salas, E. (2004). Review: Leadership capacity in teams. *The Leadership Quarterly*, 15, 857-880.

George, B. (2003). *Authentic leadership: Rediscovering the secrets to creating lasting value.* San Francisco, CA: Jossey-Bass.

Goleman, D., Boyatzis, R., & McKee, A. (2002). *Primal leadership: learning to lead with emotional intelligence*. Boston: Harvard Business Press.

Heifetz, R. (1998). *Leadership without easy answers*. Boston: Harvard University Press.

Heifetz, R., Grashow, A., & Linsky, M. (2009). *The practice of adaptive leadership: tools and tactics for changing your organization and the world.* Boston: Harvard Business School Press.

Johnson, K., Greenseid, L.O., Toal, S.A., King, J.A., Lawrenz, F., & Volkov, B. (2009). Research on Evaluation Use : A Review of the Empirical Literature From 1986 to 2005. *American Journal of Evaluation,* 30(3), 377-410.

Kouzes, J. & Posner, B. (2002). *The leadership challenge*. San Francisco: Jossey-Bass.

Linsky, M. & Heifetz, R. (2002). *Leadership on the line: staying alive through the dangers of leading*. Boston: Harvard Business Review Press.

Northouse, P. G. (2012). *Introduction to leadership*. Thousand Oaks: Sage.

Northouse, P. G. (2010). *Leadership: theory and practice*. Thousand Oaks: Sage.

Patton, M. Q. (2011). *Developmental evaluation: applying complexity concepts to enhance*

innovation and use. New York: The Guilford Press.
Patton, M. Q. (2001). *Qualitative research and evaluation methods*. Thousands Oaks: Sage.
Patton, M. Q. (1997). *Utilization-focused evaluation: the new century text*. Thousands Oaks: Sage.
Richtel, M. (2011, September 3). In classroom of future, stagnant scores. *The New York Times*. Retrieved from: http://www.nytimes.com/2011/09/04/technology/technology-in-schools-faces-questions-onvalue.html
Tourmen, C. (2009). Evaluator's decision making: the relationship between theory, practice, and experience. *American Journal of Evaluation*, 30(1), 7-30.
Zaccaro, S. J., Rittman, A. L., Marks, M. A. (2001). Team leadership. *The Leadership Quarterly*, 12, 451-483.

Appendix A - Selection Criteria for Leadership Theories

In order to select from a large array of leadership theories and models, we developed a set of criteria narrowing the range of analyzed theories to those most directly applicable to increasing use of evaluation findings. In this paper, we address only literature that:

- Offers practical steps and invites useful questions for the evaluator as she considers strategies that are likely to improve the use of evaluation processes and findings to "make things better".
- Focuses on leadership from the perspective of an individual with or without titular 'leader' status. We will not focus on models that emphasize charismatic traits, or on selecting the "right leader" for the job (given that evaluators rarely are in a position of selecting a positional leader).
- Emphasizes leadership as an emergent act that *any* agent (including evaluators) can perform, with the aim of helping individuals achieve a common goal.
- Provides strategies for situations that are not merely *complicated*, but also *complex*: where there is high uncertainty and high potential for social conflict (see Patton, 2011, p. 90). For this reason, we eschew management literature in favor of leadership literature, which tends to focus on highly complex situations where there are high levels of technical and social uncertainty.

Appendix B - Table of Leadership Theories and Guiding Questions

Leadership Lens	Key Idea	Key Questions
Adaptive Leadership (Heifetz, 1998; Linsky, Heifetz, 2002; Heifetz, Grashow, & Linsky, 2009).	Distinction between *technical challenges* and *adaptive challenges*. Provides the evaluator with language to distinguish categories of problems requiring different leadership to address.	• What kind of challenges does this evaluation address: technical or adaptive? What are the consequences of each type of challenge for the type of leadership one should or might choose? • Can evaluation findings be made relevant and drive change for the type of problem at hand? That is, the results do not provide technical suggestions to solve adaptive

		challenges, or vice-versa. ● Can the evaluator open the space to address adaptive challenges as a part of the evaluation process itself? By this we mean, can the evaluator help to "Mobilize the System" of the organization toward adaptive change through how the evaluation is conducted? For example, by including an advisory group or intended users in the evaluation process?
Style Approach (Blake and McCanse, 1991)	Distinction between *concern for results* and *concern for people*. Provides the evaluator with a schema for pinpointing stakeholders' salient concerns, and for expanding the scope of evaluation inquiry to include both results and relationships.	● What kind of concerns does this evaluation address: task-oriented, relationship-oriented, or a mix of both? Are there additional questions the evaluation should address, in order to get a more rounded picture of task and relationship behaviors in an organization? ● Where do the evaluator's stakeholders fit along a continuum from *concern for results* to *concern for people*? Are they predisposed to focus more on *results* or more on *people*? How can an evaluator be mindful of this when presenting evaluation results to stakeholders? ● Leadership literature tells us that both task and relationship behaviors are necessary for effective leadership and change. How can an evaluator concerned with "making things better" address both *concern for results* and *concern for people* in an evaluation design?

Yearning for Learning? Conditions for Learning from Evaluation in Human Service Non-Government Organisations (NGOs)

James Herbert; University of Western Sydney, Sydney, Australia
PhD Candidate in the School of Social Sciences and Psychology

Abstract

The aim to foster learning is prominent in contemporary evaluation theory (Schwandt, 2005; Shaw & Shaw, 1997; Taut, 2007; Torres & Preskill, 2001) and practice (Carman & Fredricks, 2008; Preskill & Caracelli, 1997). However, there are considerable challenges in advancing this intent to actually creating opportunities for learning from evaluation. Drawing on a preliminary analysis of two human service programs delivered by NGOs in Sydney, Australia, this paper will discuss the difficulties involved in promoting learning from evaluation. Under (sometimes) difficult conditions, human service practitioners expressed an acceptance of the aims of evaluation and a willingness to engage with and reflect on evaluation information. Even with this willingness, challenges presented in terms of the way evaluations are conducted and practitioners' experiences of them that affected the potential for learning.

[1]*Keywords:* Evaluation influence; Human Service Practice; Learning from Evaluation

Introduction

Past evaluation practices have been characterised as relying on top-down change (Torres & Preskill, 2001). Driven by the need to engage with organisational context (Preskill, 1991), formative and participatory approaches to evaluation have become increasingly prominent (Shulha & Cousins, 1997). As such, discussions about organisational learning (e.g. Owen & Lambert, 1995; Torres & Preskill, 2001), practice oriented evaluation (e.g. Schwandt, 2005; Shaw & Shaw, 1997), and evaluation capacity building (e.g. Carman & Fredericks, 2010) have become mainstream in the evaluation literature. Evaluation practitioners seek to influence programs, policies, and practices in broad and multifaceted ways and the potential for learning to directly improve the social conditions of service users (Schwandt, 2005; Shaw & Shaw, 1997; Taut, 2007) makes it an attractive mode of influence.

There are inherent challenges in advancing the intent of evaluators to foster learning. While evaluators may value program learning, evaluations still tend to be oriented towards providing upwards accountability (Ebrahim, 2005a), and the information important to funders and program managers (Carman, 2007; Hoole & Patterson, 2008). Even where an evaluation aims to foster learning, this can be confounded by the complex socio-political environment of human service organisations. Efforts to provide human service practitioners with useful feedback about their work is situated in a wider milieu of contestation about the role of NGOs as service providers and their relationship with government funders (Maddison, Denniss, & Hamilton, 2004; O'Shea, 2007). Added to the traditional unfamiliarity of evaluation procedures in the NGO sector (Taut & Alkin, 2003), there are broad institutional challenges in fostering learning from evaluation.

1

[1] *The author would like to acknowledge the assistance of the NGOs involved and their staff that generously provided their time for interviews, and the research supervision of Professor Natalie Bolzan and Dr. Mick Houlbrook.*

From the perspective of human service practitioners, evaluation may be associated with broader developments in human services that have led to the reduction of autonomy and professional discretion (Davies, Nutley, & Smith, 2000). Stepney (2000) suggests that evaluation can serve to reinforce mechanistic responses to service users, and stifle creative and innovative practice. Empirical studies of human service providers' perceptions of evaluation have found that it is seen as an irrelevant undertaking that serves only bureaucratic and managerial imperatives (Meagher & Healy, 2003; Taut & Alkin, 2003).

Practitioner Learning from Evaluation

Conceptualising practitioner learning from evaluation presents a challenge as evaluation is essentially a trans-discipline (Scriven, 2003), working across disciplines with different conceptualisations of learning. While there are some linkages between organisational learning and evaluation (Preskill, 1994; Torres & Preskill, 2001), the evaluation literature also emphasises the importance of building the capacity of individuals to understand and reflect on evaluation information (Carman & Fredericks, 2010). Reflecting the multiple levels at which an evaluation can affect change, typologies of the use/influence of evaluation include individual, group, and organisational processes (Mark & Henry, 2004).

While human services are delivered in the framework of a program, the workers that implement program aims have considerable autonomy and discretion (Evans & Harris, 2004). Organisational learning systems can retain and disseminate information (Daft & Weick, 1984), and affect implicit values and attitudes across members of an organisation (Schulz, 2001; Weick & Roberts, 1993). However changes in practices occur from "the liberation of knowledge from self-reflection and questioning" (Antonacopoulou, 2001, p. 328), which can be prompted by change processes at the individual, group, or organisational level. The consideration of information from an evaluation must be interpreted alongside experiential or constructivist learning "... where the learner reflects on lived experience and then interprets and generalises this experience to form mental structures" (Fenwick, 2001, p. 248). This reflection process encourages individuals to think about their own assumptions and values, but also about the wider context and effect of practice (Schwandt, 2005). While this approach to learning emphasises the agency of practitioners in interpreting new information, this is not to underestimate the influence of organisational systems to create shared mental models that can in turn influence the reflective process (Hedberg, 1981; Pawlowsky, 2001). Torres and Preskill (2001) emphasise the importance of integrating learning with work activities, the culture/systems within the organisation, and the power of learning to align values and attitudes amongst individuals.

This paper represents a preliminary analysis of an extensive case study of two evaluations of human service programs provided by NGOs.

Methodology

The research took place within two NGOs operating in Sydney, Australia, funded to deliver human service based programs. The following criteria were used in selecting programs:

- sustained and established with reasonably secure funding and ongoing support for their continuation;
- partly or primarily funded by an external agency to which the NGO is accountable to;
- involve a human interaction between staff and clients in a community setting as the central program activity; and

- externally evaluated within the past year (or have ongoing efforts to promote the use of previous evaluations) and the evaluation at least in part identified internal learning and improvement as a goal.

Following contact with a number of large NGOs in the state, two programs were selected that were similar in terms of their functioning and structure (table 1).

Table 1. Comparison of Program Characteristics

Program A	**Program B**
Funded following a competitive tending process	Funded through a combination of individual grants and contracts for the delivery of parts of the network
Targets families at risk of involvement in the child protection system	Targets families with identified child protection issues, but also at risk and socially isolated families
Program constituted by the complimentary delivery of family services: case management, childcare, parenting programs, and home visiting	A combination of early intervention services combined with a secondary prevention intensive parenting program
Delivered across five planning areas (metro and regional)	Delivered at one site in outer metro Sydney
Delivered by a large NSW based NGO with an explicit commitment to research and evaluation	Delivered by a large NSW and ACT based NGO with an explicit commitment to research and evaluation
External evaluation conducted by a university research centre and managed by state government funder	External evaluation conducted by a university research centre and funded through a external collaborative fund by the NGO and the university

From each organisation interviews were conducted with two sets of NGO staff[2]. The first group were management or research staff involved in the internal management of the evaluation. These individuals had a high level of knowledge of the planning and conduct of the evaluation, the dissemination of findings and the organisational impacts. The second group were human service staff that had direct experience of the evaluation. They were able to provide a service level perspective, as well as some reflection on the impact the evaluation has had on human service practice in the programs. The evaluation reports, and key organisational documents identified by the participants were also included in the analysis.

The full research will draw on Mark and Henry's (2004) evaluation influence framework in order to provide a context for accounts of human service practitioners' experiences of evaluation. For this paper the analysis concerned purely the identification of factors the participants suggested may have bearing on the influence of the evaluation for learning. Participants' descriptions of how the evaluation influenced the program were collected and reduced to examples where participants indicated that the influence resulted in an affective, motivational, behavioural, or general change (Mark & Henry, 2004) at the practice level. A

2

[1] At this point in the research a total of fourteen interviews across the two organisations have been undertaken. It is anticipated that approximately 25 will be undertaken by the end of the research.

directed content analysis (Hsieh & Shannon, 2005) of these changes and the descriptions provided by participants about why these changes were influential was undertaken organised by the following categories: program context, organisational context, conduct of the evaluation, dissemination, and impacts of the evaluation. At the time of writing at one of the sites, only the practitioner interviews were complete as the organisation was leading into disseminating and discussing the findings of the evaluation.

Findings

The analysis revealed a number of key themes that present significant challenges for the use of evaluations for learning. A brief summary of events at each site provides some context to these.

Study site A
There was a sense that although the NGO was part of discussion around how the evaluation should be undertaken, the interests of the funder drove the evaluation. These interests were not contrary to the interests of the NGO, indeed the purpose of the evaluation from the funder's perspective was to be able to present a case to treasury for a continuation of funding. However, the NGO's interest in fostering internal improvement and learning was in part obstructed by the funder. After it became clear that evaluation data was not going to be shared, the NGO went to some effort and expense to obtain access to site-specific evaluation data in order to foster internal learning. Their work with disseminating and working with the implications of this data is ongoing.

Study site B
The evaluation is still underway, meaning that the data collection for this site is far from complete. The most significant event for the human service practitioners was the release of phase two findings at a conference by the evaluators without consulting them. The dissemination of findings and some of the work around the implications have been delayed due to a number of other ongoing evaluation projects and other broader developments in the organisation. The piloting of social impact bonds in NSW (The Centre for Social Impact, 2011) has opened up alternative systems of funding NGO delivered human services; the program included is one such that has been earmarked by the organization for such an arrangement.

In both case studies, it quickly became clear that what seemed to be fairly ideal conditions for learning from evaluation were in fact much more complicated than anticipated. Both NGOs had stated values around evidence based practice, had strong internal evaluation capacity, and human service practitioners with an interest in feedback about their practice. Despite these conditions, the learning that has taken place so far has been limited primarily due to the approach of the evaluations. While it might be easy to attribute this to the evaluators it must be recognised that what occurred in the evaluation was a product of the interrelationship between different types of organisations and within the different layers of the NGOs.

Influence of the Evaluations

While it is beyond the scope of this article to discuss the impact the evaluations had on the programs, organisations, and individuals involved in detail, it is worth highlighting the influence the participants identified:

- Both evaluations confirmed people's expectations that the program was working,

while also providing improved confidence around the continuation of the program;

- Each evaluation served to highlight some of the challenges the practitioners were engaged with, such as keeping families in the service, particularly indigenous families in case A. Highlighting the challenges for the organisation and for the practitioners themselves in some cases led to reported changes in practices.
- For Case A, the evaluation:
    - Highlighted some of the substantial differences between the communities the program operated within, which had implications for program delivery and the comparability of outcomes across sites;
    - Fostered evaluation capacity through the recruitment of practice support managers, and ongoing interface with practitioners about site level evaluation data. The underpinnings of an evaluation culture were laid through regular discussion of data sets, and the reduction of anxiety about evaluation;
    - Recognised that the client group were beyond the specifications of the program, which meant addressing how to properly support these workers who often had to deal with difficult and potentially traumatic situations;
    - Highlighted characteristics of client groups that had implications for practice;
    - Changed the funder's practices around the referral and intake of clients; and
    - Resulted in part of the program that the evaluation suggested was not effective being phased out;
- While the evaluation for Case B is ongoing and the organisation has yet to disseminate and work with the evaluation findings, participants identified some impacts:
    - Outcomes informed the ongoing planning for the program, in particular the types of data collected for the shift to social impact bonds;
    - Highlighted the advantages of referring within the network, and the importance of the Community Connecter role;
    - Identified the challenges of sharing the space across the programs/services, which served to legitimise some of the difficulties experienced by the practitioners;
    - Resulted in the increased collocation of services; and
    - Resulted in some anxiety about the way human service practice is presented and represented.

Challenges in Fostering Learning from Evaluation

From the two evaluations a number of themes emerged that seemed to challenge the influence of evaluation on practice:

The Political Meaning of Evaluation and Evidence

The political implications of evaluation within organisations and in the broader context of human services create potential challenges in fostering learning from evaluation. Where the evaluation process, results, and dissemination take place without attending to the political context, its influence is diminished (Markiewicz, 2008), particularly where findings come up against strong pre-existing beliefs and attitudes (Christie, 2007; Fleming, 2011). One of the regional managers reported coming up against attitudes amongst practitioners that evidence based practice was 'middle-class' and 'elitist'. The sense that an evidence informed program reflects a managerialist or mechanistic approach to human services fits into a broader narrative around the perceived marginalisation of practitioners and the value of their tacit and

experiential knowledge (Meagher & Parton, 2004). It was also mentioned that parts of the evaluation could be quite contentious, particularly the release of site-level data which could be used to compare different NGOs' performance, and the comparison of government and NGO service delivery. Each of these fit into broader debates around the effects of results based accountability in the human services sector (Carman, Fredericks, & Introcaso, 2008), and about the role of NGOs in the modern welfare state (O'Shea, 2007). Without engaging with broader political issues, evaluation information can easily be dismissed and not enter into discussion and reflection about practice.

Validity of Evidence to Inform Practice

An important precondition for learning from evaluation is the positive appraisal of this information in terms of its validity and usefulness (Christie, 2007; Fleming, 2011). In both evaluations, the human service practitioners had concerns about the ability of the evaluation to reflect and measure their work. A number of concerns were expressed about the validity of the evaluation data: the ability to systematically measure highly individualised behavioural outcomes in children; the lack of an understanding of the context practice was taking place in; the effect of response bias, particularly where caseworkers had been used as defacto-researchers; the long timeline leading to findings that were no longer relevant; the lack of a control group to compare families in the program to; the vastly different contexts and manner of program delivery in case A (which was discussed in the report to an extent); errors in the entry of evaluation data; the lack of systems in place to monitor and prompt the completion of the survey at some sites; inadequate numbers of responses; using pre-assessment vulnerabilities that did not reflect the circumstances of the client; and that the evaluation captured 'the program' in a formal sense, but ignored the impact of informal interactions and promise of safety the centre offered. While some of these were addressed in the evaluation, each of them represents a potential challenge to the validity of the information to inform practice either at and individual or organisational level. The credibility of evaluation information from the perspective of the person was a strong condition for any changes to practice.

The interviews also emphasised how practitioners from different disciplines had a different sense of what types of evidence were valid and meaningful to inform practice. At one site with workers and managers from psychology and health disciplines there was a significant commitment to evidence informed practice, however due to concerns about data quality, the evaluation was viewed with suspicion. The participant described some conflict between psychologists and teachers and some of the different ideas around the meaning and purpose of assessment. Particularly in regional sites where there are challenges in recruiting qualified professional staff, learning seems to have been affected by the limited capacity to process and understand evaluation information. The high rates of staff turnover at these sites reflected in part a clash between the organisational values of evidence based practice and localised practice values, but also ongoing challenges in building receptivity and capacity to respond to evaluation information.

Inter-Organisational Priorities

Both cases described a different arrangement between the NGO, evaluator, and in case A the government service funder. This interrelationship brought many challenges in terms of the planning, conduct, and dissemination of the evaluation, which affected the level of influence the process and findings had on practice. Firstly, while the evaluations had a broad set of priorities and the NGOs were involved in initial discussions about the direction of the

evaluation, the funder's imperatives seemed to take precedence. In case A, while building the evidence base for the program and strengthening the case for continued funding were important to the NGO, the use of the data for internal learning and improvement not only fell by the wayside but was actively resisted by the funder. Secondly, coordination between the three organisations was a constant challenge to the smooth conduct of the program and the evaluation. There were a number of times when there was confusion around who was responsible for what, with the responsibility usually falling to the NGO. In case B, the practitioners expressed an expectation that the evaluator would discuss and feedback findings with them prior to dissemination, and felt very strongly this was part of the work the contracted evaluator should be responsible for. Academics may be under pressure to cease their involvement in contracted evaluations unless there are clear funding arrangements in place for follow-ups or dissemination (Michaux, 2010). There are ongoing challenges for NGOs trying to negotiate their relationship with funders, and the potential for coercion (Ebrahim, 2005b). The clash of these different organisations with very different values and organisational contexts (O'Shea, 2007) creates challenges both for the influence of evaluation, and for the development of a sustainable evaluation culture within NGOs. In short, without the imperative for practitioner learning firmly on the agenda for all organisations involved, evaluations may not produce the information of interest to practitioners or else have limited capacity to disseminate the information through the organisation to the practice level.

Equitable and Participatory Engagement

Across both cases there were significant issues brought up by the participants in regards to the role of human service practitioners in the evaluation. Particularly over the last ten years, evaluation researchers have drawn attention to the importance of process use (Shulha & Cousins, 1997) or the influence that comes from the experience of or engagement in an evaluation. This type of learning in particular is important for the development of evaluation capacity (Preskill & Boyle, 2008). A negative experience of participation in an evaluation potentially reduces the influence of that and future evaluations (Fleming, 2011). In each of the cases negative experiences and the lack of a substantive role for practitioners affected receptiveness to learning from evaluation.

In case A, practitioners were required (through the service contract) to administer a lengthy survey instrument. The inconvenience and challenge of explaining the purpose of the questionnaire to clients was made worse in that there was initially no obvious benefit for either practitioner or client as the data was sent off to the funder. In addition, while the findings of the evaluation were mostly consistent with practitioner experiences, there was some frustration with some of the more contentious findings. There was a sense that more engagement and a greater role for practitioners in the evaluation were needed to better reflect the complexities of practice.

As mentioned in the summary of case B, the most significant part of the evaluation for the practitioners was the presentation of findings, including the description of clients and work-roles, at a practitioner oriented conference. The participants felt aggrieved by this for a number of reasons: they felt that as the staff of the program they should have known what the findings were before they were disseminated broadly amongst their colleagues; they felt that the results didn't reflect well on their work, that the limited change in their clients on the instruments was due to the long term and individualized nature of their work; and that there should have been a place for them to discuss and provide feedback about the data before conclusions were drawn from them.

Conclusion

The paper summarises the preliminary findings from two case studies of evaluations in NGOs providing human service programs. Despite both cases having seemingly ideal conditions for learning, the way the evaluation was conducted and the organisational context seem to have limited practitioner learning from evaluation over the medium term (6-12 months). In particular, a lack of attention to the broader political implications of the findings, different perceptions of the validity of the data to inform practice, the challenge of working between different organizations, and limited engagement of practitioners in the process of the evaluation all served to challenge the possibility of practitioner learning. The full study will attempt to chart the pathways of influence (Mark & Henry, 2004) in these two cases in order to provide a context for practitioners' experiences of the evaluation and the medium term impacts on their daily practice.

References

Antonacopoulou, E. P. (2001). The paradoxical nature of the relationship between training and learning. *Journal of Management Studies, 38*(3), 327-350.

Carman, J. G. (2007). Evaluation practice among community-based organizations: Research into the reality. *American Journal of Evaluation, 28*(1), 60-75.

Carman, J. G., & Fredericks, K. A. (2010). Evaluation capacity and nonprofit organizations: Is the glass half-empty or half-full? *American Journal of Evaluation, 31*(1), 84-104.

Carman, J. G., Fredericks, K. A., & Introcaso, D. (2008). Government and accountability: Paving the way for nonprofits and evaluation. *New Directions for Evaluation, 119*, 5-12.

Carman, J. G., & Fredricks, K. A. (2008). Nonprofits and evaluation: Empirical evidence from the field. In J. G. Carman & K. A. Fredricks (Eds.), *Nonprofits and evaluation, New Directions for Evaluation* (Vol. 119, pp. 51-71).

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions. *American Journal of Evaluation, 28*(1), 8-25.

Daft, R. L., & Weick, K. E. (1984). Toward a model of organizations as interpretation systems *The Academy of Management Review, 9*(2), 284-295.

Davies, H., Nutley, S., & Smith, P. (2000). *What works?: Evidence-based policy and practice in public services*. Bristol: The Policy Press.

Ebrahim, A. (2005a). Accountability myopia: Losing sight of organizational learning. *Nonprofit and Voluntary Sector Quarterly, 34*(1), 56-87.

Ebrahim, A. (2005b). *NGOs and organizational change: Discourse, reporting, and learning*. Cambridge: Cambridge University Press.

Evans, T., & Harris, J. (2004). Street-level bureaucracy, social work and the (exaggerated) death of discretion. *British Journal of Social Work, 34*(6), 871-895.

Fenwick, T. (2001). *Experiential learning: A theoretical critique from five perspectives.* Columbus: ERIC Clearinghouse on Adult, Career, and Vocational Education.

Fleming, M. A. (2011). Attitudes, persuasion, and social influence: Applying social psychology to increase evaluation use. In M. Mark, S. I. Donaldson & B. Campbell (Eds.), *Social psychology and evaluation* (pp. 212-243). New York: The Guildford Press.

Hedberg, B. (1981). How organisations learn and unlearn. In P. Nystrom & W. Starbuck (Eds.), *Handbook of organisational design: Adapting organisations to their environment.* (Vol. 1, pp. 3-27). Oxford: Oxford Publishing.

Hoole, E., & Patterson, T. E. (2008). Voices from the field: Evaluation as part of a learning culture. In J. G. Carman & K. A. Fredericks (Eds.), *Nonprofits and evaluation. New Directions for Evaluation* (Vol. 119, pp. 93-113).

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288.
Maddison, S., Denniss, R., & Hamilton, C. (2004). Silencing dissent: non-government organisation and Australian democracy. *65*. Retrieved from https://www.tai.org.au/file.php?file=discussion_papers/DP65.pdf
Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1), 35-57.
Markiewicz, A. (2008). The political context of evaluation; What does this mean for independence and objectivity? *Evaluation Journal of Australasia, 8*(2), 35-41.
Meagher, G., & Healy, K. (2003). Caring, controlling, contracting and counting: Governments and non-profits in community services. *Australian Journal of Public Administration, 63*(3), 40-51.
Meagher, G., & Parton, N. (2004). Modernising social work and the ethics of care. *Social Work & Society, 2*(1), 10-27.
Michaux, A. (2010). Integrating knowledge in service delivery-land: A view from The Benevolent Society. In G. Bammer, A. Michaux & A. Sanson (Eds.), *Bridging the 'know-do' gap: Knowledge brokering to improve child well-being*. Canberra: The Australian National University E-Press.
O'Shea, P. (2007). A discursive study of institutionalisation in community organisations. *International Journal of Sociology and Social Policy, 27*(11), 483-493.
Owen, J. M., & Lambert, F. C. (1995). Roles for evaluation in learning organisations. *Evaluation, 1*(2), 237-250.
Pawlowsky, P. (2001). The treatment of organizational learning in management science. In M. Dierkes, A. B. Antal, J. Child & I. Nonaka (Eds.), *Handbook of organizational learning and knowledge* (pp. 61-88). Oxford: Oxford University Press.
Preskill, H. (1991). The cultural lens: Bringing utilization into focus. In C. Larson & H. Preskill (Eds.), *Organizations in transition: Opportunities and challenges for evaluation.* (pp. 5-15). San Francisco: Jossey-Bass.
Preskill, H. (1994). Evaluation's role in enhancing organizational learning. *Evaluation and Program Planning, 17*(3), 291-297.
Preskill, H., & Boyle, S. (2008). A multidisciplinary model of evaluation capacity building. *American Journal of Evaluation, 29*(4), 443-459.
Preskill, H., & Caracelli, V. J. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *American Journal of Evaluation, 18*(1), 209-225.
Schulz, M. (2001). The uncertain relevance of newness: Organisational learning and knowledge flows. *Academy of Management Journal, 44*(4), 661-682.
Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation, 26*(1), 95-105.
Scriven, M. (2003). Evaluation in the new millennium: The transdisciplinary view. In S. I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millenium*. Mahwah: Lawrence Erlbaum.
Shaw, I., & Shaw, A. (1997). Keeping social work honest: Evaluating as profession and practice. *British Journal of Social Work, 27*(6), 847-869.
Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: Theory, research, and practice since 1986. *Evaluation Practice, 18*(3), 195-209.
Stepney, P. (2000). Implications for social work in the new millennium. In P. Stepney & D. Ford (Eds.), *Social work models, methods and theories: A Framework for practice*. Lyme Regis: Russell House.
Taut, S. M. (2007). Studying self-evaluation capacity building in a large international development organization. *American Journal of Evaluation, 28*(1), 45-59.

Taut, S. M., & Alkin, M. C. (2003). Program staff perceptions of barriers to evaluation implementation. *American Journal of Evaluation, 24*(2), 213-226.

The Centre for Social Impact. (2011). Report on the NSW Government Social Impact Bonds Pilot. Retrieved from http://www.csi.edu.au/assets/assetdoc/0b6ef737d2bd75b9/Report_on_the_NSW_Social_Impact_Bond_Pilot.pdf

Torres, R. T., & Preskill, H. (2001). Evaluation and Organizational learning: Past, present, and future. *American Journal of Evaluation, 22*(3), 387-395.

Weick, K. E., & Roberts, K. H. (1993). Collective minds in organizations: Heedful interrelating on flights decks. *Administrative Science Quarterly, 38*(8), 357-381.

Employee-driven evaluation in change and innovation – a multi-case study of examining different representations of knowledge

Anne Kallio; Lappeenranta University of technology, Lahti School of Innovation
Anne Pässilä; Lappeenranta University of technology, Lahti School of Innovation
Tuija Oikarinen; Lappeenranta University of technology, Lahti School of Innovation*
**presenting author*

Abstract

Designing evaluation in the workplace is being governed by the conceptions of how knowledge and learning as knowledge creation are seen. In the context of practice-based innovation activities, this paper examines how evaluation is affected by different kinds of representations of knowledge. Evaluation that aims to involve employees as a driver of change needs to tackle the actual work process instead of official work descriptions. Employee-driven evaluation is suggested as a starting point towards a more holistic picture of evaluation in an organizational context. In order for the employees to take an active role in evaluation, empowerment is suggested as (a first) element of employee-driven evaluation. We are suggesting that empowerment evaluation could facilitate the balance between the practical, analytical and interpretative dimensions of work process.

Keywords: Evaluation, Employee-driven evaluation, knowing, Practice-based innovation

Introduction

This paper examines evaluation in heterogeneous knowledge and knowing in the context of practice-based innovation. It highlights the factors that should be acknowledged when designing evaluation that enhances employee-driven change. The concept of employee evaluation is usually attached to performance evaluation, assessment and selection. In this paper the authors state that aside from employee evaluation, an employee-driven evaluation should be considered. In other words, employees become active agents in designing the practices on how their work is being evaluated.

Evaluation can be objective or subjective; objective is based on quantitative operational information whereas subjective is based on collected surveys (e.g. Kemppilä and Lönnqvist, 2003). As a basic function, measurement provides information about factors that are being held important. They signal the employees as to what is important and where they should focus their energies. Furthermore, the information can be used to control that aims are reached and activities are done as planned. Measures and evaluation can also be used for learning and provide enhanced understanding about a phenomenon. (Jääskeläinen, Kujansivu, and Lönnqvist, 2009) Since evaluation (and measurement) has consequences, the benefits and burdens caused by it should be examined before choosing the measures to a situation (Lönnqvist and Mettänen, 2005).

The employee initiated evaluation studies have focused mainly on health care issues and empowerment. Lippin et al. (2000) call for studies that develop methods for assessing the processes of worker-initiated change in the workplace. They conclude that an empowerment-based approach to training (as a part of participative learning process) is effective in creating employee-initiated change. (Lippin et al., 2000)

From an innovation viewpoint, the continuous innovation (Boer et al., 2000; Bessant et al., 2001) and high-involvement innovation (Bessant, 2003) stressed every employee's role in innovation. Employee-driven innovation is driven by the trend of globalisation; the trade unions in Denmark wanted to keep a high employment rate even though relocations to low cost countries (LO, 2007). Thus the employees were not just objective of innovative efforts;

they took more active role as subjects of innovation. In employee-driven innovation the individual does not only take part in innovation but is the driving force (Kesting and Ulhoi, 2010). However, the employees are not used to drive the change. An organization evaluates something that it thinks is important. Therefore, in order to foster employee-driven innovation, evaluation that encourages and empowers employees to continue change efforts is needed.

The evaluation of employees in the workplace is designed according to the explicit work process, whereas the unveiled potential of workplace evaluation is still hidden in the actual work processes. According to Ellström (2010) explicit work process is formally described and the procedures can be codified whereas implicit work process contains the interpretation of an individual on how work is actually executed. Oikarinen and Pässilä (2011) examined that in practice, work has three interdependent layers: Practical, analytical and interpretative. In order to get a more holistic picture of work and evaluation, the characteristics of knowledge and knowing in each layer is examined. In this paper the concept of knowledge is based on knowledge forms according to Heron and Reason (2001): Propositional knowledge, practical knowledge, experiental knowledge, presentational knowledge.

The research question of this study is "How does heterogeneous knowledge affect in designing employee-driven evaluation in change and innovation?" The data is from two action research processes in industrial organizations. Our approach to innovation is practice-based innovation (Melkas & Harmaakorpi, 2012; Ellström, 2010).

First we discuss evaluation in the workplace and present the knowledge and knowing presented by Heron and Reason. Grounding on empirical observations, we discuss the knowing in the three different layers of work, and how evaluation could be designed so that it would support the initiativeness. We conclude by discussing the essential factors in designing employee-driven evaluation practices.

Employees as drivers of change

Evaluation practices are related to general values. The key question is what is being held valuable in current times. In the employee evaluation context, what is the knowledge that is regarded as important; is it more valuable to produce as many end products as possible by myself, or is it more appreciated if we get fewer end products now but since they are made together in groups we can achieve more in the future?

Training is a traditional way to increase the awareness of employees of what are currently the valuable practices in the organization. "The better we are able to identify and refine training evaluation measures of workplace change processes, and to understand them in the unique contexts in which they operate, the better training programs will be able to equip workers, their unions, employers… " (Lippin et al., 2000, p. 706)

Kesting and Ulhoi (2010) state, that employee-driven innovation performance is positively related to the rewards that appreciate the collective innovation activities. They list three important aspects in employee-driven innovation (p.78):

"(1) The general acknowledgement of ordinary employees: Are they regarded as mere inferiors or as partners whose opinion is respected?
(2) The power game: Are employee initiatives perceived as a loss of power and prestige? Are they perceived as an attack on management's authority (since they question existing practice)? In this respect, do managers suppress innovative talent from below only to ensure their own interests or reinforce existing positions?
(3) Failure culture: Can failures be utilized as weapons in internal political quarrels?Are failures tolerated and accepted as potential stepping-stones to success?"

Kristiansen and Bloch-Poulsen (2010) talk about employee-driven innovation in the context of communication and dialogue. They state that in order to facilitate employee-driven innovation, it is important to create dissensus into communication. Dissensus, as opposed to consensus means *"that team conversations must be organized in ways where silent or unspoken, critical voices speak up"* (p. 156). Organisations can thus develop skills of "dissensus sensibility to open up for more voices, for indirect criticism, and for more democracy in the decision process trying to balance dialogues in multidimensional tensions between consensus and dissensus" (p. 156).

It is to be considered whether evaluation is focused on individual worker behavior or willingness of employees to raise or win concerns (Lippin et al., 2000). As Kallio et al. (2012) state, it is important that employees feel they are evaluated by factors that they feel are important. In organizational context, the individual is only part of it. However, individuals form groups and organizations, and their incentives do have an effect on the outcomes.

Oikarinen and Pässilä (2011) have examined the layers that are involved in the actual work process. In figure 1 the three levels are presented: Practical, analytical and interpretative. The analytical layer consists of documented practices. It is much researched and there are measures to evaluate that. The practical layer includes how the work is done as in routines and situations. In the bottom of work processes there is the interpretative layer that has an effect to the other layers. Interpretation deals with the human aspect of work process, for example power tensions, emotions and social interaction.

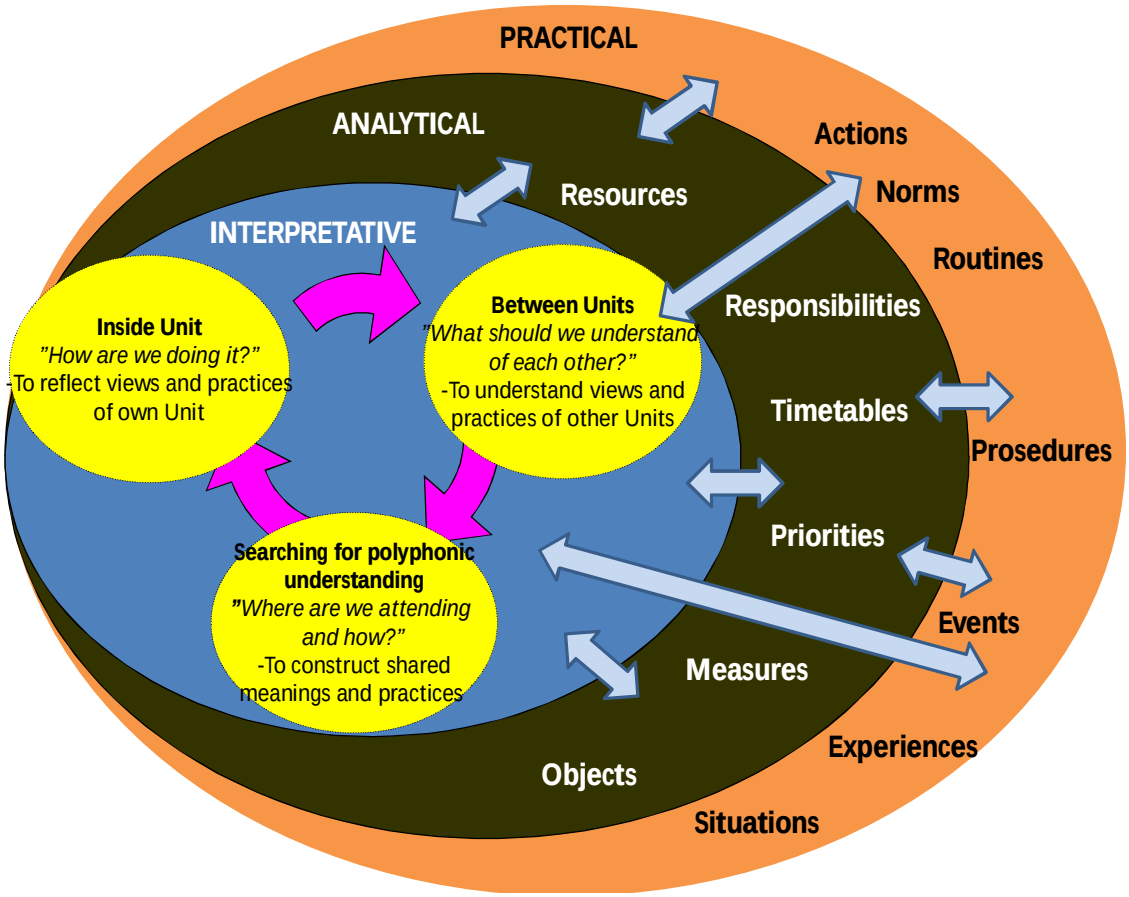


Figure 1. The layers of actual work process (Oikarinen & Pässilä, 2011)

Evaluation should somehow acknowledge individual aims and targets. Everyone should be

able to recognize themselves in the joined targets that groups or organizations have. Empowerment evaluations offers and insight into how employees could take the change as their own matter. However, the evaluator does not empower anyone, people empower themselves often with assistance and coaching (Fetterman, 1996a). Empowerment evaluation should take process and outcome into consideration. The snapshots should then be placed within the process where the individual/group is. In empowerment literature the focus is on both the means by which employees are attempting to drive change and the outcomes as consequences of these efforts (Fetterman, 1996b). In building evaluation that aims to empower employees, some themes emerge: an emphasis on building on strengths (rather than focusing on problems and finding fault), collaboration, participation and self-determination (Fetterman, 1996b, p.380). The evaluation is an activity of a group, not an individual (Fetterman, 1996a). Together the groups form a holistic picture what seems to be the state of things in the current case.

Evaluation is traditionally based on recognizing causalities. However, complexity is present in every way and this should be acknowledged. Instead of presenting causality, complexity should be examined. How to shift from static evaluation to more complex contextualized evaluation in uncertain environments? A metaphor for evaluation could be like throwing a small stone to still water. When you hit the water surface, it creates waves around the spot you hit. Sometimes in evaluation these waves are hidden from the management, i.e. those who traditionally conduct the evaluation. And if the stone is continuously thrown into the same spot, eventually the significance of the waves can grow. Probably these waves cannot be fully taken away, but they need to be recognized. The movement in each of the layers presented above has an effect on the other layers.

Knowing and knowledge in the work process

The potential of heterogeneous knowledge in the creation of new knowledge and knowing is well recognized in the field of research methodology (e.g. Taylor, 2003; Heron, 1992; Heron & Reason, 2001), management and organization theory (e.g. Phillips, 1995; Vickers, 2008; Schreyögg & Geiger, 2006; Hargrave & Van de Ven, 2006), and learning and innovation literature (e.g. Nonaka & Takeuchi, 1995; Cook & Brown, 1999; Amin & Cohendet, 2004).

Knowledge and knowing in this study is based on Heron's and Reason's division (Heron, 1992; 1996; Heron and Reason, 2001) of heterogeneous knowledge. The four natures of knowing are experiential, presentational, propositional and practical and they each provide incomplete understanding on their own but form holistic picture together. According to Heron (1996) the integration of different forms is created as following: experiential knowing through meeting and encounter; presentational knowing through the use of aesthetic, expressive forms; propositional knowing through words and concepts; and practical knowing-how in the exercise of diverse skills.

Table 1. Type of participation in a process of knowing by Heron (1996) and Heron and Reason (2001, p. 184)

Nature of knowing	Participation of knowing	Congruence of knowing
Propositional knowing	= "about" something; knowing through ideas and theories, expressed in informative statements = involves seeing the entity in terms of the concepts and identifying names that come with the acquisition of language, as having certain describable qualities and as being in certain describable relations with other entities. = expressed in statements that something is the case	knowing understood through theories which make sense
Practical knowing	= "how to" do something, is expressed in a skill, proficiency, knack or competence, whether physical and/or mental = cannot be fully reduced to any set of written instructions = is grounded on and empowered by all the prior forms of knowing, and is immediately supported by propositional knowing	knowing expressed in worthwhile action
Experiential knowing	= emerges through direct face-to-face encounter and interaction with a person, place, process or thing; = is knowing through the immediacy of perceiving, through empathy and resonance, through sustained acquaintance = creative shaping of a world through transaction of imaging it and participative empathy, through which we commune with the inner experience of beings, their mode of awareness. The transaction of imaging a world is not restricted to sense perception, but includes productive imagination and extrasensory perception	knowing grounded in experience
Presentational knowing	= expressing meaning and significance through forms of imagery like movement, sound, drawings, paintings, photos, sculptures, musical forms, mime, dance, ritual, poetry, story, allegory, drama, thick multi-sensory descriptions = an intuitive grasp of the significance of patterns as expressed in nonverbal art-forms and verbal forms used for expressive, evocative-descriptive and metaphorical effect = emerges from experiential knowing, and can never be fully accommodated in language	knowing expressed through stories and images

Methodology

To explore how to evaluate heterogeneous knowledge on the different layers of work processes we studied two case organizations. They both operate in forest industry in Finland. Both have been part of a longer research and development processes. Even though this paper presents only part of the rich data gathered, the whole processes have an effect on how the researchers analyze and interpret the data excerpts.

The first case company (Case 1) is in the forest industry. Five different units participated in the action research process; four factories and one administrative unit. All four factories are situated in different locations. The focus is on studying the analytical layer. Initially twenty semi-structured interviews were conducted in five units. Sixteen of the interviewees were from the shop floor and four from management. For this paper, four of them are at the centre of attention, while the rest have had an effect analyzing the data. The interviews included open questions about individuals' facilities and motivation to give ideas, awareness of the idea management system and perception of the atmosphere in general. Interviews were recorded and the length varied from 60 to 90 minutes. The interview material was analyzed with help of ATLAS.ti software. The data was gathered in spring 2007.

The second case company (Case 2) is a family company with 1.000 employees and a history spanning a century. Case 2 focuses on studying the interpretative layer. The data was collected by group interviews and facilitated by storytelling techniques. We organized separate group interviews for the representatives of sales, production and R&D. The management group in its entirety participated in the group interview. The chair of the board was interviewed personally. Altogether, there were 48 informants and the interviews lasted 22 hrs. They were videotaped and observed by a researcher. The data was gathered during April-

May 2011.

These two case organizations form the unit of analysis and the paper thus relies on a two case study setup. Following Yin (2003) different methods can be applied when carrying out a case study that first and foremost signals the boundaries of the phenomena in question. These different ways of collecting the data, as well as the nature of the processes illustrates the different knowing that is generated with using different methods.

In this paper we use cross-case analysis. The knowledge and knowing is examined to identify factors that have an effect to whether employees take initiative to drive change for innovation.

Knowing in the layers of work process and its effects on employee initiativeness

In order to involve employees to be the drivers of change, a more holistic understanding is needed of the underlying knowing of work processes. Which kind of knowing is attached to the layers of work process?

Practical layer

The external layer and practical layer represent something that can be seen and described, even though they require action in order to become visible. This kind of action can be for example routines or events. Thus the knowing that is characteristic to this layer is practical knowing in the exercising of different skills. This can be observed, even though the interpretations of the observations are dependent on who interprets. However, the interpretation of an observation can be reduced and thus make the observation more objective.

Experience is an essential part of work processes and should be acknowledge in organizations. As new technology takes over, the skills needed in the work may change. Still, many repetitions over time and seeing many different situations can be acquired only through time.

"Even though you have studied as an engineer and you have the theoretical knowledge of the work process, before you really have assimilated what needs to be done, it takes about three to five years… before you can say how this should be done and not asking how things are done here."

However, the experience does not always come with age. In fact, there are situations where a long career in too narrow a field can lead to an inability to comprehend the bigger picture.

"That the beginning (of the production process) knows what happens in the end. But, sometimes we encounter people who have been working for 30 years in that department and they have no clue what happens in the beginning… "

There is a lot of small group activity going on which may not be documented.

"It is this kind of development work and… that employees can be involved in the purchasing and designing. … they are not submitted as suggestions but done in small groups. … those are very practical things that we have done in those small groups… that… they just have been implemented on the spot."

These kinds of situations are usually documented only if they end up e.g. submitting a suggestion. However, this activity spreads practical knowing from one person to the whole group. Recognizing these situations and upraising the value in them facilitates the organization eventually to maintain their level of practical knowing even though individual employees would leave the organization.

Analytical layer

The analytical layer presents material, measurable knowledge. The knowing that is governing this layer is propositional knowing through words and concepts. It is assumed that anyone can observe the same things in the same circumstances.

Suggestion system is one of the examples that aim to engage employees to take part in the development activities of an organization. There evaluation is based on knowledge that is made as computational/ calculative as possible. For example, measuring the amount of suggested ideas, products, actions or events which are considered as valuable.

".. first we make an action plan and then we make the reports and… this activity is stated in the law… … we count the activity to make suggestions, the amount of them..
..How much did the number of suggestions increase within a year, how much did we pay rewards. Then we measure the ratio between rejected suggestions and rewarded suggestions…
… this new practice enabled us to… …so the desicion comes faster, and the person who has made the suggestion will be paid faster…
… we also have these widely effecting suggestion that we have been obliged to follow-up even 2-3 years… … the person who made the suggestion can wonder why it takes so long. But we have to be certain if we are going to pay thousands of euros. That it really is that effective.
(Industrial factory, Interview of the chairman of the suggestion board in 2007)

The suggestion box is one channel for the employees to take part in organisation's innovation activities. However, the evaluation of amount of initiatives can lead to the situation that others try to create as much initiatives as possible. The critical question hereby is whether employees are creating ideas for the suggestion box or are they creating ideas how to renew their own work practices. Also, the paid rewards of the initiatives encourage employees to do the things that are rewarded. In other words, what is rewarded is being held valuable in the organization.

Interpretative layer

The interpretative layer is not visible. In fact, some aspects of it may not be recognized in the everyday activities at all. Some methods, for example research-based theatre (Pässilä et al., 2012; Pässilä and Oikarinen, forthcoming) help to reveal this layer into a more conscious level. In interpretative level two types of knowing are highlighted: experiential knowing through meeting and encounter and presentational knowing through the use of aesthetic, expressive forms. The observation of interpretative layer is always socially constructed and dependent on the viewpoint of interpreter.

The next is taken from a work story session, where participants ponder their own work practices through visual method called theatrical image.

Alice: I have an image where three fellows are trying to sit down at the same box. They are three salesmen who have all sold the same box and now compete with whose client will get it. Salesmen sell eagerly but don't communicate among themselves. They may sell the same product many times. Then they quarrel whose customer will get the product and whose customer will be disappointed. And whose fault it is, you always have to find the one to blame.
John: But the reason behind this is the production-oriented way of working: you have to sell what is produced, what is stocked. If we operate we should have the on-line information of the stock situation.
Jack: Or you should be able to book on-line the stuff you have sold and block it so that it can't be sold many times.

Alice: Yes, and if there is the situation that the same product is sold many times, we should have rules on how to decide which customer will get the product. We should agree on the practices together so that all the departments will have the same practices.
Iris: But to agree on the shared practices demands work, meetings, new ways of working -
Jack: - and organizational changes
Jane: - and that people change

The images enable the employees to take a step back and look at their own practices in a metaphorical level. More sensitive issues can be brought up. Thus, expressive forms of presentational knowledge help to make existing cognitive structures and presuppositions visible, discussable, and eventually changeable (Oikarinen and Kallio, 2012). Also, presentational knowing can also act as a bridge between the other forms of knowing and thus facilitate creation of new knowledge (Heron and Reason, 2001; Oikarinen and Kallio, 2012).

Discussion and conclusions

This paper aims to open a conversation about employee-driven evaluation aiming at make the employees as drivers of change. More specifically, it presents the three interdependent layers of work process and knowing what characterizes each layer.

A holistic approach on employee evaluation would in its best bring up all the three layers of work process. These can be better examined through the knowledge and knowing that characterizes each layer. For the practical layer, practical knowing in the exercising of different skills is characteristics. Propositional knowing through words and concepts is governing the analytical layer. In the interpretative level two types of knowing are highlighted: experiential knowing through meeting and encounter and presentational knowing through the use of aesthetic, expressive forms.

Evaluation includes a reference to values: what is being held important is evaluated. Not all values are said out loud, part of them are hidden in everyday processes and work roles. For example, in the case of salespeople, the interpretative layer revealed that there are different expectations as others prioritize customer needs, other production deadlines, etc. In the case when evaluation is conducted only based on practical or analytical layer, there are those who do not feel the evaluation is suitable for their work. Thus, empowerment evaluation is part of building suitable evaluation practices as every employee will have the chance to reflect on work practices. In the process of building evaluation practices emerge knowing which is embodied in people and practice, interaction and the complex relationships.

This paper suggests factors that should be taken into consideration when designing evaluation practices that aim to employee-driven change in innovation.

- Employee evaluation should include both; process efforts and outcomes (Fetterman, 1996b)
- Evaluation should be dynamic and participative
- Empowerment evaluation could facilitate the balance between the practical, analytical and interpretative dimensions of work process
- Use of presentational knowing in evaluation could help bridging between other forms of knowing (Heron and Reason, 2001; Oikarinen and Kallio, 2012)

In the future, more studies, both theoretical as well as empirical, should be focused on the thematic of holistic evaluation. How evaluation could be designed so that it takes into consideration both; individual process of development and organizational growth targets. Future aspects of evaluation could include for example, the use of wisdom of groups aside of expert evaluation.

References

Bessant, J., Caffyn, S. & Gallagher, M. (2001). An evolutionary model of continuous improvement behavior. *Technovation, (21),* 67-77.

Bessant, J. (2003). *High-Involvement Innovation: Building and Sustaining Competitive Advantage Trough Continuous Change*. Chichester: John Wiley & Sons.

Boer, H., Berger, A., Chapman, F. & Gertsen, G. (2000). *CI changes: from suggestion box to organisational learning. Continuous improvement in Europe and Australia*. Ashgate Publishing Company: Aldershot.

Cook, S. & Brown J. (1999). Bridging epistemologies: the generative dance between organizational knowledge and organizational knowing. *Organization Science, 10(4),* 381-400.

Ellström, P-E. (2010). Practice-based innovation: a learning perspective. *Journal of Workplace Learning, 22(½),* 27-40.

Fetterman, D. (1996a). *Empowerment evaluation. An introduction to theory and practice*. In Fetterman, D., Kaftarian, S. and Wandersman, A. (Eds.) Empowerment evaluation. Knowledge and tools for self-assesment & accountability. London: Sage Publications. 3-46

Fetterman, D. (1996b). *Conclusion. Reflections on emergent themes and next steps.* In Fetterman, D., Kaftarian, S. and Wandersman, A. (Eds.) Empowerment evaluation. Knowledge and tools for self-assesment & accountability. London: Sage Publications. 379-384.

Heron, J. (1992). *Feeling and Personhood: Psychology in Another Key*. London: Sage.

Heron, J. (1996). *Co-operative Inquire: Research into the Human Condition.* London: Sage.

Heron, J. & Reason, P. (2001). *The Practice of Co-operative Inquiry: Research ´with` rather than ´on` people'* in, In P. Reason & H. Bradbury (Eds.), *Handbook of action research. Participative inquiry and practice*. London: Sage. 144-154.

Jääskeläinen, A., Kujansivu, P. & Lönnqvist, A. (2009). Learning from existing performance information in public service organizations, Paper presented at the 5th *Conference on Performance Measurement and Management Control.* September 23–25, 2009. Nice, France.

Kallio, A. (forthcoming). Enhancing absorptive capacity in a non-research and development context- An action research approach to converting individual observations into organizational awareness. Dissertation. Lappeenranta University of Technology.

Kallio, A., Kujansivu, P. & Parjanen, S. (2012). Locating the Weak Points of Innovation Capability before Launching a Development Project. *Interdisciplinary Journal of Information, Knowledge, and Management*, 7, 21 – 38.

Kemppilä, S. & Lönnqvist, A. (2003). Subjective Productivity Measurement. *The Journal of American Academy of Business, Cambridge, 2(2),* 531-537.

Kristiansen, M. & Bloch-Poulsen, J. (2010). Employee driven innovation in team (EDIT) - innovative potential, dialogue, and dissensus. *International Journal of Action Research, 6(2-3),* 155-195.

Lippin, T., Eckman, A., Calkin, K. & McQuiston, T. (2000). Empowerment-Based Health and Safety Training: Evidence of Workplace Change From Four Industrial Sectors. *American Journal of Industrial Medicine, 38(6),* 697-706.

Lönnqvist, A. & Mettänen, P. (2005). *Criteria of Sound Intellectual Capital Measures, Business Performance Measurement: Intellectual Capital - Valuation Models*. Le Magnus University Press, 97 - 120.

Oikarinen, T. & Pässilä, A. (2011) Presentation slides.

Oikarinen, T. & Kallio, A. (2012) Absorption and creation of new knowledge – A multi-case study of different forms of knowledge impacting on absorptive capacity. A conference paper presented in OLKC, Valencia 25-27th April.

Pässilä, A., & Oikarinen, T. (in press). *Research-based theatre as a facilitator for organisational learning*. P. Meusburger, A. Berthoin, & M. Ries (Eds.), Learning organizations: The importance of place for or-ganizational learning. Dordrecht: Springer Verlag.

Pässilä, A., Oikarinen, T., & Vince, R. (2012). *The role of reflection, reflection on roles: Practice-based innovation through theatre-based learning*. In H. Melkas & V. Harmaakorpi (Eds.), Practice-based in-novation: Insights, applications and policy implications (173-191). Heidelberg: Springer.

Vince, R. & Pässilä, A. (forthcoming). Aesthetic distancing and critical reflection.

Robson, C. (2002). *Real world research, second edition*. Oxford: Blackwell.

Saunders, M., Lewis, P. & Thornhill, A. (2009). *Research methods for business students*, fifth edition. Italy: Prentice Hall.

Yin, R. (2003) *Case study research, design and methods, 3rd ed*. Newbury Park: Sage Publications

II

Ethics and Methodological Dilemmas of Evaluation

The slippery slope of evaluation: Ethics, issues, & methodological challenges using the case study of a housing development

JoDee Keller, PhD, Pacific Lutheran University
Janice Laakso, PhD, University of Washington, Tacoma
Christine Stevens, PhD, University of Washington, Tacoma
Cathy Tashiro, PhD, University of Washington, Tacoma

Abstract

Using an evaluation of a public housing development in the United States Pacific Northwest as a case study, the authors explore the politics of evaluation as well as ethical considerations and methodological challenges in evaluation research. As this particular housing development was made up of a large percentage of immigrants and refugees, the authors also discuss the importance of developing methodologies that give voice to culturally and economically diverse populations. Notably, ethical considerations occur at multiple levels, from the broad assumptions underlying housing policy, to the vulnerability of public housing residents, particularly with regard to health issues, to the redevelopment process itself, to hardship faced by individual residents. The authors raise a number of questions that are important to consider in conducting ethical research with marginalized populations.

Keywords: evaluation research, public housing, ethics, methodological challenges

Introduction

Housing Opportunities for People Everywhere (HOPE VI) is a competitive grant program initiated in 1992 by the U.S. Department of Housing and Urban Development to eradicate "severely distressed" public housing and disperse pockets of poverty by creating new mixed-income communities. It was expected that mixed-income housing would provide better quality housing for low-income families, and physical and economic revitalization of the inner city (Joseph, Chaskin & Webber, 2007). A primary goal of HOPE VI was to support individual and family mobility out of public housing and toward self-sufficiency (Wexler, 2001). This program has radically changed the face of public and affordable housing in the U.S., resulting in the displacement and relocation of tens of thousands of low-income households. This paper focuses on emergent issues from a 2-stage evaluation using mixed methods to assess the impact of relocation on residents of the Tacoma, Washington Salishan HOPE VI project. Before HOPE VI, Salishan residents were poor, culturally diverse, and many were disabled and/or elderly. When relocation began, over half were refugees, many not fluent in English.

The interdisciplinary evaluation team consisted of researchers from three different Pacific Northwest universities and translators from the different language groups represented in the Salishan Housing population. It was important for the translators to be team members from the beginning to guide the methods and development of the interview instruments to be culturally sensitive.

HOPE VI required evaluation of self-sufficiency, housing, and service outcomes for housing residents who were relocated during demolition and reconstruction of housing units. In addition, the evaluation team planned to explore health effects of relocation as well as to promote discussions among residents, housing staff, and administration (Bodonyi, et. al. 2007; Stevens, 2010). In 2003, the Public Housing Authority (PHA) began resident relocation

as the demolishing of the old housing began.

The methods in this evaluation included stakeholder and community leader interviews, interviews with case managers, Geographic Instrument Mapping (GIS), key economic and census data. While these methods provided an overview of the 811 families in the Salishan housing community, the evaluation team wanted to interview a representative sample of families that reflected the range of populations. An indepth health status and healthcare needs survey with 60 households was conducted in 2005 with residents who had not yet been relocated.

Additionally, the team wanted to explore in-depth several key areas that included current housing situation, strengths and challenges of the old and new housing community, sense of community, how relocation information was shared with residents, community and support services and health status of family members as well as contextual factors such as immigration, socioeconomic status, disability, gender, and age. Recruitment began with letters sent by the PHA to all residents as well as announcements in community meetings in their primary language – Vietnamese, Cambodian, Russian and English - explaining that a member of the evaluation team would contact them in their primary language to invite them to participate in interviews and schedule appointments. Fifty-two families agreed to be interviewed in 2007 and 2009. At the first interview, the participants were asked if they wanted to participate in photovoice project. In addition to the interviews, focus groups were held with different populations in Salishan.

Overall, respondents had positive comments about the relocation process, feeling they had been provided sufficient help with moving expenses and explanation of the process. They generally found better physical living conditions but struggled to pay utility bills. However Salishan, before HOPE VI, had a strong sense of community and strong social ties, and these have been damaged. Also, some residents who had requested a return to the new Salishan were found to be ineligible due to income requirements, insufficient suitable housing for elderly and disabled or ineligibility based on unacceptable behaviors of family members. Residents who relocated to the new Salishan have found an absence of community and the addition of social controls that restrict favored activities, thus reducing opportunities for social interaction. Focus group members noted that there is a sense of isolation in new Salishan since community activities have decreased. Based on other studies of HOPE VI projects, there is limited evidence that mixed income neighborhoods increase social mobility.

Using Salishan as a case study, the interdisciplinary team of researchers from the University of Washington Tacoma and Pacific Lutheran University, examine (1) general ethical issues for evaluators in human services; (2) methodological challenges and opportunities in conducting research with multicultural populations and the ethical implications of methodology; (3) the poor health of public housing residents as the embodiment of structural inequality; and (4) flaws in housing policy based on ideological assumptions about the poor. Each of these areas raises ethical issues.

Ethical Issues for Evaluators in the Human Services

Evaluation is by its very nature a political act. While Royse, Thyer, Padgett & Logan (2006) assert that the ideal goal of evaluation research is to provide the highest quality services to clients and consumers, other motives invariably find their way into the research process, including personnel issues and countering negative media attention. Evaluation is intrinsically threatening to those being evaluated, who want to be reviewed positively and want the best possible outcomes. “Any evaluation may be perceived as a political activity by those being evaluated” (Royse, et. al., 2006, p. 378). Thus, administrators and staff may want to stifle the conversation if there are concerns that the program may not be working as

intended.

In the case of the Salishan study, line staff, in preliminary interviews, were very open in identifying concerns with the implementation of the HOPE VI redevelopment and its effects on residents, anticipating challenges around such issues as paying for utilities while off site, being promised that they could return to the site following redevelopment even though it was unlikely that most would, and having to adjust to less understanding landlords. Royse, et. al. (2006) suggest that in evaluation research, some staff will be helpful, others will not, for a variety of reasons, including fears about results and the feeling that the evaluation is being imposed upon them. The evaluation team in Salishan had difficulty obtaining basic, potentially helpful information from staff members, particularly mid-level supervisors, for a variety of reasons, illustrating some of the challenges that can occur in evaluation research. These staff may have felt vulnerable, not knowing what they could and could not share. Additionally, staff turnover meant that the team was working with several different HOPE VI coordinators as well as housing managers. Also, some data were simply not accessible because of the way in which the PHA collected and stored data. Regardless of the reasons, this lack of data compromised the quality of the evaluation, as some important information simply was not made available.

Royse, et. al. (2006) also identify the agency's investment in obtaining the best possible results because of its importance for future funding. Again using the Salishan example, at a very tense midpoint meeting of evaluators and PHA administrators, it was clear that the administrators did not approve of the draft evaluation report for a number of reasons. A draft was presented because the evaluators wanted feedback from the housing authority to shape the final report. Though a draft, the administrators used this opportunity to critique the report as if it were the final product. Beyond that, though, the report and verbal presentation identified both positive and negative aspects of the PHA's implementation of the HOPE VI program, even making suggestions for mid-course corrections. The administrators desired a report based on quantitative outcome data, such as test scores at a local school, though those results could not have been linked to the housing redevelopment at this stage of the redevelopment process. Ironically, much of the outcome data they desired needed to be based on data they had, and which they had been unable to provide us, despite numerous requests! The evaluators, though, felt a commitment to making recommendations for mid-point corrections.

Additionally, the team chose mixed methods, including quantitative as well as qualitative data. The team members were committed to a qualitative approach in order to include the stories of the residents. The agency administrators, however, were less interested in this approach and expressed preference for statistics on employment of residents, housing outcomes, etc.

Evaluation must be situated within contexts of culture and structure (Schwandt, 2007). In any evaluation, it is not only beneficial but also is an ethical imperative to obtain the perspectives of a variety of stakeholders (Greene, 1997). In addition to housing residents, the team interviewed community and government leaders and service providers, and housing authority line staff. However, this raises the ethical question of whether equal inclusion of stakeholders who are inherently unequal privileges the powerful, and whether ethical evaluation inquiry demands advocacy for the least powerful.

Specific ethical challenges

Beyond the political challenges involved in evaluation research, specific ethical concerns may emerge. Holosko, Thyer & Danner (2009) identify the NASW Code of Ethics research guidelines. First among those is considering the consequences of participation, fully

informing participants of risks and benefits up front, and taking precautions to protect them. However, as the team discovered, it is not always possible to accurately anticipate the effects of participation, particularly with vulnerable populations who may have experienced significant prior life trauma. During interviews, a number of participants identified sometimes quite intense feelings of grief and loss around the forced relocation. In one focus group, an adolescent commented on how he didn't realize, until talking about it, how sad he was about what happened to the community where he was raised. In another focus group of older women, all long-term residents, several commented on their sadness in talking about the community that used to be (Keller, 2011). Though this sadness existed prior to the evaluation, the process of evaluation highlighted and intensified these painful emotions for participants. Other emotional responses, including anger at the PHA staff, emerged as well. In light of the edict to do no harm to participants, evaluators were left with challenges in how to manage these intense feelings. Finally, the bi-lingual, bi-cultural interpreters/research assistants, too, were exposed to personally challenging situations. Many participants were refugees with traumatic histories, and the interpreters shared this past. When questions were asked about residents' housing history, it was often very emotional and difficult for the interpreters, as well. In fact, as the effects on the interpreters/assistants intensified following multiple interview**s**, the team found it necessary and important to debrief with the interpreters. This debriefing process broadened the understanding of the evaluators, but also gave the interpreters the opportunity to process feelings.

Another challenge, was the process of getting the research proposal and instruments approved by the Institutional Review Board (IRB). Swauger (2009) asserts that ". . . local IRB members, usually from medical and hard science fields and concerned with legal implications of human subject protection, often approach research from positivistic and quantitative stances." (p. 65) She articulates the challenges for feminist qualitative researchers, stating that ethics are never fixed but continuously are reflected upon as the process unfolds. If a proposal passes the IRB, it is assumed that one has an ethical project, but "feminist qualitative researchers are concerned with deeper ethical standards throughout all stages of the research process." (p. 66).

Swauger (2009) also states that the knowledge claims of disadvantaged groups have been ignored historically. Traditional research tends to speak about these populations, but feminist qualitative researchers want to speak for them and with them. The Salishan evaluation team sought a strategy that empowered housing residents rather than silencing them or losing their stories. In addition to semi-structured interviews, the team utilized focus groups and photovoice to obtain information. These methodologies can be empowering, giving some control over the structure and course of the interview to the participants. Indeed, in one of the focus groups, participants were able to question each other and the evaluators ("I want to know how this is going to benefit me"). However, the IRB, to protect participants, changed the format of photovoice, ultimately making it a less empowering process for housing residents. Photovoice is intended as a methodology whereby participants join together in a small group to tell stories about their photos. The IRB, presumably to ensure confidentiality of participants, did not allow for the discussion of photos in groups, but rather required participants to meet individually to share photos with an evaluator. This changes the nature of the conversation and the potential mutual support and empowerment of a group.

Additionally, the team discussed the impossibility of being totally objective or bias-free. Team members felt concern for participants, providing lists of resources, and for one resident, a requested case of nutritional supplement. The Salishan evaluation raises a number of questions, including how to manage the tension between advocacy and evaluation, the (im)possibility or even desirability of objectivity, and how to effectively bring concerns to the attention of policy makers.

Methodological Challenges & Opportunities in Research with Multicultural Populations

There are a number of challenges in accurately representing the perspectives and concerns of vulnerable populations, including people of color and low-income persons. Swauger (2009) suggests that people from disadvantaged backgrounds may be skeptical of researchers, fearing exploitation or misrepresentation. Indeed, historically, there are multiple examples of research abuses and exploitation with vulnerable populations. In Salishan, one challenge was in designing and implementing methodologies that would accurately give voice to the residents. As the residents were from a number of different cultural and language backgrounds, the team worked to develop a questionnaire that could be implemented with speakers of four different languages, and found that there were concepts that could not be translated consistently.

Royse et. al. (2006) make the point that social programs often are designed for white, English speakers. The tools and methods that measure effectiveness of these programs can be biased, as well. It is important to obtain as much cultural knowledge about groups being investigated as possible so that recruitment procedures, etc. are applicable across groups. The team made a point of researching the immigration history of the different cultural groups represented in the housing development, even noting differences among different waves of immigrants from the same country, as a way of enhancing understanding of their experiences.

Royse, et. al. (2006) also suggest that economically disadvantaged groups as well as immigrants and people of color may be mistrustful or not clear about research processes. They may not trust that responses will be confidential. If respondents have not come from western countries they may have no idea of methods of research, a fact that evaluators don't always take into consideration. In the Salishan study, evaluators found that even while using interpreters/research assistants from similar language, cultural, and immigration backgrounds, participants were fearful about sharing certain types of information.

Photovoice and Community Evaluations: The Gaze behind the Camera

Common assumptions about the lives of U.S. public housing resident affect public policy and health interventions aimed at serving them. Beliefs about responsibility and choice, and the role of immigration, race, ethnicity, class, and gender all shape our understanding of the lives of public housing residents (Stevens, 2010).

In community evaluations, several different strategies are employed to elicit the views of the residents. However, the questions and intent of the evaluation are shaped by the researcher and even the funding source of the researcher. While the data gathered in these research strategies are useful, it can ignore the knowledge and perspective of the people who live in the community. Photovoice is a method that can contribute to community evaluations as the participants shape the conversation about their lives and concerns (Newell, Berkowitz, Deacon & Foster-Fishman, 2006; Stevens, 2010; Stevens, 2006; Wang, Yi, Tao, & Carovano, 1998; Wang & Burris, 1997; Wang, Morrel-Samuels, Hutchison, Bell & Pestronk, 2004). Asking participants to document their lives using cameras may be one way to address unequal power in research relationships. The photographs and the discussion that they inspire attempt to produce a mutual understanding of context. Through discussion of the photographs, the participants shape what they wish to share and it can show spaces where they actively resist and use the circulating discourses about their community (Freire, 1970).

Photovoice was used in the Salishan evaluation, as one method to encourage residents to share their experience of relocation and their visions for the new housing community. Participation in photovoice was affected by different immigration trajectories and prior

refugee experiences for the three populations that included Vietnamese, Cambodian, and Russian. For many ethnic groups in communities, their strength lies in their unity and support of one another. While photovoice was to encourage conversation, the Vietnamese community saw the photographs as a possibility of offending people in their community.

The Cambodian population had a different story of immigration that included long years of persecution and refugee camps before arriving in the United States. The Cambodian community contained both refugees and members of the Khmer Rouge who had a history of grievous violence. Photographs taken by this population was considered too risky in light of their history of persecution by Khmer Rouge and therefore felt it could threaten their housing and families if they could be identified by their photographs (Marshall, Schell, Elliott, Berthold, & Chun, 2005; Stevens, 2010).

Focus groups also were utilized in the Salishan evaluation, as a way of including un- and under-represented populations. Groups were formed by language, age, housing tenure, and housing type, and included both Cambodian and Vietnamese elders, English-speaking long-term residents, Russian-speaking homeowners, Cambodian young adults, Russian, Vietnamese and Cambodian adolescents. Though participants had the opportunity to shape the conversation and respond to each others' comments, challenges emerged with focus groups, as well. With both groups of elders, the translator shaped a great deal of the conversation.

Though photovoice is a successful technique for involving communities in evaluation, we must realize "traditional research methods have inadvertent consequences for participants" (Chin, Mio & Iwamasa, 2006). Methodological questions remain around the best approaches to give voice to people marginalized by race, ethnicity, language, and poverty, as well as the ways in which diversity mediates the use of methodologies such as photovoice and focus groups.

Poor Health of Public Housing Residents as the Embodiment of Inequality

We were interested in resident perceptions of their health as one indicator of the well-being of the Salishan community. Self-reported health has been shown to correlate well with actual health. A core goal of HOPE VI is to improve resident self-sufficiency, i.e., the ability of residents to achieve financial independence. This presupposes a population that has the mental and physical capacity to work. If a substantial part of that population is unable to be self sufficient, the ethics of a program as disruptive to the lives of residents as HOPE VI is must be scrutinized. Indeed, high rates of chronic illness, disability, and self-rated poor to fair health were reported at 3 data points in the Salishan evaluation through a survey focused on health prior to relocation (Brennan, Tashiro, & Brusco, 2005), interviews conducted during the first phase of relocation in 2007 (Bodonyi, et al., 2007), and a second wave of interviews conducted in 2009 of the sample of households interviewed in the first phase of the evaluation. This pattern of poor health is similar to findings from other HOPE VI projects (Harris & Kaye, 2004; Howell, Harris & Popkin, 2005; Manjarrez, Popkin, & Guernsey, 2007). This similarity is striking given that the population of Salishan, as in other Northwest HOPE VI sites, differs substantially from the predominantly African American populations of HOPE VI projects in other locations experiencing poor health. At the time of the initial Salishan evaluation, Southeast Asian refugees were half of the total resident population, with Russian-speaking refugees, African Americans, and native-born whites making up the other half. Many respondents with chronic conditions reported limitations affecting their mobility, ability to hold a job, or do household chores. A sobering number reported exposure to trauma, particularly violence, and many suffer from mental health problems, calling into question the HOPE VI program's goal of resident self-sufficiency.

Manjarez, et. al. (2007) rightly point out the need for much more attention to health issues of HOPE VI residents. Their recommendations include adopting improved resident health management as an expansion of the criteria for self-sufficiency, and better coordination between public housing and public health agencies. They also begin a much-needed conversation about venturing beyond traditional health services to address larger environmental concerns that affect health, such as violence. Addressing these recommendations could significantly improve the health and well-being of HOPE VI residents. As unrealistic as the achievement of their goals might seem in today's U.S. economic and political environment, their recommendations are narrow in breadth when compared with the kinds of broad "intersectoral" support recommended by the World Health Organization (WHO) and many population health experts (Kickbush, 2003). An intersectoral approach acknowledges that health and living conditions are inextricable, requiring structural intervention in sectors like transportation, housing, workplace safety, education and employment, which are not usually included in discussions of health policy in the U.S. Conditions under which we live, work, and play influence health in complex, interactive ways, and there is increasing evidence of cumulative effects of multiple risk exposures over the life course for the poor (Evans, Wetherington, Coleman, Worms & Frongillo, 2008). At a population level, the poorer you are, the sicker you are. The inverse hazard law, by which "the accumulation of health hazards tends to vary inversely with the power and resources of the populations affected" is clearly at play for the residents of Salishan (Krieger, et al., 2008).

The majority of the households of people we interviewed had some access to health care. Clearly, access to health care does not assure good health (Booth, 2010). There is also increasing evidence that inequality itself at the national level is associated with adverse health effects at all levels (Wilkinson & Pickett, 2007). Inequality has accelerated in the last two decades. The poor health of public housing residents provides a point of entry for a broader critique of U.S. inequality. Many public housing residents have suffered from exploitation and trauma, and their poor health is the visible manifestation of the destructive effects of poverty and inequality. What roles may evaluators play in disseminating information about inequities that occur outside the scope of the evaluation? Given the well-established relationship between poverty and poorer health, what approaches to the poor health of public housing residents can address underlying structural inequalities?

Flaws in Housing Policy Based Upon Ideological Assumptions about the Poor

The ideological basis of HOPE VI to address urban poverty was a mixed income approach that relied on the decades-old culture of poverty theory of Lewis (1959). This theory suggests that the poor culturally transmit values that go against mainstream societal values. To change this culture, one must change behaviors of the individuals and families who live in poverty neighborhoods (Laakso, in press). More focus is placed on the study of problems than ways to avoid problems. The Salishan researchers learned from qualitative interviews that residents had many strengths that had not been recognized including collective efficacy and a sense of place. Further, a significant number of residents did not fit into a category that would allow for self-sufficiency, a key goal of HOPE VI. At the same time, as noted by Laakso (in press), in spite of age, disability, language barriers, a number of these residents had achieved a degree of self sufficiency and/or had inculcated this value into the next generation. Contrary to a culture of poverty framework, the data demonstrated that these families had the same values as other families.

From an ethical standpoint, as evaluators, the challenge became how to address/challenge the flaws in the assumptions upon which the policies of HOPE VI have been based. Is it the role of the evaluator to address these flaws in order to productively help

those who have been directly affected**?** How can evaluators effectively help future planners and policy makers recognize the strengths and value the voices of those directly affected by flawed policy assumptions? One example of how the residents of Salishan were not heard was in the restrictive rules that were established in the new Salishan. These rules reflected what was supposedly a middle class lifestyle, and included a prohibition on gardening, hanging clothes out to dry, and parking cars on the street. How can we, as evaluators, effectively speak to these issues and others that will impinge on the quality of life for residents if they are not addressed?

Conclusion

Evaluation research poses its own sets of challenges, some of which have been highlighted here, including the inevitable political pressures, ethical issues, IRB procedures designed around medical and natural science research, and developing methodologies that give voice to marginalized populations. While much program evaluation seems to grow out of funding requirements, it is important that evaluators and program administrators look to the more important issue of providing effective, quality services for their constituents. Additionally, it is the contention of the authors that evaluators must engage in thoughtful reflection of the larger context of their research, including underlying assumptions behind social and programmatic policies, and the obvious inequities faced by the populations being studied. The responsibilities extend beyond the simple presentation of findings and need to include the well-being of those persons receiving services.

References

Booth, C.M. (2010). The impact of socioeconomic status on stage of cancer at diagnosis and survival. *Cancer, 116*(17), 4160-4167.

Brennan, K.D., Tashiro, C.J. & Brusco, E.E. (2005). *Salishan household health survey*. Seattle, WA: Northwest Institute for Children & Families, University of Washington School of Social Work.

Bodonyi, J., Brennan, K., Laakso, J., Tashiro, C.J., Stevens, C., Brusco, E. & Keller, J. (2007). *Salishan HOPE VI Redevelopment Midpoint Evaluation. Part 1: Resident Perspectives and Outcomes.*Northwest Institute for Children and Families, University of Washington School of Social Work.

Chin J, Mio J, Iwamasa G (2006). Ethical conduct of research with Asian and Pacific Islander American populations. In: Fisher JTC, Ed. *The Handbook of Ethical Research with Ethnocultural Populations and Communities*. Thousand Oaks: Sage, 117-137.

Evans, G.W., Wethington, E., Coleman, M., Worms, M. & Frongillo, E.A. (2010). Income health inequalities among older persons: The mediating role of multiple risk exposures. *Journal of Aging and Health, 20*(1), 107-125.

Freire P. *Pedagogy of the Oppressed*. (1970) New York: Continuum

Greene, J.C. (1997). Evaluation as advocacy. *Evaluation Practice, 18*(1), 25-35.

Harris, L.E. & Kaye, D.R. (2004 October). How Are HOPE VI families faring? Health. *Metropolitan Housing and Communities Center*, Washington, DC: Urban Institute.

Holosko, M., Thyer, B., Danner, J. E. H. (2009). Ethical guidelines for designing an conducting evaluations of social work practice. *Journal of Evidence-Based Social Work.* 6: 348-360. doi:10.1080/15433710903126778.

Howell, E.M., Harris, L.E. & Popkin, S.J. (2005). The health status of HOPE VI public housing residents. *Journal of Health Care for the Poor and Underserved, 16*(2), 273-285.

Joseph, M. L., Chaskin, R. J., & Webber, H. S. (2007). The theoretical basis for addressing poverty through mixed-income development. *Urban Affairs Review,* 42, 369-409. doi:10.1177/107808740629443

Keller, J. (2011). Experiences of public housing residents following relocation: Explorations of ambiguous loss, resiliency and cross-generational perspectives. *Journal of Poverty,* 15(2), 1-23. doi: 10.1080/10875549.2011.563170

Kickbush, I. (2003). The contribution of the World Health Organization to a new public health and health promotion. *American Journal of Public Health, 93*(3), 383-388.

Krieger, N., Chen, J.T., Waterman, P.D., Hartman, C., Stoddard, A.M., Quinn, M.M., Sorensen,G., & Barbeau, E.M. (2008). The inverse hazard law; Blood pressure, sexual harassment, racial discrimination, workplace abuse and occupational exposures in US low-income black, white and Latino workers. *Social Science & Medicine, 67*, 1970-1981.

Laakso, J. (in press). Flawed assumptions and HOPE VI. *Journal of Poverty.*

Lewis, O. (1959). Five families: Mexican case studies in the culture of poverty. New York: Basic Books

Manjarrez, C.A., Popkin, S.J., & Guernsey, E. (June, 2007). Poor health: Adding insult to injury for HOPE VI families. Metropolitan Housing and Communities Center, Urban Institute.

Marshall, G. Schell, T., Elliott, M. Berthold, S.& Chun, C. (2005). Mental Health of Cambodian refugees two decades after resettlement in the United States. *Journal of American Medical Association. 294*(5), 571-579.

Newell B, Berkowitz S, Deacon Z, Foster-Fishman P. (2006) Revealing the cues within community places: stories of identity, history, and possibility. *American Journal of Community Psychology*. *37*(1-2):29-41.

Royse, D., Thyer, D., Padgett, D., Logan, T. (2006). *Program evaluation: An introduction.* (4th ed.). Belmont, CA: Thompson Brooks/Cole.

Schwandt, T. (2007). Expanding the conversation on evaluation ethics. *Evaluation and Program Planning, 30*(4), 400-403.

Swauger, M. (2009). No kids allowed!!!:How IRB ethics undermine qualitative researchers from achieving socially responsible ethical standards. *Race, Gender & Class.* 16(1-2): 63-81.

Stevens, C. (*2010).* Lessons from the field: Using photovoice with an ethnically diverse population in a HOPE VI Evaluation*. Family and Community Health*. 33(4): 275-284.

Stevens, C. (2006). Being healthy: Voices of adolescent women who are parenting. *Journal for Specialists in Pediatric Nursing.* 11(1), 28-40.

Wang C, Burris M.(1997) Photovoice: Concept, methodology, and use for participatory needs assessment. *Health Education & Behavior.* 24(3):369-387.

Wang C, Morrel-Samuels S, Hutchison P, Bell L, Pestronk R.(2004) Flint photovoice: Community building among youths, adults, and policymakers. *American Journal of Public Health.* 94(6):911-913.

Wang C, Yi W, Tao Z, Carovano K.(1998) Photovoice as a participatory health promotion strategy. *Health Promotion International.*13(1):75-85.

Wexler, H. J. (2001). HOPE VI: Market means/public ends--The goals, strategies, and midterm lessons of HUD's urban revitalization demonstration program. *Journal of Affordable Housing,* 10(3), 195-233.

Wilkinson, R. & Pickett, K. (2007). The problems of relative deprivation: Why some societies do better than others. *Social Science & Medicine, 65*(9), 1965-1978.

Reflexivity of Evaluation Research

Esa Jokinen, Work Research Centre, University of Tampere

Abstract

Processes and competences relating to the policy evaluations have not been much studied previously, although reviews of evaluation studies have appeared increasingly lately. While the emergence of evaluation studies in the 1990s has taken place at the same time as the discursive change in social sciences, the mainstream of evaluations is quite non-reflexive and method-driven in many ways. The possible dialogic and self-reflexive aspect of evaluation research is discussed by employing a textual example from the evaluation of local government reform (Paras). It is shown that the ability to address evaluation methodological dilemmas in reports and presentations shortly is a vital part of any complex evaluation research as well as the ability to position evaluation researcher's role. It is suggested that it would be useful to consider these abilities as a central part of the organizational evaluation capacity and researcher competence.

Keywords: evaluation capacity building, evaluation competence, evaluation process, local government reform, reflexivity

1. Introduction

This article discusses evaluation as a research practice contrary to an institutionalized, academic or formally driven activity (about institutionalized evaluation see e.g. Varone, Jacob & De Winter 2005). The roots of my thinking lie mostly in self-evaluation models and empowering and developmental evaluation debate (e.g. Patton, 1997; Seppänen-Järvelä & Vataja, 2009).

By employing a case as an example, I aim to shed light on a very practical aspect of evaluation, namely the communication of evaluation research to the wider audience. Leaning on the literature and on some empirical observations, I argue that the objectivity and truthfulness of evaluation, which are pronounced in this type of research, cannot be achieved by doing evaluation "by the book" but by using creatively different discursive tools or frames in communicating the evaluation research to the public.

Guided by legislation (The Act on Restructuring Local Government and Services 169/2007), restructuring of local government and services (the Paras reform) has been a historical development initiative in Finland. It has been followed up by an evaluation research project "ARTTU" in 40 (in 65 until 2009) municipalities since 2007, which represents an exceptionally multi-method-, multi-disciplinary- and multi-researcher-based evaluation approach.

One of the main rationales underlying the Paras reform is the assumption of the economies of scale, meaning that in order for a municipality to be able to maintain the services needed it has to have a large enough population living in the governed area. Population growth is achieved either through municipal mergers, or through cooperation contracts between adjacent municipalities' social and health service organizations. The size of the population in one common area should be at least 20,000, which means that most of the Finnish municipalities are below that size. (Meklin & Pekola-Sjöblom, 2010, p. 2.)

In Finland, many mergers and partnership areas have been established during the 2000s. By the year 2010, 66 partnership areas had been planned, of which 48 were already functioning, involving 172 municipalities and one third of Finnish inhabitants (Heinämäki, 2011, p. 8). However, the structural reforms have not always been made because of the Paras legislation, and local governments have been allowed to act according to their own schedules and consideration in these matters.

The aim of this article is to describe and analyse the dynamics of the challenging research field of evaluation. The ARTTU evaluation programme has published over 10 research reports already, so those interested in the substance of evaluation can find it on the website of the programme[1] (see also Appendix). The research programme includes eight different research modules ranging from the municipal economy to the linguistic (equality) implications of the reform. It also strongly aims to make practical use of the results of ARTTU along the way. The research programme has been coordinated by the Finnish Local and Regional Authorities (ALFRA) and executed by different universities.

2. Local government reform evaluations reviewed

According to the categorization proposed by Roininen and Valovirta (2009, p. 38), this article belongs to the category of evaluation know-how research, which is said to be the least prosperous field of evaluation studies. Methodologically this study is based on the social sciences and substantially on the administrative sciences and it concentrates on the reflexive aspects of an evaluation research programme in question.

Even though plenty of evaluation studies on the public sector have been conducted during the 2000s, there are only a few previous examples of this type of research design which concentrate on evaluation processes.

Although there are new reforms constantly under way, the local government reform (Paras) in particular has been a media issue lately. Nyholm and Airaksinen (2011) have gone through several evaluation studies concerned with the local government reform, of which some belong to ARTTU research programme. They define evaluation in relation with the complexity of reform process. They say that, on the one hand, evaluation should bring up different kinds of perspectives and interpretations of the ongoing reform but, on the other hand, evaluation also needs to adapt to the complexity and the contextual nature of reality. Still, Nyholm and Airaksinen call for a solid and wide framework for new evaluations (ibid., p. 309).

Niiranen (2011) examines the ways in which municipal decision-makers search and use information. Her article deals with the social aspects of evaluation processes in municipal administration but not so much with the underlying principles or features of evaluation. She summarizes that although the future decisions rely more and more on the empirical data and evaluations, any synthesizing evaluation may be difficult to achieve and the use of such information would be most likely limited (ibid., p. 313, p. 320).

One additional example of an article dealing with evaluation practices in Finland is the writing by Pesonen (2009) who has taken a more distanced perspective to the evaluation practices relating to the Finnish Funding Agency for Technology and Innovation (Tekes).

1

[1] http://www.kunnat.net/fi/palvelualueet/arttu/Sivut/default.aspx (in Finnish).

Therefore, at their best, the different evaluation frameworks seem to arise from a researchers' thorough understanding of real life organizations and personal experiences of the change processes of researchers. Hence, depending on the field of the evaluator's interest, the evaluation may be characterized as more economical, psychological, social psychological, sociological or administrative in nature. The legitimizing function of evaluations is also being brought up in the articles from time to time, and it has been seen as an important aspect to bear in mind while assessing the use of evaluations (e.g. Nyholm & Airaksinen, 2011, p. 299, p. 308).

3. Evaluation dilemmas

Evaluation is not a rigid discipline but rather a field of diverse professional practices (Greene, 2001, p. 181) done by multiple actors and organizations drawing from many different methodological schools of thought. However, regardless of the background theories and interests, most evaluation studies have some universal phases and features like setting evaluation questions and indicators, collecting data, analysing and reporting of results.

In its history, evaluation research has gone through many stages of "pursuing the impossible", namely tackling with the "does the intervention work" type of questions. Yet little is known and understood about the wider functions of evaluation in society and the delicate interaction between the evaluation researchers, the objects of the evaluations and the rest of the community. (Cf. Giel, 2012.)

Recently, the policy evaluation studies themselves have become more and more reviewed from different perspectives which override the inherent policy related logic and questions (e.g. Lloyd & Harrington, 2012). It really has not been common to take the evaluations "out of the box", despite the seeming efforts to stay objective and rational while doing the assessment. However, even a policy evaluation is not made only for decision makers but also for the wider audience. Evaluation may reveal challenges and many controversial issues and dilemmas relating to the governing efforts of nations.

For instance, the case of long-term "Sure Start", a multi agency policy initiative in Great Britain aiming at supporting children and their families, appeared to be difficult to evaluate as a whole. Despite the fact that according to "on-the-ground" experience of workers and families, the initiative had had a great impact, it was difficult to demonstrate these impacts on the national level. Many of the reasons for this were very practical: there was not enough first-hand data and resources for conducting the evaluation, not to mention the integration and coordination of local and national evaluation levels. It was difficult to define any "outcomes" because the evidence gathered was often of too poor quality. Evaluation was not the top priority for practical actors and it was difficult to engage key partners. Overall the evaluation know-how was far too inadequate compared to the programme complexity. (Lloyd & Harrington, 2012.)

A new understanding of the possibilities of policy interventions and development programmes is emerging, while evaluation is being seen as part of the social phenomena and governing practices rather than a separate set of research methodologies.

4. Discursive change

Social psychologists have written about discursive change since the late 1980s. The attention of social scientists turned (for a while) to the texts and to the everyday use of linguistic repertoires. (E.g. Potter & Wetherell, 1995.) The world changed as well, and the massive emerging of policy and other evaluations also took place during this time.

Still, the constructionist understanding of social phenomena – including research or evaluation itself – has been giving way to more *realistic* evaluation methodologies (e.g. Pawson & Tilley, 1997). It has appeared that, in the world of continuous change and disappearance of solid concepts and belief structures, one of the most important "hidden curricula" of evaluation has been to construct the evaluation object and give it a shape in order to serve the respective interest groups.

What has happened in the policy context lately is a kind of overcharge of evaluation; instead of taking any intervention or policy "as it is", there has risen a constant need to fit it into some narrow – most often economically and efficiency-driven – framework, which is then operationalized through evaluation.

An example of when the given evaluation framework becomes explicit (because the phenomenon does not fit into it) comes from a Finnish labour policy researcher Simo Aho (2008, p. 45):

> "Under these conditions, evaluation inevitably produces results that show a low average employment impact (of ALMP[2]) even when the measures as such are relevant for the improvement of employability (if adequately targeted). (…) ALMP's have several additional, more or less informal and sometimes controversial, goals, such as to prevent social exclusion, provide incentives to work and legitimate the prevailing "employment regime"…"

From the perspective adopted in this article, the question is not whether the ALMP is really considered effective in terms of improved employability but how the evaluation dilemma in question is communicated in short. An important part of the evaluator competence is being able to move out from the pre-set (economical) frame in order to tell meaningfully – and truthfully – about the relationship between the evaluation methods and the real phenomenon to the wider audience.

5. The ARTTU evaluation as a concise text

The three examples illustrating "evaluative reflexivity" were identified in the common introduction chapter used in all ARTTU research modules' intermediate evaluation reports (published in 2010–2011) written by the programme coordinators (Meklin & Pekola-Sjöblom, 2010). Three of the most evident evaluation dilemmas are brought forth through them. The excerpts also represent the need to clarify the nature of evaluation research on such a (supposedly) dramatic reform in local government and service systems for the public.

First, there arises a need to steer the evaluation process adjacent to reform process itself. It has been said that

> "(t)he means of implementing the reform prescribed by the Act — mergers, establishment of partnership areas and cooperation between urban regions **— do not, as such, produce the expected advantages**. They are the tools for organising functions, or visible structures, and offer development potential only. (…) The particular focus of the research programme is to find out ***how local authorities have harnessed this potential*** by concentrating services or by making changes to ways that services are provided, and to observe the impacts of these measures from different perspectives." (ibid., bold added.)

---

2 Active labour market policies.

This positioning not only gives the evaluation programme the possibility to distinguish "real changes" from "mere constructions" or "illusions" in a methodological sense, but also highlights municipal actors' role and responsibility in interpreting the reform. In this paragraph, the role of evaluation switches from assumed "defining and assessing the reform" to "being a companion in pursuing the reform objectives".

This interpretation of the role of evaluation research is confirmed also by a second observation of maintaining the pre-set categories. It has been said in the introduction (ibid.) that

> "the participating local authorities were placed in PARAS categories[3] in autumn 2007. Local authorities may have later made **decisions which would place them in some other category than where they were originally placed**. However, this is in **keeping with the spirit of the Paras reform** and we can expect that, during the research period, several changes will take place in local authorities that will make it necessary to review their categorisation." (bold added.)

This paragraph refers again to the methodological problem that the evaluation categories are not factual but only changeable tools in analysis, implying, however, at the same time, that actually evaluation does not attempt to define the reform elements but only to point out the options for municipalities defined in the Paras Act. Simultaneously the role of evaluation as a companion (of municipalities) is further constructed.

A third observation is that evaluation will move its focus from reform effects to examining reform or reform actions themselves.

> "As the programme progressed, researchers noticed that **many other aspects of daily operations in municipalities either support or prevent the realisation of the PARAS project objectives**. (…) *While the Evaluation Research Programme ARTTU aims to determine the impacts of the PARAS reform to the extent possible, the primary goal of the programme is to find answers to the following questions, which are more general in nature:*
> - What decisions have local authorities made and what measures have they taken during the PARAS project, or what has occurred in the municipalities?
> - What impacts do these measures and events have from the perspective of the research modules?
> - What factors lie behind the differences in how the PARAS reform has progressed in different municipalities?
> - Has central government steering helped implement the PARAS reform in municipalities and how?" (Meklin & Pekola-Sjöblom, 2010; bold added.)

There are two general methodological dilemmas to be solved by any evaluation study: 1) what are the effects of intervention or changes in the system and 2) whether or not the effects are due to that change (Lindqvist 2006, p. 15). This dilemma is being addressed above in the ARTTU context. The role of evaluation as defining the good or right results of the reform is being transformed to a role of a "helper" so that it produces information according to which different actors (mainly municipalities themselves) can draw their own conclusions.

3

[1] Merged local authorities, Local authorities pursuing deepened cooperation, Other local authorities and Urban regions, of which the last one was combined with the "Other" category in 2011.

6. Discussion

The reflexive or dialogical approach is now and then used as a method of gathering other (narrative) than formal (argumentative) data during evaluation study (e.g. Abma, 2001). This approach relates to the empowering and developmental evaluation debate (e.g. Patton, 1997) and also to the action research debate (Arnkil, 2004, p. 76). Abma (2001) states that evaluation has been lagging behind other disciplines in applying dialogical methods, and according to Greene (2001, p. 182, p. 186)

> "dialogue in evaluation contexts refers to engaged, inclusive and respectful interactions among evaluation stakeholders about their respective stances and values, perspectives and experiences, dreams and hopes, and interpretations of gathered data related to the evaluand and its context. The purpose of such evaluative dialogue is to enable stakeholders to more deeply understand and respect, though not necessarily agree with, one another's perspectives. (…) Dialogic evaluation seeks to be of the world, not just to report on it."

What I suggest, however, is that dialogue as described above is just an expression of a wider concept of reflexivity or self-reflexivity of evaluation. With reflexivity I mean appreciation of other perspectives and the ability to review one's one evaluation methods and conceptions of the evaluand. This I consider to be an essential element in constructing plausible evaluation, regardless of the school of thought.

While Abma (1998), for example, is proposing reflexivity and polyvocality as a form of reporting on complex evaluation themes, I suggest that reflexivity can also be utilized more subtly or prosaically in addition to stakeholder dialogues or poetic presentations. In this way, reflexivity can also be used to mask certain aspects of using power while doing evaluation.

The proposed discursive perspective on the evaluation capacity building and know-how comes in a way close to (pragmatic) realistic evaluation (Mark, Henry & Julnes 2000). Both consider it important to distinguish between the world as it is and our conceptions of it. However, while realistic evaluation doctrine seems to suggest that we should *research* the layers and the interrelations of (social) contexts more or less thoroughly, I suggest that evaluation research should *communicate* these relationships in a way that fits the evaluative context. In addition, as I have attempted to show, this is exactly what has been done in the ARTTU local government reform evaluation research and in some other evaluation studies as well. There is only too little awareness of this aspect, and it is a hardly acknowledged part of evaluation competence.

One of the main challenges relating to this kind of publicly very noteworthy evaluation research is how to keep distance from the pre-set debate dimensions. I argue that especially this kind of significant evaluation serves best when the researchers are aware of the constructions offered to them in each part of the evaluation task and work consciously on developing their own constructions.

What comes to the emphasized role of ARTTU evaluation research as a "helper" of municipalities, it is understandable in a sense that the programme is coordinated by ALFRA. The general point I want to make is, however, that the skill of constructing the evaluation design and the position of evaluation in the current context "makes" the evaluation, and is of course only a "top of an iceberg" of an ambitious evaluation. In the three excerpts described, the evaluation methodological dilemma, on the one hand, and the contextual positioning of evaluation itself are skilfully intertwined. Introductory text represents a very concise and "powerful" text because it is repeatedly expressed in each research module report, and as such it also represents "higher level" communication of the programme. From my point of view, these examples represent "evaluative reflexivity".

Finally it could be said that the ARTTU research programme is an expression of a vision that evaluation is not a singular activity where knowledge only accumulates in some restricted field of expertise. In such an ambitious evaluation and cooperation design it has been understood that there are many methods of doing evaluation, but there also need to be some paths "out of the evaluation". This has forced us to form plausible interpretations of the results and pave way to practical use of the information like creation of linkages between information "producers" and "users". Evaluation of such a complex reform taking place in local government in Finland is not and cannot be evaluated comprehensively in any strict sense, but evaluation is rather more like a network of operations and communication acts between different stakeholders. Manifold evaluation know-how is certainly needed here.

References

Abma, T. A. (1998). Text in an Evaluative Context. Writing for Dialogue. *Evaluation* 4 (4), 434–54.

Aho, S. (2008). Miksi työvoimapoliittisten toimenpiteiden mitattu vaikuttuvuus on keskimäärin alhainen? Työllistyvyyden parantamispyrkimysten arvioinnin keskeisten ongelmien tarkastelua. *Hallinnon tutkimus* 27, 45–60.

Arnkil, R. (2004). Keeping up with the times? A comment on comments on action research. *Concepts and Transformation* 9 (1), 75–84.

Greene, J. C. (2001). Dialogue in evaluation: A relational perspective. *Evaluation* 7 (2), 181–187.

Heinämäki, L. (2011). *Yhteistoiminta-alueiden sosiaali- ja terveyspalvelut 2010. Järjestämisen, tuottamisen ja hallinnon kysymyksiä uusissa palvelurakenteissa.* Raportti 41. Terveyden ja hyvinvoinnin laitos. Helsinki.

Julkunen, I., Lindqvist, T. & Kainulainen, S. (Eds.) (2005). *Realistisen arvioinnin ensimmäiset askeleet.* FinSoc Työpapereita 3/2005. Helsinki: Stakes.

Lindqvist, T. (2006). Johdatus tapauskohtaiseen ja realistiseen arviointiin. In Julkunen, I., Lindqvist, T. & Kainulainen, S. (Eds.) *Realistisen arvioinnin ensi askeleet.* (pp. 13–16). Stakes. FinSoc Työpapereita 3/2005.

Lloyd, N. & Harrington, L. (2012). The challenges to effective outcome evaluation of a national, multi-agency initiative: The experience of Sure Start. *Evaluation* 18 (1), *93–109.*

Mark, M., Henry, G. & Julnes, G. (2000). *Evaluation. An integrated framework for understanding, guiding and improving policies and programs.* San Fransisco: Jossey-Bass.

Meklin, P. & Pekola-Sjöblom, M. (2010). *Introduction to the research program ARTTU.* http://www.localfinland.fi/en/association/research/arttu/Documents/arttu.pdf

Niiranen, V. (2011). Arviointitieto ja sen käyttöala kuntien päätöksenteossa. *Hallinnon tutkimus* 4/2011, 313–324.

Nyholm, I. & Airaksinen, J. (2011). Kuntahallinnon uudistaminen arvioinnin kohteena. *Hallinnon tutkimus* 30 (4), 297–312.

Patton, M. Q. (1997). *Utilization-Focused Evaluation: The New Century Text*. Thousand Oaks: Sage.

Pawson, R. & Tilley, N. (1997). *Realistic Evaluation.* London: Sage.

Pesonen, P. (2009). Soveltavaa arviointitutkimusta ja kehityskumppanuutta. Tekesin ohjelma-arviointien kehitysvaiheet ja asemointi tilaajan näkökulmasta. *Hallinnon tutkimus* 28 (5), 41–59.

Potter, J. & Wetherell, M. (1994). *Discourse and social psychology. Beyond attitudes and behaviour.* London: Sage.

Roininen, J. & Valovirta, V. (2009). Katse arviointiosaamiseen. *Hallinnon tutkimus* 28 (5), 37–40.
Seppänen-Järvelä, R. & Vataja, K. (2009). Kehittävän itsearvioinnin taidot, kyvykkyydet ja osaaminen: empiirinen analyysi sosiaalitoimen työyhteisöistä. *Hallinnon tutkimus* 28 (5), 60–73.
Ton, G. (2012). The mixing of methods: A three-step process for improving rigour in impact evaluations. *Evaluation* 18 (1), 5–25.
Varone, F., Jacob, S. & De Winter, L. (2005). Polity, Politics and Policy Evaluation in Belgium. *Evaluation* 11 (3), 253–273.

Appendix

The ARTTU evaluation programme

The research programme aims at gathering information about reform processes, changes and effects of the changes in the reform period 2009–2012. The task of implementing the overall research has been divided into 6 different modules and research teams: 1) local democracy and leadership (Åbo Akademi University and ALFRA), 2) municipal services (University of Kuopio), 3) municipal personnel (University of Tampere), 4) municipal and regional economy (University of Tampere), 5) community structure in the urban regions (Aalto University), and 6) evaluation of the execution of the reform (University of Lapland and University of Tampere). In addition, there are two complementary studies, namely 7) the linguistic implications of the Paras reform (Aalto University and Åbo Akademi University) and 8) the gender impact of the Paras reform (Åbo Akademi University).

40 municipalities were chosen to represent Finnish local government on the basis of previous studies, the multiple constituency model including literature, statistics and qualitative data. The central criteria are that the municipalities represent one of three Paras categories, of which two first ones refer to radical structural reform and the third to non-structural reform. The categories are 1) municipal mergers, 2) partnership and cooperation areas and 3) efficiency improvers (other). (Meklin & Pekola-Sjöblom, 2010.)

For more information and publications, go to the programme website (in Finnish): http://www.kunnat.net/fi/palvelualueet/arttu/Sivut/default.aspx.

For more information on the ARTTU evaluation programme in English, go to: http://www.localfinland.fi/en/association/research/arttu/Pages/default.aspx

The ARTTU synthesis reports:

Vakkuri, J., Kallio, O., Tammi, J., Meklin, P. & Helin, H. (2010). Matkalla kohti suuruuden ekonomiaaa? Paras-ARTTU-ohjelman tutkimuksia nro 3. Acta nro 218 http://shop.kunnat.net/product_details.php?p=1711

Meklin, P. (ed.) (2008). Parasta Artun mitalla? Arviointia Paras-uudistuksen lähtötilanteesta ja kehittämispotentiaalista kunnissa. Paras-ARTTU-ohjelman tutkimuksia nro 5. http://shop.kunnat.net/product_details.php?p=360

The latest ARTTU research module reports:

Sandberg, S. (2012). Paras-uudistus kuntapäättäjän silmin. Paras-ARTTU-ohjelman tutkimuksia nro 20. Acta nro 235. http://shop.kunnat.net/product_details.php?p=2680

Jokinen, E. & Heiskanen, T. (2012). Henkilöstö uudistusten pyörteissä. Paras-ARTTU-ohjelman tutkimuksia nro 19. http://shop.kunnat.net/uploads/arttu_henkilostoraportti.pdf

Pekola-Sjöblom, M. (2011). Kuntalaiset uudistuvissa kunnissa. Paras-ARTTU-ohjelman tutkimuksia nro 9. Acta nro 229. http://shop.kunnat.net/product_details.php?p=2590

Mäntysalo, R., Peltonen, L., Kanninen, V., Niemi, P., Hytönen, J. & Simanainen, M. (2011). Keskuskaupungin ja kehyskunnan jännitteiset kytkennät. Paras-ARTTU-ohjelman tutkimuksia nro 2. Acta nro 217. http://shop.kunnat.net/product_details.php?p=1704

Mehtäläinen, J., Jokinen, H. & Välijärvi, J. (2011). Koulutuspalvelut ARTTU-kunnissa. Koulutuksen saatavuus ja saavutettavuus. Paras-ARTTU-ohjelman tutkimuksia nro 18. http://shop.kunnat.net/uploads/arttu_koulutusraportti.pdf.

III

International Comparative Approaches in Evaluation

How do social institutions influence E-Accessibility polices in the UK, US, and Norway?

G. Anthony Giannoumis1 and Rune Halvorsen
NOVA – Norwegian Social Research Institute

Abstract

The United Nations Convention on the Rights of Persons with Disabilities recognizes accessibility to information and communication technologies, E-Accessibility, as essential for full participation in the information society. Based on original research, currently in progress, this paper presents preliminary findings of a document analysis of policies from the United States, United Kingdom, and Norway. The paper examines how national and supranational policies for promoting E-Accessibility balance economic and social needs. Due to its social and political importance, the paper focuses on web accessibility. The results demonstrate how the approaches to ensuring and enhancing economic opportunities for private enterprises and social opportunities for persons with disabilities have been influenced by different national policy traditions, the distribution of roles, and the relationships between actors participating in the design and implementation of E-Accessibility policy. Future research must continue to empirically examine the mediators to effective policy implementation from within a system of multi-level governance.

Keywords: disability, E-Accessibility, policy implementation, European Union

Introduction

The United Nations (UN) Convention on the Rights of Persons with Disabilities (CRPD) recognizes accessibility to information and communication technologies (ICT) as essential for full participation as members of the information society, to be able to exercise freedom of choice and independent living (UN, 2007, Art. 9). Although a new era of disability rights is emerging promoting accessibility for persons with disabilities – both internationally and in the European Union (EU) – many European countries still experience a digital gap between disabled and non-disabled populations (Technosite, NOVA, & CNIPA, 2011). Despite numerous initiatives by the European Commission (EC) over the last decade directed at improving accessibility to ICT (E-Accessibility), Europe continues to lag behind the United States (US) (Technosite et al., 2011). E-Accessibility refers specifically to the design and utility of ICT for persons with disabilities and concerns the universal design of facilities, products and services to be usable by all people, including persons with disabilities. E-Accessibility therefore designates that persons with disabilities have access, on an equal basis with others, to ICT. A critical question is whether the producers and providers of the technology provide necessary and appropriate adjustments of the ICT to meet the needs of persons with disabilities. In order to promote E-Accessibility outcomes, E-Accessibility policies have attempted to influence the behaviour of the market (both public and private enterprise goods and services providers).

In the EU, perhaps the most critical piece of legislation impacting the regional development of E-Accessibility, the European Accessibility Act (EAA) is expected to emerge in 2012. The EAA has been working its way through the EU since 2008. The intent of the proposed act is to harmonise the European market for accessible goods and services through social regulation (EC, 2011). Here, social regulation refers to policy instruments aimed at ensuring E-Accessibility in the market (Braithwaite & Drahos, 2000; Levi-Faur, 2011; Levi-

Faur & Jordana, 2004; Majone, 1993, 2005). In the consultation document for the EAA, the EC recognized the low levels of accessibility outcomes and compliance with national policies. In February 2012 the EC closed the public consultation for the EAA. Disabled persons organizations (DPOs), academics, persons with and without disabilities, professional organizations, and advocates throughout Europe responded demonstrating the need for EU level regulatory intervention and urging that the implementation of the EAA include effective enforcement and monitoring mechanisms.

Objective

The overall aim of this study is to examine how social institutions, *i.e.* norms, values and procedures important to a society, affect the design and implementation of national and supranational E-Accessibility policies (Powell & DiMaggio, 1991). This paper will present the preliminary findings of this research in progress by examining how national and supranational policies for promoting E-Accessibility balance regional and international economic and social needs (e.g. as reflected in policies for 'reasonable accommodation' and 'undue hardship', the definition of 'accessibility', and the conditions regional authorities comply with E-Accessibility requirements). Due to its social and political importance, the paper focuses on web accessibility (Internet and intranet websites, and web-based applications). Notable are technologies which directly impact the right to education and healthcare. These technologies include web based eLearning platforms (e.g. Learning Management Systems and virtual learning environments); and health information and healthcare services. Additionally, Web 2.0 technologies such as social media, social networking, and other web applications continue to grow in social and economic importance (Luo, Wang, Hu, & Shi, 2009). These potential and frequently encountered barriers to participation and inclusion demonstrate the need for mainstreaming web accessibility in the public and private sectors.

Previous studies have established the policy landscape, legal casework and E-Accessibility outcomes in Europe and the US; however gaps in the literature still exist regarding the mechanisms that mediate effective implementation. (Blanck, 2008; Timmers, 2008; Myhil et al., 2008; Waddington & Quinn, 2010; Aasen, Halvorsen & Silva 2009; Technosite et al., 2011).

This paper contributes to developing the field by systematically comparing three countries with highly contrasting approaches to E-Accessibility, and examining the relation between supranational and national policy developments and implementation in E-Accessibility policy. Preliminary data suggest that the US has given priority to statutory social regulation of E-Accessibility while Norway (until recently) has given priority to the provision of assistive technology (Aasen et al., 2009). The UK's approach may be considered an intermediary case (Lawson, 2008). The choice of cases (the UK, US and Norway) demonstrate contrasting approaches for addressing E-Accessibility within differing legal cultures, regulatory environments and policy instrumentation. While the UK-US comparison resembles what George and Bennett (2005) call the "most similar case design", the US-Norway comparison resembles what the authors have called the "most different case design".

The comparison is framed by implementation research which attempts to assess policy provisions and enforcement by analysing policy tools (legislation, financial incentives and persuasion strategies) as a component of policymaking and enforcement (Hill & Hupe, 2008). These tools vary in the character of their enforcement when applied to private enterprises from directive based, judicially enforceable policies to non-binding, normative, hortatory policies. The Multiple Governance Framework (MGF) redirected scientific inquiry from the inputs and outputs of policy implementation to the multi-layered framework of governance

and the complex social and political networks that shape policy implementation (Hill & Hupe, 2008). This paper builds off of these frameworks by applying policy implementation theory to a new and unique value system, the accessibility of ICT, specifically web-based information and social services.

Methods

Since national E-Accessibility policy is embedded in multiple levels of governance, a comparative case study was used to provide both a supranational macro level, and regional micro level analyses (Dobbin, Simmons & Garrett, 2007; Hill & Hupe, 2008). This paper will identify the variety of meaningful patterns that exist between cases, provide an historical interpretation through policy traditions, and describe the social mechanisms of policy implementation through patterns of constant association and invariance (Ragin, 1987). Although these relationships provide limited external validity they do provide analytical generalizability through the application of results to a broader theory, in this paper the application of defined characteristics of mediating institutions to public policy research and implementation theory (Mitchell, 1983; Yin, 1994).

It would be impossible to comprehensively account for all mediating factors in policy implementation on a national, much less, supranational level. In this paper, particular conditions and cases were deliberately chosen to explain and question established inferential relationships. Cases for this paper provide a locus of investigation for E-Accessibility policy within and outside the EU where the relationship between E-Accessibility outcomes and mediating institutions will be examined through a document analysis. Cases include the UK, US and Norway and were selected to expose unique conditions acting as potential mediators in policy implementation. These include the heterogeneity of regulatory environments (including the use of policy instruments) and policy enforcement traditions (Burke, 2002), and the relationship with supranational governance in the EU (the UK, an EU Member State; Norway, an EEA Member State; and the US, a non-European state) and UN (the US has traditionally been a late adopter of UN human rights conventions including the CRPD). The analysis is based on a combination of a document analysis of primary sources and a literature review. Documents for the literature review were selected in the following topic areas,

- The policy environment including, policy valuation (e.g. political investment, measures of success), and policymaking processes and traditions (e.g. legislative approach, the role of welfare policy, and policy development characteristics)
- Legislative and legal structures (e.g. implementation, monitoring, and enforcement characteristics) including infrastructure, scope, and procedures (e.g. litigation, regulation, incentivization, voluntariness, means of redress, and complaint mechanisms).

Primary sources include key pieces of legislation from the UK (The Disability Discrimination Act (DDA) and The Equality Act including associated binding and nonbinding policies), the United States (Section 508 of the Rehabilitation Act and the Americans with Disabilities Act including select pieces of case law), Norway (the Anti-Discrimination Accessibility Act in unofficial English translation) the UN (Convention on the Rights of Persons with Disabilities), and the EU (various policies addressing E-Accessibility and E-Inclusion spanning 1996 to 2012).

Results

A paradigm shift in disability policy in the EU began in the 1990's as redistributive policy gave way to regulation (Hvinden & Halvorsen, 2003). Here, redistributive policy refers to the transfer of services in cash and in kind aimed at equalizing life chances for persons with disabilities. The Amsterdam Treaty established disability as a civil rights issue and the rights based approach has expanded to include the failure to provide reasonable accommodation as an act of discrimination (Waldschmidt, 2009; Waddington & Quinn, 2010). Therefore establishing a threshold for reasonable accommodation is one of the most critical factors in implementing E-Accessibility policy. Since the 1990's, EU policies have been largely hortatory focusing on social participation, the right to non-discrimination and equal opportunities, leading to few administrative or judicial decisions (Hvinden & Halvorsen, 2003; Waldschmidt, 2009). During this period, the EU became an active promoter of the US approach to social regulation motivated partly by the limitation of EU governance to engage in redistributive policy. In 2010, the EU ratified the CRPD creating an obligation to exercise existing competences over Member States to fulfil the requirements of the CRPD. The objective of the CRPD was not to create new rights; it instead aimed to ensure that all existing rights are made equally effective for persons with disabilities.

To support the harmonization of E-Accessibility policy, the Monitoring eAccessibility in Europe study has provided an evidence base for comparing E-Accessibility policy and status in the EU. Relevant policy engagement and E-Accessibility status indicators were selected from the 2011 Annual Report (web accessibility, mobile web, and educational environment) (Technosite et al., 2011). Table 1 includes an elaboration of these indicators (Policy Environment – Valuation) as part of the cross case comparison of Policy Environment, and Legislative and Legal Structure. The results of the paper demonstrate that the UK, US, and Norway have all responded to the demand for more inclusive design of ICT through social regulation. The cases are uniquely oriented in their approaches to legislation, social protection systems, and policy implementation. In addition, the implementation, monitoring and enforcement of E-Accessibility policies in the UK, US, and Norway are also characteristic of their divergent and convergent public policy systems.

Discussion

The results (Table 1 Policy Environment – Valuation) show that UK policy has engaged with web accessibility through policy implementation and enforcement, and the education environment through eLearning platform accessibility. However UK policy has inadequately addressed web accessibility monitoring, and improved outcomes related to private sector websites, public and private sector mobile websites, and eLearning platforms have not been realized. US policy does not show particularly high levels of engagement with any of the selected indicators, and policies have so far inadequately addressed the mobile web. Outcomes related to web accessibility including private sector websites, the mobile web and the educational environment including eLearning platforms, still remain unachieved. In Norway, policy has addressed web accessibility, however falls short of addressing the mobile web, and the educational environment including eLearning platform accessibility. Unique among the cases presented, Norway has been successful in achieving accessibility in the educational environment including eLearning. This status produces a counterintuitive argument for addressing E-Accessibility through targeted policy provisions. Like the UK and US, Norway has also not produced achievements in mobile web accessibility, particularly in the public sector.

In its approach to legislation and policy development, the UK proves to be the median case. Policymaking is neither subject to full public scrutiny as in the US, nor is it a completely private affair. Fundamental to policymaking in common law countries, the broadly defined provisions of the Equality Act rely on judicial interpretation for implementation and enforcement. In the UK, policy development more generally appears to be particularly sensitive to the ideological transitions and social pressures inherent in the electoral voting system.

In the US, policymaking is subject to intense public scrutiny where the media plays a substantial role in interpretation and dissemination. The historical and social propensity in US disability policy has been to provide limited financial support and rehabilitation (Berkowitz, 1989). As a part of this orientation, fiscally conservative political parties in the US pioneered social regulation, despite limited private sector support, as a policy solution for achieving social objectives without spending federal revenues, increasing taxes, or increasing the size of the government. The Americans with Disabilities Act (ADA) and Section 508 of the Rehabilitation Act are both judicially enforceable; and similar to the UK, have relied heavily on the interpretation of the courts. Policy development in the US is highly cyclical and dependent on the political party in power. However counter to their political stand against welfare state policies, critical pieces of disability policy have been legislated by the conservative politicians (Burke, 2002).

In Norway, social regulation is characterized by high levels of participation by interest groups and professional organizations and low levels of participation by political appointees (Levi-Faur, 2011). Like the UK, regulatory authorities tend to be independent from other governance bodies (Neumaier, Schweiger & Sedmak, 2008). The Anti-Discrimination and Accessibility Act (AAA) came as a result of the progressive realization of disability rights in Norway and is unique among the cases in directly addressing E-Accessibility. The AAA relies in large part on a low threshold complaint mechanism handled by the Equality and Anti-Discrimination Ombud. Compared to the UK and US, social policy in Norway appears to be much less influenced by political party transitions and judicial interpretation (Levi-Faur, 2011).

In the implementation, monitoring and enforcement of E-Accessibility policies, the UK again appears to be the median case. The UK has deferred much of the implementation of the Equality Act and social services more generally to non-governmental organizations operating within the UK. The provisions of the Equality Act apply to public and private service providers and focus generally on antidiscrimination; however the UK has recognized that these provisions also apply to E-Accessibility (Stienstra, Watzke & Birch, 2007). Notably exempt from these provisions are product manufacturers. Though this exemption has received little attention, it has important implications for how E-Accessibility policies are enforced. UK disability policy also introduced a novel concept, anticipatory reasonable accommodation, which broadened the scope for reasonable accommodation from an individual right to a collective right (excepting undue burden) re-conceptualizing reasonable accommodation as accessibility. Enforcement, while within the competence of the Equality and Human Rights Commission (EHRC), has seen few court cases for web accessibility. Though advocacy organizations may have legal capacity in judicial enforcement, their role is frequently diminished due to resource capacity and legal competence. Out of court settlements including confidentiality agreements have avoided establishing precedence. This, combined with the comparatively lower levels of monitoring required by the Equality Act than its predecessor, the DDA, have contributed to low levels of public awareness.

In the US, social policy has traditionally been deferred to the states. The implementation of Section 508 of the Rehabilitation Act and the ADA targeted public sector activities with public procurement provisions indirectly impacting the private sector. Initiating judicial

enforcement become the responsibility of persons with disabilities, and their representatives (Berkowitz, 1989). US case law has established web accessibility as a component of reasonable accommodation to goods and services (National Federation of the Blind v. Target). This is a valuable precedent since the US does not have a monitoring body for E-Accessibility, and noncompliance must be addressed legally by an individual or class action. Here also, advocacy organizations have a role in enforcement, but suffer the same limitations as in the UK.

Norwegian implementation of the AAA has seen much greater administrative investment of public sector resources compared to the other cases. The low threshold for complaints necessitates a structured approach to enforcement. The role of the courts is minimized since class action suits are not allowed and if the plaintiff loses the case, they can be held responsible for the defendant's attorney fees. In Norway advocacy organizations are more experienced in lobbying the government than in administrative or legal complaints and enforcement. The AAA also establishes a distinct responsibility or authoritative body for monitoring, and additionally embeds requirements within existing public and private sector reporting responsibilities (Norway, 2009).

Conclusion

It is clear from the cases examined that the approaches to achieving E-Accessibility have been influenced and framed by different national policy traditions, the distribution of roles, and the relationships between actors participating in the design and implementation of E-Accessibility policy. These influences have fundamentally framed and structured the ways that the US, UK and Norway have ensured and enhanced social opportunities for persons with disabilities. The challenges facing policymakers revolve around how to balance these social and economic needs often under the scrutiny of the electorate. The fiscal policy argument has centred on creating economies of scale for accessible products and services. These policies help to mitigate the inherent risk in what may be perceived as an unexplored market. While this argument offers easily quantifiable support for fiscal conservatism during times of economic insecurity, it fails to capture the value added processes that come with the Universal Design principles of enhancing usability and accessibility. Social regulation, however, has shown some promise for achieving E-Accessibility outcomes. Although the question of whether it is possible to realize these changes without effective enforcement appears to be contraindicated by the cases presented. This may be the critical factor in the regulatory approach. Enforcement has limited efficacy when faced with low-resourced regulatory agencies, and an over reliance on a complex system of judicial enforcement. These conditions may prevent a well-intentioned policy from achieving its intended impact. Effective social regulation also reinforces the right to accessible information, an area addressed throughout the CRPD.

The CRPD (UN, 2007, Art. 33) clearly identifies national monitoring responsibilities for implementation. Where the cases for this study provide for national E-Accessibility policy monitoring (by public agencies in the UK and Norway, by DPOs in the US), it is implicit that monitoring plays a critical role in enforcement. However it is unclear whether this capacity is being effectively operationalized or evaluated. Part of this ambiguity may be due to the restricted power and resources of the public monitoring institutions (as in the UK and the US) or the limited experience of the DPOs (as in Norway). However establishing an effective monitoring mechanism also depends on how disability is conceptualized and defined, and how regulatory threshold criteria are established and applied, an important issue, but beyond the scope of this paper. The cases do provide a broad understanding for the relationship between monitoring and enforcement as a determinant of effective implementation, however

falls short of demonstrating the extent that monitoring and enforcement mechanisms are integrated, the impact of monitoring outputs on the enforcement process, and how these institutions draw upon shared resources. More generally, there is also limited use of social impact assessments in policy implementation which could be seen as a means for evaluating the efficacy and outcomes for policy monitoring and enforcement.

The cases presented in this paper demonstrate the utility of judicial enforcement, the flexibility of providing a low threshold administrative complaint mechanism, and the importance of monitoring. If policymakers crafting the EEA can resolve these issues through enhanced legal actions provided on the national and supranational level, and establish sustainable monitoring mechanisms, the EU may achieve E-Accessibility, critically addressing areas such as education and healthcare. These mechanisms are supported and supplemented by the provisions of the CRPD providing further impetus for addressing E-Accessibility policy in the EU.

The capacity to achieve E-Accessibility is clearly available for many high-income countries. However, future research must empirically address the mediators to effective policy implementation. What is the relationship between effective enforcement and the capacity and legal competence of NGOs? What is the role of policy transfer and its relation to isomorphism and convergence? How has the rate of technological innovation impacted the abilities of policymakers and researchers to provide timely and vetted responses?

References

Aasen, H. S., Halvorsen, R., & Silva, A. B. d. (2009). *Human rights, dignity and autonomy in health care and social services : Nordic perspectives.* Antwerp; Portland, Or.: Intersentia.

Berkowitz, E. D. (1989). Domestic politics and international expertise in the history of American disability policy. *The Milbank quarterly, 67*, 195-227.

Blanck, P. (2008). Flattening the (inaccessible) cyberworld for people with disabilities. *Assistive technology : the official journal of RESNA, 20*(3), 175-180.

Braithwaite, J., & Drahos, P. (2000). *Global business regulation*. Cambridge [England]; New York: Cambridge University Press.

Burke, T. F. (2002). *Lawyers, lawsuits, and legal rights the battle over litigation in American society*, from http://www.ebrary.com/

Dobbin, F., Simmons, B., & Garrett, G. (2007). The Global Diffusion of Public Policies: Social Construction, Coercion, Competition, or Learning? *Annual Review of Sociology, 33*(1), 449-472. doi: doi:10.1146/annurev.soc.33.090106.142507

EC. (2011). Consultation Document European Accessibility Act.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, Mass.: MIT Press.

Hill, M., & Hupe, P. (2008). *Implementing public policy : an introduction to the study of operational governance*. London: Sage.

Hvinden, B., & Halvorsen, R. (2003). Which way for european disability policy? *Scandinavian Journal of Disability Research, 5*(3), 296-312. doi: 10.1080/15017410309512631

Lawson, A. (2008). *Disability and equality law in Britain : the role of reasonable adjustment.* Oxford; Portland, Or.: Hart Pub.

Levi-Faur, D. (2011). Handbook on the Politics of Regulation, from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=807373

Levi-Faur, D., & Jordana, J. (2004). *The politics of regulation : institutions and regulatory reforms for the age of governance*. Northampton, MA: Edward Elgar.

Luo, L., Wang, D., Hu, J., & Shi, W. J. (2009). *Accessibility in Web 2.0 technology* (pp. 14): IBM Corporation.
Majone, G. (1993). The European Community Between Social Policy and Social Regulation. *JCMS: Journal of Common Market Studies, 31*(2), 153-170. doi: 10.1111/j.1468-5965.1993.tb00455.x
Majone, G. (2005). *Dilemmas of European integration the ambiguities and pitfalls of integration by stealth*, from http://rave.ohiolink.edu/ebooks/ebc/0199274304
Mitchell, J. C. (1983). Case and situation analysis. *The Sociological Review, 31*(2), 187-211. doi: 10.1111/j.1467-954X.1983.tb00387.x
Myhill, W. N., Cogburn, D. L., Samant, D., Addom, B. K., & Blanck, P. (2008). Developing accessible cyberinfrastructure-enabled knowledge communities in the national disability community: theory, practice, and policy. *Assistive technology : the official journal of RESNA, 20*(3), 157-174.
Neumaier, O., Schweiger, G., & Sedmak, C. (2008). *Perspectives on work*. Wien; London: Lit ; Global, distributor].
Norway. (2009). *Act June 20 2008 No 42 relating to a prohibition against discrimination on the basis of disability (the Anti-Discrimination and Accessibility Act)* Unofficial translation.
Powell, W. W., & DiMaggio, P. (1991). *The New institutionalism in organizational analysis*. Chicago: University of Chicago Press.
Ragin, C. C. (1987). *The comparative method : moving beyond qualitative and quantitative strategies*. Berkeley: University of California Press.
Stienstra, D., Watzke, J., & Birch, G. E. (2007). A Three-Way Dance: The Global Public Good and Accessibility in Information Technologies. *Information Society, 23*(3), 149-158.
Technosite, NOVA, & CNIPA. (2011). Monitoring eAccessibility in Europe: 2011 Annual Report.
Timmers, P. (2008). EU e-inclusion policy in context. *info, 10*(5-6), 12-19.
UN. (2007). *Convention on the Rights of Persons with Disabilities and Optional protocol*. [New York]: United Nations.
Waddington, L., & Quinn, G. (2010). *European yearbook of disability law. Volume. 2*. Antwerp; Oxford: Intersentia.
Waldschmidt, A. (2009). Disability policy of the European Union: The supranational level. *ALTER - European Journal of Disability Research / Revue Européenne de Recherche sur le Handicap, 3*(1), 8-23. doi: 10.1016/j.alter.2008.12.002
Yin, R. K. (1994). *Case study research : design and methods*. Thousand Oaks: Sage Publications.

Table 1: Case comparisons of valuation of policy environment across E-Accessibility policy engagement, and status.

		UK	US	Norway
Valuation of Policy Environment	Policy Engagement[1]	High score - Web Accessibility - Web Accessibility Enforcement - eLearning Platform Accessibility Low Score - Web Accessibility Monitoring	Low Score - Mobile Web Policy	High Score - Web Accessibility Low Score - Mobile Web Policy - Educational Environment - eLearning Platform Accessibility
	Status[1]	Low Score - Private Internet - Private Mobile Web - Public Mobile Web	Low Score - Internet - Private Internet - Private Mobile Web - Public Mobile Web - Educational Environment - eLearning	High Score - Educational Environment - eLearning Low Score - Mobile Web - Public Mobile Web

Table 1: Case comparisons of valuation policy environment across E-Accessobility policy engagement, and status.

[1] Elaboration of scalar data with scoring based on tertile calculations from a previous study (Technosite, NOVA & CNIPA, 2011)

Europeanisation of public administration through the building of evaluation capacity in the new EU member states: introduction, scope and significance

Jaroslav Dvorak, PhD
Klaipėda University, Minijos str. 153, Klaipėda, Lithuania

Abstract

The aim of this article is to assess institutional and policy arrangements for evaluation within new EU public administration to manage the European Structural Funds support and use of evaluation requirements of EU structural funds as a limited cased study to map how these arrangements are or are not developing, and isolating the key explanatory factors. In order to achieve the objectives of this research paper, we identified and investigated two variables: the coordination of the evaluation process and the evaluation scope and significance. Methodologically, we have used semi-structured qualitative interviews, quantitative online survey applied to officials, academics and evaluators. Starting from the pre-accession program PHARE as legal obligation, evaluation is used as a tool for accountability to European Commission. Current administrative culture has an impact to the success of establishing effective evaluation capacity in new EU member states.

Keywords: evaluation capacity, utilization of evaluation, monitoring, Lithuania, Bulgaria, Poland, European Union, Structural Funds

Introduction

Purpose of the paper

The aim of this article is to assess institutional and policy arrangements for evaluation within new EU public administration to manage the European Structural Funds support and use of evaluation requirements of EU structural funds as a limited cased study to map how these arrangements are or are not developing, and isolating the key explanatory factors.

Design/methodology

Disciplined-configurative case study was combined with a *structured comparative* method and applied for the analysis of evaluation of the EU Structural and Cohesion funds in Bulgaria, Poland and Lithuania. Using the method of disciplined-configurative analysis, the existing theories were used in order to evaluate the evaluation scope and significance in the new EU states. The data of the EU Structural and Cohesion funds evaluation systems are compared with the Polish and Bulgarian data. The data for the analysis were collected and analyzed applying the *triangulation conception*: (i) document analysis (legal and administrative documents, protocols, reports and media reports); (ii) in-depth expert interview of direct contact and contact by telephone; (iii) quantitative questionnaire of public officials; (iv) SWOT analysis; (v) statistical analysis of the data; (vi) logical distribution and classification; (vii) comparative analysis of the features; (viii) rating.

Findings

The research results show that isomorphism and donor-oriented evaluation dominates in the evaluation systems of the EU Structural and Cohesion funds. Member states transfer the elements necessary for support evaluation to the public administration systems. From the intervention approach, supporting many programmes and projects, there is a change to the approach based on long-term planning, programming and consulting with the stakeholders. The methodological documents of evaluation created by the EU are used and new national methodological evaluation guides are created. The officials participate in the networks of EU evaluation and initiate national evaluation associations or networking. The mentioned circumstances denote the dominating management of evaluations, the basis for which is the institutionalization of evaluation activities. However, it should also be understood that management by evaluations should be used and guaranteed that because of the evaluation study, the government fulfills the evaluation functions analyzed in this research. The EC is dependent on the information of the member states; therefore it attempts to unify the administrative structures and behavior of the member states. The Commission officials are interested in the assimilation of administration of the member states because it facilitates communication, increases professionalism and decreases costs of the contract.

Research limitations/implications

The paper is empirical and more likely to be of interest to EU or Lithuanian practitioners. Critics may argue that presented empirical materials are not linked with any theoretical framework, however during the research I aimed to combine, integrate and consolidate the data to the initial explanation framework which is in the perspective can be raised till the type of theoretical model and further synthetic theory. I'm sure that critics agreed that if there isn't operational theory framework the method becomes the operational theory. That's why in this research the baselines is not the theory but methodology.

The rise of evaluation and its scope in new European union members

The new EU member states governments were not looking for their own efficient solutions in the area of policy evaluation. After the restitution of independence most of them did not have its own strategic programmes, and initially was looking at the templates and samples offered by international organisations. During the first decade of independence, Lithuania implemented programmes initiated by the International Monetary Fund and the World Bank. Therefore the experts hired by the above organisations performed programme assessments, though it also created preconditions for presence of local evaluation experts. Not efficiency, but legitimacy was sought by the development of assessment (Dvorak, 2008, p. 99). Such rational strategy is called *isomorphism*. New EU states governments, thanks to it conformist behaviour, earned the confidence of important external actors, and it guaranteed them access to necessary resources. The use of independent strategic measures started during preparation for membership in the EU. As can be seen, strategies based on isomorphism can become successful.

The support according to the PHARE programme started in 1998 with the aim to support economical and political changes in Central and East European countries. In order to ensure the accountability function, ex-post evaluations were carried out at the European Union level. In addition, the people from the academic community were chosen to learn evaluation from the EU experts (see table 1).

Table 1: Factors influencing the development of evaluation function in the New EU member states

Countries	Membership in the European Union	Internal demand for improving of government decision-making	International organization	Demand from national parliament
	0-non-important 1-important 2- very important	0-non-important 1-important 2- very important	0-non-important 1-important 2- very important	0-non-important 1-important 2- very important
Lithuania	2	1	1	0
Poland	2	1	1	0
Bulgaria	2	0	1	0

All respondents noted that evaluation does not play any role in the formation and implementation process of national policy making; there is an attempt to start applying programming budget, it is legally confirmed in analyzed countries but still the experience in the programmes evaluation is vary from country to country. It should also be noted that national parliament did not have any influence in the origin of evaluation in new EU member states and even today most probably does not understand what evaluation is (Dvorak, 2010, p.55).

Evaluation: the structural perspective

Due to the traditions of public administration and different delegation of functions according to the competence of regional institutions in national contexts, the systems of the EU structural funds management and implementation vary along the scales of centralized/decentralized governing and integrated/non-integrated system (ESTEP, 2006; European Policies Research Centre, 2009). With respect to this, the main ways of organizing evaluation is centralized, decentralized and mixed.

While organizing the evaluation of Structural fund evaluation, Lithuania adapted the approach of centralized evaluation (see Picture 1). Under such circumstances, the process of evaluation is coordinated by the Government of the member state or the evaluation function is delegated to the Ministry of Finance and one evaluation unit was established. Vilpišauskas and Nakrošius (2005) enumerated several advantages of this choice. On the one hand, centralized organization of evaluation provides the possibility to save while hiring less staff and ordering evaluations. On the other hand, evaluation results are more consistent and comparable. Centralized way of coordination provides more advantages as well because evaluation department is independent from the staff which implements programmes, and the employees of the department acquire skills in evaluation methodology. In addition, the department has powers, which decentralized subdivisions usually lack. However, this approach is criticized because of poor decision-making and maintenance of programme effectiveness at programme level. Decentralized evaluation coordination and performance preconditions a more suitable adaptation of evaluation contents for single programmes, and the responsible institutions participate in the evaluation process more, as well as use the results of evaluation recommendations (Vilpišauskas, Nakrošis, 2005, p. 79).

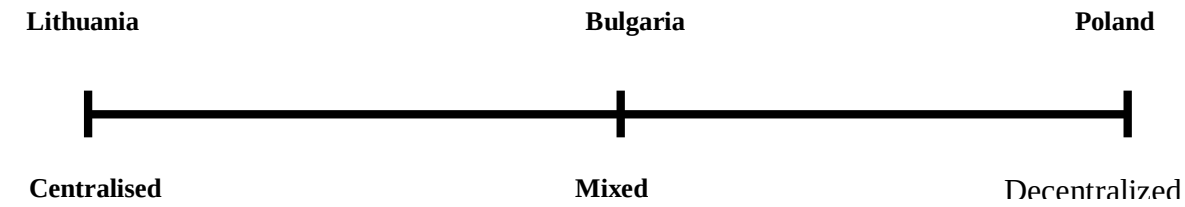


Figure 1: Evaluation approaches of the EU Structural and Cohesion funds of the New EU member states

After the beginning of the decentralization of cohesion policy implementation in Poland, the decentralization of the evaluation process appeared as well. Twenty-four units emerged in the institutional evaluation structure, in the MA. New bodies became responsible for the implementation of evaluation in each EU Structural funds and regional operational programme (henceforth, OP). In order to increase the independence and objectivity of the evaluation process, MA delegated the evaluation competences for the lower implementation level. 29 evaluation units were established in Intermediate bodies (henceforth, IB). Evaluation coordination groups were established in many action programmes, the main task of which is to help the evaluation bodies to implement evaluation process at the corresponding implementation level.

Table 2: The structural perspectives of evaluation

Country / Measure	Lithuania	Bulgaria	Poland
Evaluation units	2	3-4	56
Number of people employed	6 (3-4 persons in other ministries)	2-3 per unit	153
Evaluation management group	Yes	Yes (but not work after evaluation function transfer)	Yes

In October of 2009, it was decided to transfer the functions of general coordination and evaluation of EU Structural and Cohesion funds to the political institution of Bulgaria, the Council of Ministers. It is the strategic centre of formation, development and implementation of internal and foreign policy. Ministries, on the other hand, are specialized units responsible for the development of sector policy. Even though evaluation structure seems to be efficient in the scheme, evaluation subdivisions are *virtual* or they do not exist at all. Only one or two people work in such subdivisions, who have several different responsibilities, one of them being evaluation. Such a situation impedes the evaluation process because one person has to prepare everything from the beginning to the end.

According to Stern (2004), it is necessary to pay attention to the functions which are achieved while performing evaluations while constructing the way of coordinating the evaluation function. When the evaluation improves the implementation of programmes and policies, it is suggested to decentralize the coordination of evaluation function, and this encourages learning at decentralized level. When the function of evaluation is to reinforce central strategies and decision-making, the evaluation function is coordinated at a centralized way, but this way the staff of central governing level will learn from evaluation results, not the level of programme implementation.

Evaluation capacity

In Poland, preparatory work for evaluation system decentralization has been done. Evaluation plans are started, which is an important element in the system. Such plans are of strategic character, because there is a common evaluation plan for 2007-2013, which comprises the entire programme period and defines the main evaluation areas. The respondents noted in one accord that there is a big difference between the central and the regional level. Apparently, in such conditions, regional officers will be dependent on the pieces of advice from the central level officers while preparing technical specifications for the ordering of evaluation, preparing methodology and questions for the coordination of the observation system and organization of public buying.

All the respondents admit that the monitoring system has still remained the weakest part of the EU Structural funds evaluation system. It is claimed that during the period of 2004-2006 "the monitoring of financial expenses was working the most effectively" (the direct report to the Prime Minister, who fails to spent money within 20 days in the end of each month 18), while "the observation of effects, results and products was poorly developed." The monitoring system existed in the shape of Word and Excel files. Comparing Poland with other EU countries, it is argued that "the situation in Poland is similar <…> where the data are late in some other areas as well"(Benias, 2009, p. 125).

Apparently, it is too early to maintain that evaluation capacities exist in Bulgaria; however, it is possible to maintain that there is evaluation infrastructure. The civil servants gained the main evaluation experience while working at the preparation for the membership programme PHARE. Therefore, the forming evaluation system is mixed (more to centralization) in the present period because this was influenced by the centralized character of the PHARE programme. Evaluation abilities were created in the traditional way: the civil servants were organized training, training trips, seminars, internal evaluation exercises financed by the European Commission; twin projects were also implemented, which provided long-term help.

Previous studies (Georgiev, 2006, p. 1; Knott, 2007, p. 59) show that after the implementation of the monitoring system, some problems remain with its institutionalization at many levels. One should admit that the situation has not changed during the last several years. First, it was caused by the changes in human resources. It is apparent that after the personnel change; new people could not say much about the already implemented projects. Public administration does not understand that the methodology of the monitoring system should be very exact. Second, there is a imbalance between communication and coordination because when the implementing institutions started pursuing structural changes, they could not find the necessary documents for the evaluators. The presentation of the documents from various institutions took a long time. Third, project indicators were not described properly, which meant that the conclusions and recommendations by the evaluators could have little analytical value for evaluation customers. Four, the integrated management and monitoring system was not functional and the data were mainly collected by using an interview; therefore, the information was mainly qualitative not quantitative.

Table 3: Evaluation capacity

Country Measure	Lithuania	Bulgaria	Poland
Evaluation capacity	Strong at central evaluation unit, weak in other ministries	Weak at all levels	Strong at central level, weak at local but improve through the practice
Evaluation plan	Yes	Yes	Yes
Quality of the monitoring data	Average (problems with definition of indicators)	Low	Average (but there are improvement after 2004-2006 period)
Training	One performance audit program at university level, training monopoly of the Ministry of Finance	Course on the project management at the universities, no more training	Three postgraduates programs at university level, different courses and annual evaluation conferences

In Lithuania, the project of annual evaluation plan is discussed by an evaluation coordination group, which consists of public servants from various interim institutions. In 2004-2006, the predecessor of the groups was SPD evaluation management group. The composition of public servants of interim institutions changes, but the civil servant from the MF is the head of the group. It is likely that such *monopolization* of evaluation coordination inhibits a quicker evaluation dissemination in the public administration system. The civil servants from other ministries, who work in the evaluation system, do not have the feeling of *ownership* for the evaluation function because as long as they do not participate in managing coordination, this will be a secondary exercise. The evaluation coordination group does not have the representatives of evaluation community; it is maintained that this would not be possible to implement because the evaluation coordination group decides about administrative issues. However, the limited number of evaluators sometimes participate as invited participants. Apparently, the participation of alternative evaluators would precondition the development of the participatory model, strengthen partnership relationship between evaluators and representatives of institutions, the good practice would be exchanges and the evaluators could share their knowledge. In order to fulfill this aim, it is necessary to establish a formal association of evaluators, which would delegate its members and this way would prevent from conflicts of interests.

The monitoring system was started to be created in the programming period of 2004-2006. Foreign and local experts participated in its creation and improvement. Timetables of the monitoring system and its procedures were sufficiently clear in different level institutions; however, the implementation of the information system of Structural Funds Management was late (European Policies Research Centre, 2009). At the time, difficulties were faced in indicating and collecting information of physical implementation indicators; therefore, there were some inaccuracies in interpreting the indicators defined in SPD (European Policies Research Centre, 2009). Information system of Structural funds management was created in order to save, aggregate and prepare the data for reports. It started to function fully only in 2006 (European Policies Research Centre, 2009). However, in reality this information system started providing qualitative data only after 4-6 years of using it. ESTEP (2007) indicated that the quality of monitoring data, temporal presentation stills needs to be improved. The main concern was related to the definition of indicators and the data input to the system on time.

Towards utilization of evaluation results

It is possible to analyze the usage of evaluation recommendations in decision-making using the analytical model by Ferry and Olejniczak (2008). Its essence is that the use of recommendations depends on five main factors related to the creation of evaluation knowledge and stages of use: (i) Characteristics of learner/recipient; (ii) Characteristics of the evaluated policy; (iii) Research time; (iv) Used evaluation approach; (v) Quality of evaluation reports.

Characteristics of the learner/recipient

This factor comprises the quality of public administration human resources and the dominant tradition of public administration. It seems likely that the personnel that has evaluation knowledge, skills and experience understand evaluation advantages better and know how to use them in their work. The stability of institution, position in the political system and the experience in planning and implementing interventions can become an effective stimulus in the use of evaluation results because knowledge is needed in order to solve new complicated situations.

Characteristics of the evaluated policy

The scope of public intervention and its importance on the political process may be the critical factor in using the evaluation results. The evaluation comprising policies will possibly get more attention from the politicians, administrators and the society. It is also similar with the programmes that receive much investment because their results are important for the society, therefore, it is probable that the evaluation results will be used as well.

Research time

Evaluation is performed at different stages of the public policy cycle. While planning a policy or a programme, the ex-ante evaluation is carried out. While implementing the programme, the intermediate evaluation is performed. After the implementation is finished, the ex-post evaluation is done.

The used evaluation approach and quality of report

This factor divides it into two points of access: one oriented towards experts and another oriented towards participation. In the first case, the experts performing the evaluation analyze the programming documents, statistical data and the information provided by the partners. Evaluation customers and the interested sides remain passive during the process of evaluation; therefore the evaluator interprets the proof, provides conclusions and prepares the report. In the second case, the partners are encouraged to participate in the discussion about the programme. Their point of view is important while preparing the recommendations and conclusions. It is likely that the participating partners will understand about evaluation more and use the recommendations in their work. Qualitative preparation of the evaluation reports is the premise for its further usage in the formation of public policy formation. Apparently, this variable depends not only on the evaluators who perform the evaluation but also on the participation of the employees of the client (the managing authority).

According to Rhodes (1996), the concept of governance is very broad and may have at least 6 meanings: minimal state, corporatistic state, new public management, good

governance, socio-cybernetic system, self-organized networks. It is apparent that the distinguished indicators comprise the skills of the government to formulate and implement public policy efficiently. Governance indicators are categorized the following for the characterization of the government: (i) voice and accountability; (ii) political stability and absence of violence; (iii) government effectiveness. The quality of economical institutions is defined: (iv) regulatory quality; (v) rule of law; (vi) control of corruption. Another external indicator, which is used to evaluate the characteristics of a learner is the Corruption perception index (the scale is from zero (highly corrupt) to ten (highly clean)) used by Transparency International (see Table 3). This indicator provides information, which allows to perceive the real state of corruption in the country and is significant in determining the characteristics of the people working in the public sector.

Table 4: TOWARDS UTILIZATION OF EVALUATION RESULTS

Country / Measure	Indicators	Lithuania	Bulgaria	Poland
Characteristics of learner/recipient	WGI*	72	60	70,2
	CPI**	5,0	3,6	5,3
	TEC***	49	59	48
Characteristics of the evaluated policy	EU****	High attention	High attention	High attention
	In*****	Politicians aren't interested	Politicians aren't interested	Politicians aren't interested
Research time		Half cycle (04-06) and full 07-13	Full cycle 07-13	Half cycle (04-06) and full 07-13
Used evaluation approach		Programe theory is not used	Not known	Programe theory is used
Quality of evaluation reports		Average	Low	High

*WGI – World Governance Index 2009; ** Corruption Perception Index 2010; ***Trust in European Commision, 2009 autumn; **** EU level; ***** Internal level

Taking into consideration the obtained results, a generalization can be made that in a well-operating state office, the recommendations received during evaluation become the source of alternative information for the decision maker. Apparently, if corruption has become a legitimate tool in decision-making, no proof is necessary in the decision making process. Very much attention is paid for the EU Cohesion policy and its evaluation on the EU level. The EU budget is prepared based on the evidence; however, in the new member states the Cohesion policy lacks the local officials' feeling of possession; therefore, there is also a

lack of evidence use while reaching the changes in the country through the assimilation of resources of Structural funds. Time of the research is an important factor in the use of evaluation results. However, it was noted that interim evaluation and results are viewed as an unnecessary task because evaluation recommendations are not used as the EU priorities are provided for the member states, and they have to ensure the reflection of priorities in the strategic documents of the country. Under such circumstances, it would be useful to carry out the evaluation of the country's needs. The used evaluation approaches and quality of the report are viewed as important factors, which influence the use of the results. In fact, traditional qualitative and quantitative methods still dominate in the analyzed countries but in Poland, innovative evaluation methods are used because of deeper tradition in the social sciences, and they are also transferred to Lithuania. The quality of evaluation reports is constantly increasing because the officials' skills in project management are improving. In addition, the evaluators in Poland and Lithuania view this business seriously because there is an evaluation plan, according to which future activities may be planned. Inexperienced consultants are still sometimes hired in Bulgaria, and the officials' skills in contract management are poor.

Discussion

The research results show that *isomorphism* and *donor-oriented* evaluation dominates in the evaluation subsystem of the EU Structural and Cohesion funds. Member states transfer the elements necessary for support evaluation to the public administration systems. From the intervention approach, supporting many programmes and projects, there is a change to the approach based on long-term planning, programming and consulting with the interested parties. The methodological documents of evaluation created by the EU are used and new national methodological evaluation guides are created. The officials participate in the networks of EU evaluation and initiate national evaluation associations or networking. The mentioned circumstances denote the dominating *management of evaluations*, the basis for which is the institutionalization of evaluation function in national public administration. However, it should also be understood that *management by evaluations* should be used and guaranteed that because of the evaluation research, the government fulfills the evaluation purposes. The EC is dependent on the information of the member states; therefore it attempts to unify the administrative structures and behaviour of the member states. The Commission officials are interested in the assimilation of administration of the member states because it facilitates communication, increases professionalism and decreases costs of the contract.

As the qualitative and quantitative research shows, the EU member states have different evaluation organizing approaches, taking into consideration public management organization in the state. There is some proof that this had the impact on evaluation scope and significance. Poland chose the decentralized approach, and more than 50 evaluation branches were established where about 150 officials were employed. Decentralization makes premises for evaluation to affect politics when the Government policy control is not very centralized because there are more listeners and readers of evaluation results. In Lithuania and Bulgaria, evaluation function is not widely developed, applying the centralized evaluation approach; it is transmitted to other ministries, even though the skills of the Lithuanian Ministry of Finance are evaluated very well. Implementing political reforms in Bulgaria, evaluation functions were started to be introduced moving towards mixed evaluation organization but evaluation branches exist virtually and the officials' skills are poor because of fluctuation, corruption and lack of training. It is also possible to emphasize that the ministries still consider evaluation as an obligatory procedure and do not seek to get additional value to the public policy process, while evaluation results do not have the learning effect for the whole

institution.

The quantitative and qualitative research on evaluation shows that in all analyzed countries, it was not prepared for the collection of monitoring data and it was not planned what data will be necessary for evaluations. Monitoring of financial data worked best, as it is the inheritance of economic-financial control, which operated quite effectively during the period of socialist regime. The data of physical monitoring were not collected or there was no continuation of data collection because of staff changes. Inappropriate definition of indicators conditioned the scarcity of qualitative data of monitoring system. Even though the second programming period takes place in Lithuania and Poland and both states can be viewed as advanced compared to Bulgaria, the problems of indicator definition are still faced, and evaluation is used as a tool for monitoring data collection.

The qualitative research revealed that the supply of evaluation training differs significantly in the analyzed countries. In Poland, there are four university post-graduate study programmes, which comprise the dominant evaluation approaches: sociology, econometrics and public administration. Each year evaluation conferences are analyzed, consulting enterprises organize evaluation courses, also evaluation fairs are organized, where the evaluation customers meet with the suppliers of the services. Lithuania and Bulgaria do not have evaluation study programmes. One Lithuanian university has an activity audit Master study programme; additional training is organized by the Ministry of Finance according to the project of evaluation skills strengthening. There is some basis to maintain that Bulgaria faces the deficit of evaluation training because a seminar of two directorates was organized by the EC in order to provide the main information for the officials, after which an intensive feedback was received. Conferences on evaluation are not organized and it is early to talk about networking because the human resources which consider themselves as evaluators work abroad. Finally, the evaluations carried out in the analyzed countries may be used for the training of the state officials

References

Dvorak, J. (2008). A Theoretical Interpretation of Policy Evaluation in the Context of Lithuanian Public Sector Reform. *Baltic Journal of Law and Politics* 1(1), 95-110.

Dvorak, J. (2010). Evaluation of the European Union Structural Funds‘ support in Poland; Scope and Significance. *Baltic Journal of Law and Politics* 3 (1), 53-75.

ESTEP (2006). *Lietuva ir ES šalys: struktūrinių fondų paramos lyginamoji analizė. Europos Sąjungos parama: Lietuvos galimybės.* Pilietinės visuomenės institutas: Versus aureus

ESTEP (2007) Final Report on the Framework to analyze the Development of Evaluation Capacity in the EU member states. Retrieved February 12, 2008, from http://ec.europa.eu/regional_policy/sources/docgener/evaluation/pdf/report_integrated_2007.pdf.

European Policies Research Centre. (2009). *Ex-Post Evaluation of Cohesion Policy Programmes 2000-2006*. Co-Financed by the ERDF (Objective 1 and 2). Task 2- National Assessment Report Lithuania. University of Strathclyde: Glasgow.

Ferry, M., Olejniczak, K. (2008) The Use of Evaluation in the management of EU programmes in Poland. Warsaw. Retrieved December 18, 2009, from http://webapp01.ey.com.pl/EYP/WEB/eycom_download.nsf/resources/Evaluation_EU_Funds_Poland.pdf/$FILE/Evaluation_EU_Funds_Poland.pdf

Georgiev, B. (2006). Synergy of Government and Nongovernment Bodies Involvement in Monitoring and Evaluation. (paper presented at the Ideas Workshop: Evaluations and Systems: Practical Experience of the Central and Eastern Europe Region. Prague, Czech

Republic, June 19-20, 2006). Retrieved from http://www.ideas-int.org/documents/file_list.cfm?DocsSubCatID=8;
Knott, J. (2007). The impact of the EU accession process on the establishment of Evaluation Capacity in Bulgaria and Romania. *International Public Policy Review* 3 (1), 49-68
Rhodes, R. A.W. (1996). The New Governance: Governing without government. *Political Studies* 44(4), 652-667
Bienias, S. (2009). Application of economic modeling results in the process of Cohesion Policy Evaluation. In A. Haber & M. Szalaj (Eds.), *Evaluation in the Making Contexts and Methods*.(pp.125-133). Ministry of Regional Development, PARP: Warshaw.
Stern, E. (2004). Philosophies and types of evaluation research. In: P. Descy, M. Tessaring (eds.) *The Foundations of Evaluation and Impact Research*. Luxemburg, CEDEFOP.
Vilpišauskas, R.; Nakrošis, V. (2005). *Ko verta politika?* Vilnius: Eugrimas.

IV

Evaluation of Education Outcomes and Learning

Lessons learned from evaluation practice in Indigenous education: practitioners' comparisons between First Nations and federal government approaches to evaluation of First Nations education in Quebec, Canada

Blair Stevenson, Faculty of Education, University of Oulu, Finland
Nancy Doddridge, First Nations Education Council, Canada

Abstract

This paper draws on the specific case of a First Nations organization in Canada which initiated an evaluation project to assess education outcomes in member communities in 2010-2011. This project was developed parallel to an evaluation approach being developed at the same time by the federal government to assess education in those same First Nations communities. The purpose of this paper is to summarize lessons learned about evaluation practice in an Indigenous context as drawn from the practitioner experiences and observations of the First Nations organizations' evaluation team. These observations focus on a comparison of separate First Nations and federal government evaluation approaches and related lessons learned.

Observations from this process suggest a number of key lessons learned: to address divergent definitions of education outcomes, success and accountability; to undertake appropriate consultation and protocols; and to clarify long-term data governance and ethical oversight. These lessons have implications for both the practice of culturally competent evaluation and for Indigenous organizations and governments attempting to assess education outcomes within an Indigenous context.

Keywords: First Nations, evaluation, Indigenous education,

Introduction

The Indigenous peoples of Canada are often referred to as Aboriginal peoples. The term Aboriginal, however, obscures the distinctiveness of the First Peoples of Canada as specified in the 1982 Constitution Act of Canada — First Nations, Inuit and Métis. This paper focuses on the First Nations peoples in Canada and in particular First Nations communities in the province of Quebec.

Overall, First Nations peoples live throughout Canada in over 630 First Nations communities (AFN, 2012) as well as sizable populations in major centres across the country. The most widely used national data set on population in Canada is the Canada Census.[1] Based on 2006 data, the Canada Census estimates that the Aboriginal population of Canada had reached approximately 1.1 million people, with First Nations representing the largest percentage (61 percent) (Sharpe et al, 2009). Census data also suggests that the total Aboriginal population is growing significantly faster than the non-Aboriginal population with approximately 300,000 Aboriginal children and youth who could enter the labour force over the next 15 years (Standing Committee, 2007).

With regards to education, it must be recognized that, for millennia, the forms of education used by First Nations imparted all that was necessary to survive and perpetuate their own cultures, languages and livelihoods. More recently, as outlined within treaties between First Nations and the British Crown and later through land claims and agreements between the

[1] Note: Census estimates for Aboriginal populations are seen by some organizations as problematic as a result of non-participation by some Aboriginal populations and differences of definitions for Aboriginal peoples.

Canadian federal government and individual First Nation communities, more institutionalized schooling has become established in First Nations communities. Today, it is because of on-going treaty and land claims obligations as well as sections 114 – 122 of the federal Indian Act that the federal government holds the responsibility toward First Nations education rather than the provinces. Under current policies, Aboriginal Affairs and Northern Affairs Canada (AANAC) is the key department responsible for the federal governments' role in First Nations education.

Historically, however, federal policies for implementing their fiduciary responsibilities have been characterised by attempts at assimilation and the passing of responsibility on to churches and provincial governments. These early policies also included the development of a residential schooling system for over a century in which approximately 150,000 Aboriginal children were separated from their families and communities (AANDC, 2012). The Indigenous scholar Marie Battiste summarizes the contemporary results of this legacy well when she states that "while much money and work have been focused on First Nations education in the last century, contemporary schools have not corrected or confronted the lessons of the residential schools and the residual negative stereotypes of First Nations people" (Battiste, 2004, p.1).

In the face of this legacy, Aboriginal communities have become more politically active while calls have grown from First Nations for increasing control of education within their own communities. An early example of this call came in 1972 with the tabling of the policy paper *Indian Control of Indian Education* (AFN, 2010) by the National Indian Brotherhood, which is now known as the Assembly of First Nations (AFN). This document, revised in 2010 by the AFN, along with other recent efforts by educational and political organizations established by First Nations at the national and regional levels reflect philosophies and positions that call for the establishment of more appropriate First Nations education systems.

Evaluation Context

Currently in Canada, program evaluation is increasingly being viewed as a necessary aspect of monitoring and assessing the education received by First Nations students living in First Nations communities. This view parallels general policy within the Canadian federal government in which evaluation is used as a tool within an overall results-based management (RBM) method. To a large degree, evaluation in the realm of First Nations education is synonymous with the funders' (the federal government) need to assess the effectiveness of their financial contributions. As a result, First Nations regularly provide reporting information to the federal government for management or accountability purposes.

However, underpinning this vision for evaluation is the "question of how to satisfy the evaluation needs of funders without trampling on, or otherwise marginalizing, the Aboriginal ways of knowing and communicating" (Johnson, 2008, p.2). Furthermore, McKenzie (1997) suggests that measures used for evaluation in Indigenous contexts should not just be for external accountability.

This question points to the need for balancing government requirements for accountability with First Nations' needs that are grounded in their own cultures and epistemologies. Grover (2008) details this issue specifically with respect to the use of culturally valid measures by outlining the need "to balance the state's desire for pre- and post-survey data with measures that will be more credible to the community" (p.47). Fundamentally, this scenario points to the need for viewing evaluation within First Nations education contexts as an inter-cultural process.

Furthermore, Hopson (2003) suggests that "the challenge is for evaluators to understand how awareness and knowledge of cultural differences in evaluation work can contribute to

different kinds of understandings about what evaluation is and what it can be" (p.3). Chouinard and Cousins (2007) similarly recognize this challenge calling for culturally competent evaluation which recognizes the "relationship between power, knowledge, evaluation use and questions of validity" (p. 54). In order to address these issues, an increasing number of participatory evaluation and research approaches have been designed which attempt to address issues of culture, power and Indigenous participation (Chino & DeBruyn, 2006; First Nations Centre, 2007; Fisher & Ball, 2002). The following summary of a case study involving First Nations in Canada will attempt to inform research surrounding issues of culture and power as drawn from observations made by evaluation practitioners working for First Nations communities.

First Nations Education Council Case Study

Established in 1985 with offices in Wendake, Quebec, Canada, the First Nations Education Council (FNEC) is an association of 22 First Nation communities across Quebec with the goal of defending the interests of its members and striving for full jurisdiction and self-governance in the area of education. It receives its mandate from a General Assembly made up of representatives from the 22 participating communities. FNEC provides pedagogical support services, services for school administration and professional development training opportunities for educators and administrators in member communities. It also offers support toward community implementation of Information and Communications Technologies (ICT) and assists in the development of technology support programs. As part of its role, FNEC also works with its member communities to implement a number of federal initiatives relating to evaluation such as the Education Information System (EIS). The following case study involves FNEC's recent participation in development activities associated with the EIS, which is part of the federal government's long-term Reforming First Nation Education Initiative.

Within the context of First Nations education, it can be generalized that the federal government is the primary funding source which dictates budgets and has an ultimate fiduciary responsibility and accountability to the Treasury Board of Canada. At the same time, First Nations responsibility and accountability is to their community members with First Nations authorities acting as education service providers receiving annual contributions from the federal government. Reporting has, therefore, reflected the government's requirement for financial oversight and monitoring which in turn has involved regular reporting from First Nations to the federal government in the following areas: financial audits, detailed activity reports, teacher information, and personal and service-related student data. This reporting system and the corresponding data management structure has been described by the government itself as a "patchwork" and has most recently been recognized by a recent report from the Auditor General of Canada that called for the federal government to develop "better operational and performance indicators, and the alignment of financial and non-financial data collection for education programs" (Auditor General of Canada, 2011).

In large part as a result of this lack of a coherent system to manage the data, the federal government has recently embarked on the development of the Education Information System to act as a single data management system and warehouse to collect, compile, analyze, and communicate data, as well as track performance and measure success for eventual use by both AANDC and First Nations. The development of the EIS, which is expected to be in operation by September 2012, reflects increasing attention on measurements of achievement and outcomes. It is precisely this current effort to define and measure education outcomes that sets the context for this case study and is of critical importance as the government embarks on a long-term reform of the First Nations education sector.

A critical component of the development and implementation of the EIS has been the establishment of a series of joint Assembly of First Nations/AANDC Experts' Groups tasked with accepting a set of indicators intended to measure improvements. Under this process, representatives from First Nations across Canada, including the FNEC, have been invited to the table by the federal government to finalize a standard set of performance measures of education success. And yet as Nelson-Barber et al. (2005) remind us, "simply inviting everyone to the table does not ensure that the power differential recedes" (p. 71). Consequently, as observed at meetings of these Expert Groups, a number of critical differences have emerged between how First Nations and the federal government believe consultation[2] and evaluation should take place, and what success in First Nations education means. It is these differences, as demonstrated by the divergent evaluation approaches used by First Nations and the federal government, which will be the thematic focus of this paper.

Federal Government Evaluation Approach

Current programs initiated by the Canadian federal government in the area of assessment for First Nations education are strongly influenced by the governments' overall approach to evaluation. This approach is most recently reflected in the 2009 Treasury Board of Canada policy for program evaluations in effect throughout government departments. This policy indicates that evaluations are to be conducted to supply "a comprehensive and reliable base of evaluation evidence that is used to support policy and program improvement, expenditure management, Cabinet decision making, and public reporting (Auditor General of Canada, 2011).

Overall, this approach is grounded in specific definitions of both evaluation and a performance measure. As drawn from Treasury Board documentation, evaluation is defined by the federal government as "the application of systematic methods to periodically and objectively assess effectiveness of programs in achieving expected results, their impacts, both intended and unintended, continued relevance and alternative or more cost-effective ways of achieving expected results" (Treasury Board of Canada, 2012). Performance measure is defined as "a qualitative or quantitative means of measuring an output or outcome, with the intention of gauging the performance of an organization, program, policy or initiative. Quantitative performance measures are composed of a number and a unit" (Treasury Board of Canada, 2012).

Notable in the above definitions is the focus on evaluation being an 'objective' assessment with a corresponding preferential use of quantifiable measures as demonstrated in the above definition on performance measure which only provides an example of a quantitative performance measure as opposed to any qualitative measures. Accordingly, federal evaluations are often limited to quantitative measures that can be readily used in results-based management frameworks for numerical assessment of progress and to determine accountability. This vision of current evaluation practice by the Canadian federal government is echoed by the authors in Gauthier et al. (2009) who suggest that "the current views of evaluation tend to be rigid…and the focus is on logic models and performance measurement. Evaluation is used to appease funders, and their emphasis is on accountability" (p.8). Gauthier et al. (2009) also propose that federal evaluations have become "a form of audit" (p.10).

[2] Note: Both First Nations and the federal government do not call the EIS meeting process a consultation since there is a formal policy on consultation that the government has enacted. In fact, the federal government continues to claim that they do not have to formally consult First Nations on the EIS. The duty to consult policy can be found at: http://www.aadnc-aandc.gc.ca/eng/1100100014664

First Nations Evaluation Approach

In order to begin a discussion of formal program evaluation approaches within a First Nations education context, it is important to ground the discussion in two significant factors. Firstly, it must be recognized that the federal government continues to exact considerable control and power over the system as a whole. As a result, current evaluation and related initiatives are firmly rooted in the legacy of colonialism that continues to heavily influence the design, implementation, and management of educational programming within First Nations communities.

Secondly, evaluations are strongly influenced by the fundamental differences between Indigenous epistemology and ontology and those of the 'Western' tradition. Fundamental to this difference is that Indigenous knowledge systems are not easily bound by 'Western' definitions and instead are complex and tied to the context within which they are being discussed. Ermine (1995) describes this as Western science's propensity toward the "fragmentation of the constituents of existence…into neatly packaged concepts" (p. 103). Ermine (1995) goes on to suggest that this process of division has led to a "vicious circle of atomistic thinking that restricts the capacity for holism" (p. 103). Consequently, evaluation in the First Nations context does not easily fit into the formal definitions and practices in use within the federal government's approach which commonly includes measures that are 'numbers and units'.

Likewise, Chouinard and Cousins (2007) suggest that it is difficult to determine evidence-based progress in many Aboriginal contexts because "outcome indicators cannot be so neatly demarcated and contained, as outcomes are often integrated into the culture of the broader community" (p. 49). In response to this differing cultural and epistemological grounding, Chouinard and Cousins (2009) go on to suggest that more elaborated strategies are required for "developing culturally and contextually appropriate approaches to outcome measurement" (p. 49). Therefore, they suggest that it is culturally and contextually appropriate methods which better reflect evaluation in First Nations contexts. We tend to agree.

And yet, what is a culturally and contextually appropriate evaluation method? It is this type of approach which FNEC attempted to initiate. With the above two factors in mind and based on the ultimate goal of offering an alternative to the educational evaluation processes initiated by the federal government, a separate evaluation process controlled by First Nations was undertaken concurrently by FNEC between February and May 2011. This process was established as an alternative evaluation process and reflected FNEC's attempts to "take ownership for the process of defining success" (Lafrance and Nichols, 2008, p.18).

This alternative process was governed and managed by a Performance Measurement Committee made up of representatives from a number of FNEC member communities. Once the governance structure was established, two products were designed and implemented which FNEC viewed as the key activities in their evaluation: 1. a set of draft education performance indicators, and 2. a corresponding data management protocol. These two documents were intended for use by First Nations working with the First Nations Education Council (FNEC) in contrast to the tools and procedures in development through the EIS process initiated by the federal government. Additionally, it must be noted that, as part of this process, FNEC hired an external evaluator (co-author Blair Stevenson) to offer advice and training with the objective of clarifying federal evaluation procedures and supporting FNEC capacity building efforts in the area of evaluation. This contract evaluator, however, worked under the guidance and direction of a project governance structure that was fully under First Nations control and management. Ultimately, this use of an outside evaluator was predicated on the principle of developing capacity building similar to what Grover (2010) suggests as

the importance of an evaluator working with the community to increase its capacity for planning and evaluation.

Education Performance Indicators and Data Management Protocol

First Nations education indicators of success were developed to be used by schools and communities to gauge performance in the delivery of education programs in their own schools and to create a regional aggregate profile of educational success. A process was initiated to bring together common indicators as agreed by communities to be measured using data collected and analyzed by the FNEC through a new web-based, data collection tool.

The development of indicators was begun by undertaking a scan of research into existing First Nations education performance indicators and research protocols. Based on this literature review, examples of indicators previously used by First Nations to gauge educational success were organized under five thematic areas: Academic, Cultural, Linguistic, Holistic, and Systemic Indicators. After this research scan was completed, a series of consultations were undertaken during March 2011 with individual First Nations in order to receive comments and feedback on the design of an initial draft listing of performance indicators. This process was repeated as later versions of the indicators were drafted.

The second activity conducted for the FNEC evaluation process was the design of a data management protocol to set out the parameters under which data would be collected and used for those education performance indicators measured. The purpose of this protocol document was to establish a framework of principles and procedures to guide the collection and use of common indicators data. This document acted as an agreement between the community and the FNEC and was drafted using a similar consultation process as that used for the design of the education indicators. Furthermore, this process was comparable to what LaFrance and Nichols (2010) suggest as "protocols appropriate to tribal practices" (p. 17).

Grounding this entire process were the principles of Ownership, Control, Access, and Possession (OCAP). Originally introduced in 1998 by the National Steering Committee of the First Nations and Inuit Regional Longitudinal Health Survey, OCAP principles are increasingly recognized in Canada and have been described as 'self-determination applied to research' (Schnarch, 2004) when applied in First Nations contexts. Furthermore, one of the key aspects of the OCAP approach is ensuring a capacity building component in order that First Nations can conduct their own research and become more fully involved in the research process. For the purposes on this evaluation process, OCAP was defined as the following:

Ownership: The notion of *ownership* refers to the relationship of a First Nations community to its cultural knowledge/data/information. The principle states that a community or group, in this case a community, owns information collectively in the same way that an individual owns his or her personal information.

Control: The principle of *control* asserts that First Nations, their communities and representative bodies are within their rights in seeking to control research and information management processes which impact them.

Access: First Nations people must have access to information and data about themselves and their communities, regardless of where these are currently held. The principle also refers to the right of First Nations communities and organizations to manage and make decisions regarding access to their collective information.

Possession: While *ownership* identifies the relationship between a people and their data in principle, the concept of *possession* is more literal. Although not a condition of ownership, possession (of data) is a mechanism by which ownership can be asserted and protected (First Nations Centre, 2007, p.5).

Discussion

As a result of participating in the development of evaluation practices for the EIS with the federal government, the authors have observed a number of key issues influencing the overall success of the process. It is suggested here that these issues arise from fundamentally divergent approaches for evaluation and data collection outlined above between First Nations organizations and the federal government. The authors propose that these divergent approaches are grounded in differences with respect to the key responsibilities of the agents involved – the fiduciary responsibility of the federal government and the responsibility for community accountability of First Nations.

Based on observations of the FNEC practitioners, three key lessons learned have been drawn from this process. These lessons are: 1. Address divergent definitions of education outcomes, success and accountability; 2. Undertake appropriate consultation and protocols; and 3. Clarify long-term data governance and ethical oversight.

1. Address divergent definitions of education outcomes, success and accountability

As noted above, FNEC and the federal government draw from significantly different approaches for evaluation. However, these differences are not commonly recognized by federal programs leading in turn to a 'one-size-fits-all' approach. An example of this approach can be seen in the federal government's performance measurement strategy for education in which eighteen non-negotiable indicators have been demanded under the EIS program. Examples include indicators related to literacy, numeracy, student retention, student attendance, and graduation rates. By demanding these indicators, it must be supposed that the federal government holds the assumption that these indicators can in fact be measured and compared across a system that is fundamentally diverse (a highly questionable assumption based on the fact that Canada is made up of multiple provincial and territorial education systems with distinctive systems for educational assessment).

The demand for these quantitative indicators also suggests that the federal government views literacy and numeracy skills as the foundation for success in education. And yet, this choice leads to the questions: Why only these indictors? Why not alternative indicators that may measure success in a different way? In the end, the choice for these required indicators points to a fundamental vision for how education should be evaluated and reflects a process of program design that has not questioned the epistemological and ontological grounding of the system that it attempts to assess.

From the perspective of FNEC practitioners, what has been lacking in this process is significant inter-cultural dialogue that focuses on the epistemological and ontological differences between the 'Western' model of assessment as used by the federal government and a more 'holistic' and context-based vision of assessment for use in First Nations communities.

2. Undertake appropriate consultation and protocols

Ultimately, the EIS represents a program unilaterally designed by the federal government and developed as a result of limited discussions with First Nations representatives. The small group who were involved in the AANDC/AFN Expert Groups learned of the development of the EIS and related performance indicators in December 2009. By March 2010 a series of national EIS information sessions with First Nations Educational Directors had concluded, however as of December 2011, community-based political leaders had yet to be informed by AANDC of the EIS and its intended purpose.

From the perspective of FNEC, these consultations were limited in comparison to the consultation protocols necessary for appropriate and ethical consultation practices used commonly within First Nations organizations and associated with the federal government's duty to consult. Furthermore, the content of the EIS discussions were a review and fine-tuning of products already developed by the government rather than discussions of the fundamental purpose, objectives and structure for an information system. The authors view that it is critical that these types of discussions take place and that appropriate time is given in order for First Nations organizations to consult with their membership appropriately, especially since data collected within the EIS may in fact have far-reaching effects in the future on possible funding levels and program planning.

3. Clarify long-term data governance and ethical oversight

During the discussions initiated during the development of the Education Information System, the issue of data governance often came to the forefront. First Nations organizations such as FNEC focused on the importance of recognizing the OCAP principles as underlying any discussion on data governance. By grounding any oversight process on these principles, the authors submit that concerns relating to issues such as third party access and the publication of aggregate data can be alleviated. Developing appropriate and ethical oversight and governance agreements is, therefore, critical especially in an environment such as First Nations education in which maintaining confidentiality in small-population communities is of the utmost importance and ensuring that data is not easily misrepresented.

Conclusion

Ultimately, the lessons outlined above have significant implications for both the practice of culturally competent evaluation and for First Nations organizations and federal governments attempting to assess education outcomes within First Nations contexts. By addressing them head on, a more equitable process can be developed. If left untouched, these factors will feed further mistrust and continue to challenge future efforts at reaching a robust and appropriate system of assessment.

If taken further into consideration, then programs such as EIS would more closely parallel the components for assessment and research in general outlined in a growing number of research protocols developed specifically by First Nations organizations (AFNQL, 2005; First Nations Centre, 2007). Notable examples also exist of specific national level projects which incorporate governance and protocol structures more reflective of First Nations needs such as the First Nations Regional Longitudinal Health Survey (RHS). These examples should be more readily referenced as a model for on-going evaluation processes.

Conclusions point to the need for more sustained dialogue between First Nations organizations / communities and the federal government about the basic assumptions behind and objectives for the assessment of Indigenous education before evaluation systems are formally established. This dialogue should be based on an inclusive, joint agenda and the principles of First Nations OCAP must be at the forefront of every discussion.

If the goal of the EIS is in fact to support First Nations educators and improve school results, then the connection must be obvious and transparent with respect to how the activities supported under this program will directly improve the work of teachers and schools. It is ultimately proposed by the authors that, in essence, it is teachers and schools, not the federal government, that are in the best position to influence practice in the classroom and assessment tools should be primarily targeted to aid the work of those that work at that level.

References

Aboriginal Affairs and Northern Development Canada website (2012). *Residential Schools Apology Statement*, Retrieved January 14, 2012 at http://www.aadnc-aandc.gc.ca/eng/1100100015644

Assembly of First Nations website (2012), *About AFN*, Retrieved January 15, 2012 at http://www.afn.ca/index.php/en/about-afn/description-of-the-afn

Assembly of First Nations. (2010). *First Nations Control of First Nations Education*. Ottawa: Assembly of First Nations. Retrieved February 15, 2012 at http://www.afn.ca/uploads/files/education/3._2010_july_afn_first_nations_control_of_first_nations_education_final_eng.pdf

Assembly of the First Nations of Quebec and Labrador. (2005). First Nations of Quebec and Labrador Research Protocol. Wendake: Author.

Auditor General of Canada. (2011). *June 2011 Status Report, Chapter 4: Programs for First Nations on Reserves*, Ottawa: Government of Canada, Retrieved January 30, 2012 at http://www.oag-bvg.gc.ca/internet/English/parl_oag_201106_e_35354.html

Battiste, M. (2004). *Animating Sites of Postcolonial Education: Indigenous Knowledge and the Humanities*, University of Saskatchewan CSSE Plenary Address May 2004, Retrieved January 25, 2012 at http://www.usask.ca/education/people/battistem/csse_battiste.htm

Chino, M., & DeBruyn, L. (2006). Building True Capacity: Indigenous Models for Indigenous Communities, *American Journal of Public Health*. April; 96 (4), 596–599.

Chouinard, J.A. & Cousins, J.B. (2007). Culturally Competent Evaluation for Aboriginal Communities: A Review of the Empirical Literature. *Journal of MultiDisciplinary Evaluation*, 4 (8).

Ermine, W. (1995). Aboriginal Epistemology. In M. Battiste & J. Barman (Eds.) First Nations Education in Canada: the Circle Unfolds (pp. 101-112). Vancouver: UBC Press.

First Nations Centre. (2007). *OCAP: Ownership, Control, Access and Possession*. Sanctioned by the First Nations Information Governance Committee, Assembly of First Nations. Ottawa: National Aboriginal Health Organization.

Fisher, P. A., & Ball, T. J. (2002). The Indian family wellness project: An application of the tribal participatory research model. *Prevention Science*, 3(3), 235-240.

Gauthier, B., Barrington, G., Bozzo, S., Chaytor, K., Dignard, A., Jahey, R., Malatest, R., McDavid, J., Mason, G., Mayne, J., Porteus, N. & Roy, S. (2009). The Lay of the Land: Evaluation Practice in Canada in 2009. Canadian Journal of Program Evaluation, 24(1), 1-50.

Grover, J. (2008). Challenges in Applying Indigenous Evaluation Practices in Mainstream Grant Program in Indigenous Communities, *Canadian Journal of Program Evaluation*, 23(2), 33-50.

Hopson, R. (2003). *Overview of multicultural and culturally competent program evaluation: Issues, challenges and opportunities*, Social Policy Research Associates, Oakland: The California Endowment.

Johnson, A. (2008). Aboriginal Ways of Knowing: Aboriginal-led Evaluation – Introduction, *Canadian Journal of Program Evaluation*, 23(2), 1-12.

Lafrance, J. & Nichols R. (2008). Reframing Evaluation: Defining an Indigenous Evaluation Framework, *Canadian Journal of Program Evaluation*, 23(2), 13-32.

McKenzie, B. (1997). Developing First Nations child welfare standards. *Canadian Journal of Program Evaluation*, 12(1), 133-148.

Nelson-Barber, S., LaFrance, J., Trumbull, E., & Aburto, S. (2005) Promoting culturally reliable and valid evaluation practice (Chapter 5), In S. Hood, R. Hopson, H. Frierson

(Eds.), *The role of culture and cultural context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice* (pp. 61-85.). Greenwich, Connecticut: Information Age Publishing.

Sharpe A., Arsenault, J., Lapointe, S., & Cowan, F. (2009). *The Effect of Increasing Aboriginal Educational Attainment on the Labour Force, Output and the Fiscal Balance.* Ottawa: Centre for the Study of Living Standards.

Schnarch, B. (2004). *Ownership, Control, Access, and Possession (OCAP) or Self-Determination Applied to Research: A Critical Analysis of Contemporary First Nations Research and Some Options for First Nation Communities.* Journal of Aboriginal Health, January 2004, 80-95.

Standing Committee on Aboriginal Affairs and Northern Development. (2007). *No Higher Priority: Aboriginal Post-secondary Education in Canada*. Ottawa: Government of Canada.

Treasury Board of Canada website (2011). *Results-Based Management Lexicon*, Retrieved February 15, 2012 at http://www.tbs-sct.gc.ca/cee/pubs/lex-eng.asp.

Australian higher education evaluation through assurance of learning

By Shelley Kinash, Trishita Mathew, Romy Lawson, James Herbert, Erica French, Tracy Taylor, Cathy Hall, Eveline Fallshaw & Jane Summers

The authors acknowledge research funding awarded by the Australian Learning and Teaching Council (now Office of Learning and Teaching) Strategic Priority Fund.

Abstract

A collaborative research project conducted by five Australian universities inquired into the philosophy and motivation for Assurance of Learning (AoL) as a process of education evaluation. Associate Deans Teaching and Learning representing Business schools from twenty-five universities across Australia participated in telephone interviews. Data was analysed using NVIVO9. Results indicated that articulated rationale for AoL was both ensuring that students had acquired the attributes and skills the universities claimed they had, and the philosophy of continuous improvement. AoL was motivated both by ritualistic objectives to satisfy accreditation requirements and virtuous agendas for quality improvement. Closing-the-loop was emphasised, but was mostly wishful thinking for next steps beyond data collection and reporting. AoL was conceptualised as one element within the larger context of quality review, but there was no evidence of comprehensive frameworks or strategic plans.

Introduction

Universities worldwide are watching Australia to see the process unfold and the outcomes revealed, as bold new reforms are recreating higher education evaluation. In order that the Australian context might be used as a global case study of education evaluation, this paper begins by describing stakeholders, documents and reforms in higher education. The paper then proceeds to describe the outcomes of a research project whereby twenty-five Australian Business Schools shared their approaches to education evaluation and closing-the-loop through assurance of learning.

The first significant entry on the evaluation reform timeline was the 2008 publication of what is colloquially referred to as the *Bradley Review* (Bradley, Noonan, Nugent, Scales, 2008). Many of the recommendations from this *Review of Australian Higher Education* necessitated a reform of the higher education evaluation system. As follow-on from the review, the 2011-2012 Australian Budget included the formation of the *Advancing Quality in Higher Education* (AQHE) initiative (Department of Education, Employment, and Workplace Relations Australia, 2011).

In December 2011, AQHE distributed three Discussion Papers to diverse stakeholders in the higher education sector, with the response deadline set midway through February 2012. The paper titled *Development of Performance Measurement Instruments in Higher Education* (Department of Education, Employment, and Workplace Relations Australia, 2011) is an overview document, outlining processes and describing the evaluative context. Embedded throughout the discussion paper, three main purposes of education evaluation reforms are described; these include accountability, consumer choice through transparency and comparison, and performance improvement. The discussion paper posed multiple questions for sector response, raising such issues as centralisation versus institutional administration of evaluation, balancing parsimony with complexity, and avoiding harm from misinterpretation and de-contextualized results.

Table 1
Key stakeholders, documents and reforms in Australian higher education

Key terms/bodies in overarching Australian higher education
Review of Australian Higher Education (Bradley, Noonan, Nugent & Scales, 2008)
Advancing Quality in Higher Education (AQHE)
MyUniversity Website (myuniversity.gov.au)
Discussion papers released by AQHE
Development of performance measurement instruments in higher education
Review of the Australian graduate survey
Assessment of generic skills

Another of the AQHE discussion papers, titled *Review of the Australian Graduate Survey* (Department of Education, Employment, and Workplace Relations Australia, 2011) identified, discussed and queried reform options for existing and proposed surveys, primarily of graduates. The content of this paper implied that one efficacious approach to evaluating higher education is to survey graduates a few months after their ceremony. The evaluative information sought is whether the graduates are employed in their discipline of study and their post-study perception of their university experience, specifically learner engagement, teaching and support, and educational development (Radloff, Coates, James, & Krause, 2011). The key problem of this evaluative approach is the diversity of graduates and destinations. For example, the educative experience cannot be considered the independent variable leading to unemployment of domestic students returning to regional remote Australia, or international students to countries with low socio-economic status. In addition, evaluating higher education on the basis of early graduate employment socially constructs universities as performative manufacturers of employees (Marginson, 2009).

As compared to the other discussion papers, the third and final, titled *Assessment of Generic Skills* (Department of Education, Employment, and Workplace Relations Australia, 2011) presented the most contentious approach to higher education evaluation. This approach suggests consideration of a single test to assess the skills of students in the final stages of their respective degrees as a reflection of the value-added by their university education. Further, the approach suggests that universities will be ranked and compared through a public *My University* website (Department of Education, Employment, and Workplace Relations Australia, 2010). Sector responses to the proposed evaluative strategy query the feasibility, reliability and validity of this approach to education evaluation (Gora, 2010; Thorpe, 2011). Specifically, there are concerns about how Australian universities will be compared; will all universities be ranked on a single scale? Will similar institutions be compared? Furthermore, how will similarity or likeness be determined and who will determine the groupings? Stakeholders also questioned the usefulness of an over-simplified score to employers and graduates (Devlin, 2010). In other words, fitness of purpose may not hold for the evaluative approach, any more so than the validity of a summation score purported to reflect quality of the respective universities.

The discussion papers described above reflect the context of higher education evaluative reform in Australia. It is clear that evaluating quality in higher education is in a state of flux and that an enforced standardized process of evaluation is likely. There is widespread sector discomfort with the approaches proposed to date (Devlin, 2010; Thorpe, 2011). In summary of the context of Australian higher education evaluation described above, graduate employment may not be a valid and reliable measure of learning and teaching quality. Neither

can a test adequately measure generic exit skills of students and link these back to causal factors of learning and teaching through a given university. There is widespread dissatisfaction with the proposed means of education evaluation. The as yet unresolved question is how to efficaciously measure the quality of university education and thereby make improvements to benefit stakeholders.

This paper addresses Assurance of Learning (AoL) as an education evaluation alternative. AoL is becoming one of the most frequently discussed topics in tertiary education today (Altbach, Reisberg, & Rumbley, 2009; Martell & Caldron, 2009; Smith, Meijer, Kielly-Coleman, 2010). In the context of education evaluation, AoL refers to the capturing, monitoring and evaluating of data indicating student achievement related to specific program goals. AoL is gaining popularity as an emerging means of informing quality assurance in tertiary education through developing systems and processes for capturing and monitoring direct measures of learning achievement as related to generic cross-disciplinary attributes and program specific learning goals. AoL serves as a recursive means of explicitly depicting, evaluating, developing and enhancing university teaching and learning, as the processes include defining operational program goals (Gardiner, Corbitt, & Adams, 2010) and ensuring there is a strong interconnected relationship between the articulated learning outcomes and means of assessing their attainment (Biggs & Tang, 2007).

This paper reports and analyses the philosophy and motivation for AoL addressed in the first phase of a multi-faceted research study conducted by a collaboration of five Australian universities. The first phase collected data from personnel in business schools of twenty-five Australian universities; business was selected as the phase one discipline because AoL is salient for experts in this key content area due to AoL's inclusion as a factor in Association to Advance Collegiate School of Business (AACSB) accreditation (AACSB International Accreditation Coordinating Committee, 2007; AACSB International Accreditation Quality Committee, 2007).

The relationship between higher education *evaluation* and *quality assurance* is that the former is the process and the latter is the intended outcome. The research project described in this paper addresses both. This paper focuses on the intended outcome of AoL while an upcoming paper will focus on the process. The key question of this inquiry is:

- What are the philosophy and motivators for assurance of learning?

The sub-questions addressed in the analysis are:

- Is AACSB accreditation the driver or a by-product of quality assurance in Australian Business schools?
- How developed are plans for 'closing-the-loop' in assuring learning and thereby applying the results of evaluation to curriculum and quality improvement?
- To what extent is assurance of learning addressed in the context of a larger process of quality review in which other service components are addressed?

Method

Data was collected via semi-structured telephone interviews. The interviews were conducted by an experienced interviewer and lasted approximately 45 minutes. The sample comprised Associate Deans Teaching and Learning (ADTL) (or equivalent) from Business Schools in all Australian Universities. The participants were recruited through the *Australian Business Dean's Council Teaching & Learning Network*. All participation was voluntary and responses were treated as anonymous. The sampling frame was all 41 Australian Business Schools

ADT&Ls of which 25 volunteered to be interviewed for this study. Therefore, the response rate was 61%.

Table 2
Description of participating universities by type and location

Type of University				
Research	Technology	Regional	Other	Total
6	4	6	9	25

Location of University by State							
ACT	NSW	VIC	QLD	SA	WA	TAS	Total
2	5	6	7	1	3	1	25

Analysis

All interviews were taped and transcribed. Data was analysed at two levels using NVivo9, a qualitative data analysis software. In the first instance, open coding was undertaken to identify general themes. Following open coding, axial coding was undertaken where relationships within general themes (sub-themes) were identified. Two co-authors undertook both open and axial coding independently and then discussed and reviewed the results and agreed on the themes presented.

Results

The four main themes to emerge from full analyses of the complete data set of interview transcripts, listed in descending order of strength of theme, were graduate attributes; AoL; challenges faced; and, general suggestions by the university representatives on the AoL process. Within these four overarching themes, several sub-themes emerged. This paper reports on the analysis of the data from the AoL theme. The other three themes are addressed in other papers. The axial themes relating to the open theme of AoL were: context of AACSB, closing-the-loop, and quality review.

Processes of AoL Currently in Place

In order to contextualise the philosophy and motivation for AoL, it is important to provide a brief description of the way in which Australian business schools are operationally defining the process. There were a number of approaches used by universities to assure learning, such as specific tools developed by or for their university, designing capstone subjects, employing moderators and/or external assessors, development of rubrics, random sampling of student assessments, coordinating review teams, hosting workshops for staff, benchmarking against other universities, and creating curriculum maps. Methods will be described in detail in an upcoming paper.

Underlying Philosophy and Motivators for AoL

There were two main underlying philosophies of AoL as portrayed by the respondents: ensuring that students had acquired the attributes and skills the universities claimed they had; and the philosophy of continuous improvement. In relation to the first rationale of accountability, a representative participant comment was, "The question is: are our students learning what they should be learning and do we have evidence that they are learning those things. So what should a business graduate look like?" Regarding the second theme of continuous improvement, another respondent stated, "we see AoL as being fundamentally a strategy for continuous curriculum improvement. So the link to the exercise of AoL is not productive unless it indicates something about how you have changed the curriculum in response to that." The results of the two rationales of accountability and development are consistent with what is written in the literature regarding why universities participate in AoL. Kai (2009) articulated the objectives of higher education evaluation as "ensuring and improving quality" (p. 39).

While these philosophies describe the ideological reasons for administering AoL, another theme of the interviews was the practical or business reasons. While several university representatives acknowledged that they were driven to put AoL processes in place by external bodies such as Australian Universities Quality Agency (AUQA), Tertiary Education Quality and Standards Agency (TEQSA), or Association to Advance Collegiate Schools of Business (AACSB), they also stated that such processes were robust educational practices and married well with their philosophies of continuous improvement and ensuring that their graduates had the capabilities that they claimed they had. As succinctly stated by one university representative, "well, it's a quality management logic. You know if you say you're going to give qualities, then you need ascertain whether you're doing it."

AACSB as Driver or By-product

Several respondents emphasised that AACSB accreditation provided the initial impetus and leverage for AoL processes. With AACSB as the initial driver, other programs and faculties took on the journey and AoL data collection and mapping extended beyond that required specifically by and for the accrediting body. Other respondents explicitly stated that AACSB was never a driver and was always a by-product. "If we are successful at something like AACSB that's a by-product it's not the purpose or the intent. So really fundamentally I feel it's a moral and legal obligation that we have that we fulfil the promises or the contract that we enter into with our students." Some respondents explained that accreditations form a subset of the data that they collect under the umbrella of AoL. "The only accreditations we really focus on are professional body accreditations that have their own set of elements that they want us to look at around generic skills." Most respondents emphasised that the main driver for the AoL process was the "growing accountability of universities."

Closing-the-Loop

Respondents articulated that they were primarily collecting data regarding student performance on learning objectives within each program and that they were using this data to improve the program and the process. For example, one respondent stated that they used this data to "...look at what's actually happening, are we scaffolding enough; have we given enough support; do we need to add additional modules that people can access and students can access to help them with skill development?" Another respondent explained that their university had two rounds of AoL and then a program review every five years, where

curriculum changes took place. This respondent emphasised that changes could not be made based on just one observation. Closing-the-loop was considered an important exercise for most universities in the sample as summarised by one respondent, "the link to the exercise of AoL is not productive unless it indicates something about how you have changed the curriculum in response to that. So in a sense closing-the-loop!" Whereas sentiment strongly emphasised the importance of putting action plans into place as a follow-on from the AoL analysis, few respondents were able to describe the actual processes that they put into place. They remained occupied by the process of AoL and closing-the-loop was largely aspiration rather than achievement.

AoL in the Context of Quality Review

While most respondents did not explicitly list other standards of quality in universities in addition to AoL, other components of a quality framework were implicit in many of the statements. For example, a number of respondents underscored the importance of equivalence of delivery of courses across campuses as an aspect of quality. Another example of an alternate aspect of quality in universities was an expressed concern of one respondent in regard to the English language proficiency of international students. This university had created a screening instrument for all new students to undertake and based on the results of the instrument, students would be directed to undertake a particular course. Another respondent explicitly stated "the key issues are the AoL process, the qualifications of the faculty and increasing the number of full time faculty that are doing the teaching."

Discussion

One of the resounding discussions on Australian campuses is whether to wait for clear directions from the Tertiary Education Quality and Standards Agency (TEQSA) as to what quality data they will require, or develop and follow-through on an institutional process design in hopes that the required data has been collected when the audit information is requested. One analogy is that of audit/accreditation as *driver* or *surveyor*. Universities who conceptualise audit/accreditation as *driver* will collect and map the information articulated by the respective authority. The metaphor of *surveyor* suggests that instead of being passenger to another's journey, the university will determine its own course of action, and provide the required responses for any accreditation from a larger data set collected for their own quality agenda. Respondents in this research acknowledged interplay between higher education evaluation/audit driving and surveying their AoL endeavours.

As another means of expressing this duality, the respondents involved in this study expressed a combination of ritualistic and virtuous motivations in this decision process. AACSB accreditation held the authority and provided the leverage to develop and follow-through on rigorous AoL processes that many knew they should be undertaking anyway. These results are consistent with the results of other research studies. Menassa, Safi, and Chaar (2009) situated their research in Lebanese business schools. They conducted 88 face-to-face interviews/questionnaires with stakeholders from six universities. The authors interpreted the data to indicate that quality improvement was a concern that extended before and beyond accreditation and that the primary rationale for AACSB achievement would be to enhance international recognition and marketing.

The analysis of the research with Australian university representatives reported in this paper confirmed some of what has been shared previously in opinion papers in the higher education literature. Templin and Blankenship (2007) articulated the dilemmas of the relationship between quality assurance and quality improvement as a series of questions and

responses. One of the questions was, “do we need accreditation to conduct self-study or gain insights from external constituents or to improve our lot?” (p. 151). Their response was simply “indeed not” in that institutional research and quality enhancement would occur with or without accreditation. They asked, “do we have to be accredited to maintain excellence or professionalize our programs or students?” and replied “probably not” (p. 151). They posed the question, “does it assure quality and improvement?” and “does it lead to a data set or portfolio from which we can learn about the effect of our accredited professional preparation programs?” (p. 151). The authors were not as specific about the response to these questions. Their analysis implies that accreditation/audit is not the driver of change, which would happen anyway because universities are committed to quality enhancement. However, accreditation/audit does provide the buy-in power of externally mandated and/or defined expectations for ongoing data collection, mapping and reporting.

In addition to confirming previous research on the relationship between higher education evaluation and advancing the student learning experience, the research described in this paper has added to the debate about assuring quality. Analysis of the interview transcripts revealed that respondents firmly conceptualise AoL as only one aspect of the higher education quality agenda. This research study has identified the quality concerns of Associate Deans outside any process of AoL. This research contributes evidence to support views of AoL that were previously theorised rather than empirically researched. Gora (2010) provided a critical and sardonic metaphor in response to his self-posed question, “But what is this mysterious entity called ‘quality’?” His metaphorical response was, “on closer inspection this grand assurance exercise turns out to be a four lane highway leading to a cowpat” (p. 77). Gora’s description of the elements of quality assurance echoes a list of key themes that emerged in the research respondents’ interviews about education evaluation. “Much of it boils down to a calibration of publications, grant acquisitions, information systems, qualification and program accreditation, teaching performance and learning outcomes” (p. 77).

The salience of multi-faceted quality review for respondents is both affirming and symptomatic of the recent activity in the Australian higher education sector. The Department of Education, Employment and Workplace Relations website which hosts the TEQSA website includes documents describing five types of standards. Each of these is a different component of higher education quality assurance. The *teaching and learning standards* (Department of Education, Employment and Workplace Relations, 2011) are the category most in keeping with the concept of AoL. These standards are in the process of consultation and development. Discussion questions posed to stakeholders include such text as, “It is proposed that teaching standards and learning standards are conceptually distinct and therefore require consideration as separate sub-domains for TEQSA quality assurance and regulatory activities. Are there any problems with creating two sub-domains of this kind?” (p. 7). The *provider standards* (Department of Education, Employment and Workplace Relations, 2011) include such headings as financial viability and safeguards, corporate and academic governance, and management and human resources. The *qualification standards* (Department of Education, Employment and Workplace Relations, 2011) address such elements as “articulation, recognition of prior learning and credit arrangements” and that “certification documentation issued is accurate and protected against fraudulent use” (p. 2). To date, the only notice regarding *information standards* (Department of Education, Employment and Workplace Relations, 2011) on the TEQSA website is that “they are intended to act as a guide to information-sharing between providers and their key stakeholders, especially students.” Similarly, the TEQSA website states that the *research standards* are at the “initial stage of development” and that they will likely link to the “Australian Code for the Responsible Conduct of Research.” The further development and communication of the standards will no doubt impact the ways in which Australian universities collect, document and report data to

assure learning and other components of quality assurance.

Conclusion

Assurance of Learning (AoL) as a means of higher education evaluation is a process of collecting, mapping, compiling, reporting, and processing data about learning, teaching, curriculum, and pedagogy. Some of the key components included in AoL are learning outcomes, assessment, graduate attributes, enrolment statistics, and completion rates. AoL, in the context of education evaluation, is conducted for the dual purposes of accountability and continuous improvement. Various combinations of processes are used to inform on quality, such as rubrics, curriculum maps, sampling, staff workshops, benchmarking, teams, and data collection tools as part of the process of assuring the learning. Notably, AoL leaders are hard-pressed to articulate a comprehensive depiction of AoL that is transparent and significant for three reasons. First, there is no agreed-upon definition or guidelines for higher education AoL. Second, while there is a clear sector-wide message from Australian national higher education authorities that education evaluation, including the development of performance measurement instruments, is expected, and that it is incumbent upon universities to collect and report quality assurance data, no clear definitions, process, standards and guidelines have been established and communicated. Third, the leaders are inconsistent on the topic of what is driving their focus on assurance of learning.

Despite the levels of uncertainty in evaluating quality in higher education, Australian universities are not in holding-mode waiting for further instruction. There is a nation-wide commitment to quality improvement and widespread agreement that AoL offers a means to that end. Respondents in this study agreed that audit and accreditation provide leverage for challenging processes, but are not always the driver. Universities are also committed to closing-the-loop by ensuring continuous improvement of programs and the next steps are to cover the change processes and quality improvements revealed through the analysis of data collected in the name of quality assurance.

There are two limitations in this study. The first is that the participant pool was drawn only from Australian business schools. Business schools have a particular slant on AoL because it is a defined component of AACSB accreditation. This limitation will be addressed by further research in the next phase of the research project that includes expanding the participation to disciplines beyond business. A further limitation is that the data reported in this paper was collected through a retrospective self-reporting survey. The concern is that the respondents recall and self-select at the time of the interview and some aspects may be overlooked or forgotten. Research on AoL in Australian universities within the described research project is ongoing and involves further surveys and collecting quality assurance artefacts including tools, frameworks and rubrics to address limitations and ensure rigour.

References

AACSB. (2007). *AACSB assurance of learning standards: An interpretation.* Tampa, Florida: AACSB International Accreditation Coordinating Committee; AACSB International Accreditation Quality Committee.

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). *Trends in global higher education: Tracking an academic revolution: A report prepared for the UNESCO 2009 world conference on higher education.* Paris: UNESCO Division of Higher Education.

Biggs, J., & Tang, C. (2007). Setting the stage for effective teaching. In J. Biggs, & C. Tang

(Eds.), *Teaching for quality learning at university* (3rd ed., pp. 31-59). England & NY: Society for Research into Higher Education & Open University Press.

Bradley, D., Noonan, P., Nugent, H., & Scales, B. (2008). *Review of Australian higher education: Final report.* Canberra, Australia: Government of Australia.

Department of Education, Employment and Workplace Relations (2010). *Government to introduce 'My University' website.* Retrieved February 17th, 2011, from http://ministers.deewr.gov.au/gillard/government-introduce-%E2%80%98my-university%E2%80%99-website

Department of Education, Employment and Workplace Relations (2011). *Assessment of generic skills.* Canberra, Australia: Government of Australia.

Department of Education, Employment and Workplace Relations. (2011). *Development of performance measurement instruments in higher education.* Canberra, Australia: Government of Australia.

Department of Education, Employment and Workplace Relations. (2011). *Review of the Australian graduate survey (AGS).* Canberra, Australia: Government of Australia.

Devlin, M. (2010, March 3rd). My university website. *Deakin Speaking,* Retrieved February 17th, 2011 from http://www.deakin.edu.au/deakin-speaking/node/94

Gardiner, L. R., Corbitt, G., & Adams, S. J. (2010). Program assessment: Getting to a practical how-to model. *Journal of Education for Business, 85*, 139-144. doi:10.1080/08832320903258576

Gora, J. (2010). Watch out! Here comes the TEQSA juggernaut. *The Australian Universities' Review, 52*(2), 76-78.

Henderson-Smart, C., Winning, T., Gerzina, T., King, S., & Hyde, S. (2006). Benchmarking learning and teaching: Developing a method. *Quality Assurance in Education, 14*(2), 143-155. doi:10.1108/0968-4880610662024

Kai, J. (2009). A critical analysis of accountability in higher education. *Chinese Education and Society, 42*(2), 39-51. doi:10.2753/CED1061-1932420204

Kift, S. M., Butler, D. A., Field, R. M., McNamara, J., Brown, C., & Gamble, N. (2010). Conceptualising a capstone experience for law students. *Australasian Law Teachers Association 65th Annual Conference, 4-7 July 2010,* University of Auckland, Auckland.

Koppang, A. (2004). Curriculum mapping: Building collaboration and communication. *Intervention in School and Clinic, 39*(3), 154-161. doi:10.1177/10534512040390030401

Marginson, S. (2009). The limits of market reform in higher education. Paper presented to the *Research Institute for Higher Education (RIHE),* Hiroshima University.

Martell, K., & Caldron, T. G. (2009). Assessment in business school: What is it? Where we are, where we need to go now. *Assessment Seminar July 6-7 2009,* Sydney, Australia.

Martell, K. (2007). Assessing student learning: Are business schools making the grade? *Journal of Education for Business, 82*(4), 189-196. Retrieved February 17th, 2011, from http://zl9eq5lq7v.scholar.serialssolutions.com/?sid=google&auinit=K&aulast=Martell&atitle=Assessing+student+learning:+Are+business+schools+making+the+grade%3F&id=doi:10.3200/JOEB.82.4.189-195&title=Journal+of+education+for+business&volume=82&issue=4&date=2007&spage=189&issn=0883-2323;

Maxwell, S. (2010, Good, better, best: The use of rubrics for graded assessment. *Teacher: The National Education Magazine, 2010*(212), 34-36. Retrieved February 17th, 2011, from http://research.acer.edu.au/teacher/vol2010/iss212/14/

Menassa, E., Safi, M, & Chaar, B. (2009). A pilot study of university professors and students'

perception regarding accreditation of business schools in Lebanon. *International Journal of Business Research, 9*(2), 129-143.

Radloff, A., Coates, H., James, R., & Kerri-Lee, K. (2001). *University experience survey: Report on the development of the university experience survey.* Canberra, Australia: Government of Australia.

Smith, J. E., Meijer, G., & Kielly-Coleman, N. (2010). Assurance of learning: The role of work integrated learning and industry partners. In M. Campbell (Ed.), *Work integrated learning: Responding to challenges* (pp. 409-419). Perth, WA: Australian Collaborative Education Network (ACEN) Incorporated, Curtin University of Technology.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment time to save grading time, convey effective feedback and promote student learning.* Virginia, Canada: Stylus Publishing L.L.C.

Stolz, I., Hendel, D. D., & Horn, A. S. (2010). Ranking of rankings: Benchmarking twenty-five higher education ranking systems in europe. *Higher Education, 60*, 507-528. doi:10.1007/s10734-010-9312-z

Templin, T.J. & Blankenship, B. T. (2007). Accreditation in kinesiology: The process, criticism and controversy, and the future. *Quest, 59*, 143-153.

Thorpe, C. (2011, May 12). Cautious welcome for my university website. *ABC News,* Retrieved February 17th, 2011, from http://www.abc.net.au/news/2011-05-12/cautious-welcome-for-my-university-website/2711684;

Evaluating Finnish teacher educators as entrepreneurship educators

Heikki Hannula, HAMK University of Applied Sciences, Professional Teacher Education Unit, Finland
Elena Ruskovaara, Lappeenranta University of Technology, Finland
Jaana Seikkula-Leino, University of Turku / Lappeenranta University of Technology, Finland
Anne Tiikkala, University of Turku, Finland

Abstract

Entrepreneurship education has gained more important role in education systems. We here present and analyse data concerning entrepreneurship education practices taking place in Finnish teacher education. We see teacher educators in key role in promoting entrepreneurship education. Our main research questions are: what are entrepreneurship education activities of teacher educators in Finland and could our measurement tool be used to evaluate it?

As a method we use the Measurement Tool for Entrepreneurship Education developed in Lappeenranta University of Technology to evaluate the entrepreneurship education activities of Finnish teacher educators. It is a self assessment tool that gives information both for the respondents themselves and the data collectors in Lappeenranta. The main target group is Finnish teacher educators working both with future all round teachers and vocational or professional teachers. We here present quantitative data collected in November and December 2011.

The results direct the future activities how to develop teacher education and how to develop the measurement tool to give us even more high quality results in the future. Teachers' entrepreneurship education activities will be measured in the course of our ongoing projects and we are seeking information about possible changes

Keywords: entrepreneurship education; evaluation; teacher education

Introduction

The advancement of entrepreneurship is becoming more and more significant with regards to the national economy and from the entire European perspective. It is seen as a basis for developing the social and economic well being of the European Union (EU). Therefore, the contribution of entrepreneurship and entrepreneurship education has been greatly acknowledged at the EU level and it has been seen as a major element determining accountability in education (see for example Commission of the European Communities 2006; European Commission, Enterprise and Industry 2010; Europe 2020 Strategy.) This is even more evident today after serious economical crises in several member states of EU.

In the European Union, one of the latest core aspects is to develop entrepreneurship education in teacher education. However, entrepreneurship education still seems to be, across EU countries, quite uncommon in initial teacher training. (GHK 2011) Moreover, from this EU perspective, teacher education is yet to be fully incorporated into most national strategies and it is not a part of teachers´ continuing professional development. (ETF, 2011) In Finland, however, the *Guidelines for Entrepreneurship Education* of the Ministry of Education and Culture (2009) has been developed to indicate how entrepreneurship education should be reinforced within teacher education.

Internationally, there is hardly any evaluation research available on the entrepreneurship

education in teacher education. There is, however, research on evaluating entrepreneurship programs and their impact upon new venture creation (e.g. Hytti & O'Gorman 2004; Fayolle 2005; Barr et al. 2009; Boni et al. 2009). The Organization for Economic Co-operation and Development [OECD] report *Evaluation of Programmes Concerning Education for Entrepreneurship* states that there is no useful system or method for evaluating courses (OECD 2009, p. 22). Also, higher education is lacking this kind of evaluation system.

Entrepreneurship Education

The research of entrepreneurship education builds largely on the conceptual understanding of entrepreneurship and learning. As Gibb (2005) has stated, entrepreneurship education is about learning for entrepreneurship, learning about entrepreneurship and learning through entrepreneurship. Therefore, entrepreneurship education should be considered both as a method of learning, as well as a content of learning (see Remes, 2003).

The learning outcomes of entrepreneurship education include several layers. Entrepreneurship education introduces entrepreneurship as a career choice, it supports the entrepreneurial way of seeing and doing things and it characterizes a way of teaching and learning (Steyaert & Katz, 2004; Berglund & Johansson, 2007). Entrepreneurship education for younger students has been suggested to concern more learning the spirit and ways of doing and seeing than about business activity. The aim is that students could take more responsibility for themselves and their learning (for example Remes, 2001; 2004, Tiikkala et al. 2010) In other words, entrepreneurship education should support the students' feeling of their *internal locus of control*. As a learning outcome, the students would also try more persistently to achieve their goals, to be creative, to discover existing opportunities and in general to cope with the complicated society. This education involves the development of attitudes, behaviors, skills and attributes applied individually and/or collectively to help individuals and organizations of all kinds to create, cope with and enjoy change and innovation. (Frank, 2007) This process involves higher levels of uncertainty and complexity as a means of achieving personal fulfillment and organizational effectiveness.

While the learning outcomes of entrepreneurship education have been under careful research, the viewpoint of teaching has seemingly been underdeveloped. According to Kyrö (1997), entrepreneurship education deals with three main components: 1) self-oriented, 2) internal and 3) external entrepreneurship. Self-orientated entrepreneurship refers to an individual's self-oriented behavior and serves as the basis for developing internal and external entrepreneurship (Remes, 2004). Internal entrepreneurship deals with entrepreneurial and enterprising behavior. External entrepreneurship is about doing business (Ristimäki, 2004). Within fairly young students, self-oriented entrepreneurship is emphasized (Remes, 2001). As a consequence, the focus is not only on developing factors related to motivation, self-awareness and creativity (for example Menzies & Paradi, 2003), and responsibility for learning, but also on cooperation and interaction, which refer to internal entrepreneurship development. In comparison, in the school context, external entrepreneurship education is about developing innovation (see also Gibb 2005) and business ideas, as well as strengthening cooperation between schools and the world of work, including such activities as work experience and study trips.

The YVI project is Finnish, nation-wide, multi-science development and research project, and it aims to develop entrepreneurship education in both vocational and general teacher education. The project is financed by the European Social Fund (ESF) and the Finnish National Board of Education. Both projects co-operate very closely to each other. The purpose of the project is to develop a dynamic, virtual learning environment, aimed at the developers of entrepreneurship education. The learning environment helps teachers to

improve their skills in planning, implementing and evaluating entrepreneurship education. Additional goals are to increase the collaboration among entrepreneurship education developers, to improve the knowledge of entrepreneurship education and to help teacher educators to improve their pedagogical skills of entrepreneurship education.

Teacher Educators and Education in Finland and in European Union

In this study by teacher educators we mean people who are teachers in Finnish academic teacher education units in universities, teacher training schools and vocational or professional teacher education units in applied sciences.

The results obtained in Finnish education have aroused interest also in our teacher education (cf. Kupiainen, Hautamäki & Karjalainen 2009). Finnish teachers have a graduate degree. This means that Finnish teacher education is inclined towards a research-oriented education, in which all parts of the teaching degree have an integrated research aspect. The research orientation steps in at a fairly early stage of the education and theoretical content is woven into practice at all stages of the teaching studies. Other models that exist in teacher education are the school-based model, reminiscent in some ways of the apprenticeship model; the case-specific or practical model, whose approach is problem-oriented; and the experiential learning model, which emphasizes the learner's personal views regarding teaching (Krokfors et al. 2009). Undoubtedly, Finnish teacher education contains aspects of each of these models, but compared to other countries, Finland is characterized by its research-orientation, which means that students studying to become teachers learn to justify their decisions and actions based on both experience and theory.

Teacher education units linked to universities usually offer academic teacher education. Teacher training schools are an essential element in their networks. The teacher training schools coordinate and develop research-oriented, directed teaching practice, as well as further education (cf. Teacher Training School Strategy for 2020). Academic teacher education covers primary and lower secondary education, as well as upper secondary education. Furthermore, academic teacher education units offer opportunities for completing a special education teacher or special class teacher qualification and a guidance counselor qualification. Most academic teacher education students complete a Master's degree. Those aiming to teach primary school complete the class teacher's qualification. Those aiming to teach lower and upper secondary schools usually complete a subject teacher's qualification. However, lately there have been efforts to develop teachers' opportunities to teach widely across educational levels. (cf. Trade Union of Education in Finland, 2010).

In Finland, professional and vocational teachers work mainly at higher education institutions, vocational colleges, vocational adult education institutions and public sector organizations. They are pedagogically educated in professional or vocational units of universities of applied sciences. In practice they study in multiform programs, including both contact days and distance learning. It is also possible to study only through distance learning.

The vocational or professional teacher education program (60 ECTS) includes basic pedagogical theory and professional or vocational training; professional or vocational pedagogy; in-service teacher training in the teacher education student's respective organization, mostly educational institutions and thesis research or development work. To be accepted to the professional or vocational teacher education the teacher education students need mainly a bachelor´s degree at the minimum, and three to five years of relevant work experience, depending on the professional field. Students represent fields such as high technology and engineering, social sciences, communications, and business and industry.

Evaluation

The evaluation of education means the definition of the value or merit of the background, of processes and of the results of education. It takes place when comparing processes or results with prerequisites and targets. Further, the evaluation must determine the value of the matter examined. (Atjonen, 2007) Evaluation aims to collect information, produce feedback, go forward and build the future (Linnakylä & Atjonen, 2008). In addition, at the right time processed evaluation motivates students to study and aims to promote learning (e.g. Black & Wiliam, 1998; Brooks, 2002; Race, Brown & Smith, 2005; Pickford & Brown, 2006; Atjonen, 2007).

Coombes (1992) considers that power and accountability in evaluation are major factors which guide the ethics of evaluation. In terms of accountability, we must trust that the evaluator has the competence and qualifications that are required for the task and that allow the evaluator to be fair and responsible. In addition, it must be borne in mind that the evaluation concerns the values of people who take part in it. Atjonen (2007), Korkeakoski (2008), House (1980) and House & Howe (1999) stress that the aims and contents of education and schooling are determined on the grounds of what is valuable. The school subjects or other themes are planned, implemented and evaluated on the grounds of relevant values.

According to Pickford and Brown (2006), evaluation should be a genuine part of learning. Usually evaluations are conducted at the end of the learning process. Pickford and Brown clarify that the end-results of learning should be considered first, after which the content and the aims of learning should be derived from these end-results. The basic questions for evaluation are: What are we going to evaluate? Why do we evaluate? How do we evaluate? Who evaluates? When do we evaluate? In line, with Pickford and Brown (2006), Seikkula-Leino et al. (2010) stress in their study concerning the evaluation of teachers that the evaluation should be based on the teacher's reflection process. They emphasize that their evaluation ought to involve the development of vision, motivation, understanding and practice in order to empower the teacher's learning through evaluation.

Methods

Our empirical study builds on survey data including responses from 51 teachers working in all-round and professional/vocational teacher education. They work in universities, teacher training schools and universities of applied sciences in the Finnish language. In Finland, there are also Swedish language institutions but they are not included in this study. By the deadline we had received 51 respondents. Thirty-one of them represented all-round institutions and 20 professional and vocational institutions. Two respondents were working in leading positions and the rest were teachers or teacher educators or trainers. There were 17 male and 34 female respondents, representing nine different institutions from north to south. The data was collected during November and December 2011 through a web-based questionnaire. The questionnaire has originally been developed in the project of Measurement Tool for Entrepreneurship Education in Lappeenranta University of Technology. The data was collected for testing the web-based questionnaire itself as part of YVI-project.

In our measurement tool the statements were mostly presented in scale 0-4. 0=never and 4=continuously. Some of the statements were presented in scale yes/no. In addition we tried to find out some differences between professional and all round teacher educators. If we saw a need we used t-test to determine if the difference is significant.

The questionnaire includes four main parts: 1) teaching methods, 2) Development of the students' entrepreneurial behavior, 3) planning and implementation of teaching and teacher

training and 4) enterprising operational culture in organization.

Results

To the respondents were presented different teaching methods and they evaluated how often they use the method in their teaching. In table 1 there are presented only the most interesting results. We can see, that teacher educators use very often the methods based on problem solving and collaborative learning. But it is a little surprise that for example pedagogical drama is mot used so often.

There are some differences between the groups of professional teacher educators and all round teacher educators. The difference is very significant (t-test<0.01) in co-operative learning. The other founded significances were significant (<0.05) or almost significant (<0.1). All the differences show that professional teacher educators use more often these methods than all round teacher educators.

Table 1 Learner based teaching methods and working styles

Statement	Mean (all)	Mean (prof.)	Mean (all round)	t-test
Learning by doing	3.46			
Problem based learning	3.22			
Co-operative learning	3.39	3.70	3.19	0.009
Team learning	3.02			
Experimental learning	3.39			
Peer learning	3.31	3.60	3.13	0.038
Pedagogical drama	1.54			
Inquiry-based learning	2.98	3.20	2.84	0.088
Learning diary/portfolio	3.46	3.05	2.55	0.070

Table 2 shows us that teacher educators do not discuss with their student about their dreams but much more they discuss about the future plans. But they direct their students to take responsibility, self-regulation and many other things that can be connected with entrepreneurship and entrepreneurial behavior. The only statistically significant difference between the two groups can be seen in supporting the students for self-directivity.

Table 2 Development of the students' entrepreneurial behavior

Statement	Mean (all)	Mean (prof.)	Mean (all round)	t-test
Discuss with the students about their dreams	2.35			
Directing the students to take responsibility	3.84			
Supporting the students for self-directivity	3.69	3.85	3.58	0.043
Supporting self-regulated decision making	3.71			
Developed students trust on own abilities	3.71			
Directed student to set their own goals	3.78			

In the table 3 we can see, that the teacher educators have changed their teaching and guidance plans flexibly, created new activities in their job (3.60) and worked in an entrepreneurial way. These results show that teacher educators believe that they behave entrepreneurially and according to the objectives of entrepreneurship education when they plan and implement their teaching.

In contrast, teacher educators have rather seldom familiarized their own students to entrepreneurial learning environments and methods, such as practice enterprise or Young Enterprise, used practical case studies of entrepreneurship education, had an entrepreneurship lesson or moment, had an entrepreneurship education course, and encouraged their students to develop their own enterprise or practice enterprise.

At an almost significant level the all round teacher educators ”keep their plans as well as possible” even more seldom than the professional/vocational teacher educators. There is a very significant difference in guiding the teacher education students to networking with working or business life.

Significant or almost significant differences could be found in the following statements (how regularly I have...): used practical case studies of entrepreneurship education, discussed with the students about entrepreneurship and brought up working or business life in teaching.

Table 3 Planning and implementation of teaching and teacher training

Statement	Mean (all)	Mean (prof.)	Mean (all round)	t-test
Have changed his/her plans flexibly	3.36			
Keep plans as well as possible	1.35	1.60	1.19	0.065
Creation of new activities in job	3.60			
Worked in an entrepreneurial way	3.08			
Familiarize students to entrepreneurial environments	1.82			
Usage of practical case studies of EE	1.80	2.16	1.58	0.07
Have had entrepreneurship education lessons	1.64			
Have had entrepreneurship education course	1.24			
Encouraged students to develop an (practice) enterprise	1.33			
Guiding the students to network with business life	2.46	3.37	1.90	0.0001
Discussed with students about entrepreneurship education	2.06	2.37	1.87	0.08
Brought up business life in teaching	2.74	3.21	2.45	0.01

Table 4 shows us both entrepreneurial acting but also not so entrepreneurial ways of ways of building entrepreneurial culture in their organizations. For example they have attempted to create the mistakes accepting and expertise sharing culture. But less they have thought entrepreneurship education itself.

Table 4 Enterprising operational culture

Statement	Mean (all)
Influence in community to accept mistakes	3.08
Encouraged organization to use know-how of different specialists	3.04
Developed culture of shared expertise	3.37
Developed culture of free brainstorming	3.31
Created something new with students	3.06
Evaluated my EE-plans in practice	1.90
Evaluated EE results with students	1.71
Assessed how EE can be seen in learning outcomes	1.84
Assessed my own development as an entrepreneurship educator	1.90
Strengthened understanding of EE in my working community	1.78
Developed systematically EE-activities in my organization	1.90

Discussion

The basic idea of this study is to promote entrepreneurship education by supporting the teacher educators to teach entrepreneurship education both as contents and methods. And in this study we wanted to measure and evaluate the entrepreneurship education activities of the Finnish teacher educators. We succeeded to find out some interesting points. One is that the basis of promoting entrepreneurship education is quite good. Teacher educators use pedagogical and didactical tools that are also used in entrepreneurship education. On the other hand there could be noticed that the claims including the word entrepreneurship were evaluated with lower values than others.

The other objective was to develop the measurement tool because we want to continue this kind of evaluation also in the future. Some claims have to be developed because they might be good for teachers. But because of the different kind of nature of work of teacher educators they can't give always the valid responses. In addition there are also some claims to be deleted.

The questionnaire was distributed to the teacher educators by e-mail offering a direct link to the questionnaire. The amount of respondents was much lower than expected. The response time was increased and the managers of teacher education institutes were asked to support their teacher educators to answer. In the future, more consideration should be put into how to increase the response rate. In addition, it can be supposed that the respondents were mainly people who have a positive or neutral perspective on entrepreneurship. In order to draw reliable conclusions it should be important to get responses from every kind of teacher educator.

However, we can see many good results in the advancement of entrepreneurship education in teacher education: teacher educators seem to use many good pedagogical models and methods that also supporting entrepreneurship education. Although they did not use these methods with entrepreneurship education in mind, it can be thought that this gives a good basis for the development of entrepreneurship education as well. Teacher educators strongly support their students' self-dependence, self-confidence, responsibility, goal-orientation and self-directivity. In addition, they support their students to learn from their mistakes and to handle their successes. All these are widely connected with entrepreneurial learning and entrepreneurship education.

Our data indicates that there are a number of areas for development and raises many issues for future research in this area. The next stage is to test the question battery on a larger cohort, and at the same time gain a broader understanding of Finnish teacher training and the entrepreneurship education connected with it. Furthermore, it would be important and interesting to attempt to find ways for teacher trainers to take better advantage of different entrepreneurship education actors in their own work. However, it would also be important to increase determined efforts to bring entrepreneurship education into teacher educators' work more often.

References

Atjonen, P. (2007). *Hyvä, paha arviointi.* Jyväskylä: Gummerus.

Barr, S., T. Baker & Markham, S. (2009). "Bridging the Valley of Death: Lessons Learned from 14 Years of Commercialization of Technology Education". *Academy of Management Learning & Education*, 8(3): 370-388.

Berglund, K., & Johansson, A. W. (2007). "Entrepreneurship, Discourses and Conscientization in Processes of Regional Development". *Entrepreneurship & Regional Development*, 19:

499-525.
Black, P. & Wiliam, D. (1998). *Assessment and classroom learning.* Assessment in Education, 5, 7-78.
Boni, A.A., L.R. Weingart & Evenson, S. (2009). "Innovation in an Academic Setting: Designing and Leading a Business Through Market-Focused, Interdisciplinary Teams". *Academy of Management Learning & Education*, 8(3): 407-417.
Brooks, M. (2002). *Assessment in secondary schools. A new teacher's guide to monitoring, assessment, recordings, reporting, and accountability*. Philadelphia: Open University.
Coombes, M. (1992). *The ethics of educational evaluation.* Social Alternatives 11, 39-42.
European Training Foundation (ETF). (2011). *Working together learning for life. Teacher education for entrepreneurship: Towards a development agenda.* A report and evaluation of the pilot action on training teachers for entrepreneurship education initiated by the European Commission. ECORYS.
Europe 2020. *A European strategy for smart, sustainable and inclusive growth.* European Commission (2010). COM (2010) 2020. Communication from the Commission.
European Commission, Enterprise and Industry. (2010). *Towards Greater Cooperation and Coherence in Entrepreneurship Education.*[pdf] Report and Evaluation of the Pilot Action High Level Reflection Panels on Entrepreneurship Education Initiated by DG Enterprise and Industry and DG Education and Culture. Available at: http://ec.europa.eu/enterprise/policies/sme/promoting-entrepreneurship/education-training-entrepreneurship/reflection-panels/files/entr_education_panel_en.pdf [Accessed on 15.3.2010].
European Commission, Entrepreneurship Education in Europe. (2006). *Fostering entrepreneurial mindsets through education and learning.* [pdf] Oslo, 26.-27.10.2006. Available at: http://ec.europa.eu/enterprise/policies/sme/files/support_measures/training_education/doc/oslo_agenda_final_en.pdf [Accessed on 9.12.2011].
Fayolle, A. (2005). Evaluation of entrepreneurship education: behaviour performing or intention increasing? *International Journal of Entrepreneurship and Small Business,* Vol. 2, No. 1, pp. 89-98.
Frank, A. I. (2007). "Entrepreneurship and Enterprise Skills: A Missing Element of Planning Education". *Planning, Practice & Research,* 22(4): 635-648.
GHK. (2011). *Mapping of teachers' preparation for entrepreneurship education.* Order 129. Framework Contract No EAC 19/06. DG EDUCATION AND CULTURE. Final Report.
Gibb, A. (2005). "The future of entrepreneurship education – Determining the basis for coherent policy and practice?". in P. Kyrö and C. Carrier (eds.) *The dynamics of learning entrepreneurship in a cross-cultural university context.* Entrepreneurship Education Series 2/2005, Hämeenlinna: University of Tampere, Research Centre for Vocational and Professional Education, pp. 44-67.
House, E. R. & Howe, K. R. (1999). *Values in evaluation and social research.* London: Sage.
House, E. (1980). *Evaluating with validity.* London: Sage.
Hytti, U. & O'Gorman, C. (2004). "What is "enterprise education"? An analysis of the objectives and methods of enterprise education programmes in four European countries". *Education + Training,* 46(1): 11-23.
Korkeakoski, E., (2008). "Arvotietoisuus, teorialähtöisyys ja vaikuttavuus arviointimenetelmien kehittämisessä". In E. Korkeakoski and H. Silvennoinen (eds.) *Avaimia koulutuksen arvioinnin kehittämiseen.* Jyväskylä: Koulutuksen arviointineuvoston julkaisuja 31, pp. 201-215.
Krokfors L., Kynäslahti H., Stenberg K., Toom A., Maaranen K., Jyrhämä R., Byman, R. & Kansanen, P. (2009). *Opettajia muuttuvaan kouluun –Tutkimuspainotteisen*

opettajankoulutuksen arviointia. Kasvatus, 40, 206-219, 285.
Kupiainen, S., Hautamäki, J. & Karjalainen, T. (2009) *The Finnish Education System and Pisa.* Ministry of Education Publications, Finland 2009:46.
Kyrö, P. (1997). *Yrittäjyyden muodot ja tehtävä ajan murroksessa.* Jyväskylä Studies in Computer Science. Economics and Statistics 38. Jyväskylä: University of Jyväskylä, Finland.
Linnakylä, P. & Atjonen, P. (2008). ”Arviointi, tutkimus ja arviointitutkimus koulutuksen tietotuotannossa”. In E. Korkeakoski and H. Silvennoinen (eds.) *Avaimia koulutuksen arvioinnin kehittämiseen.* Jyväskylä: Koulutuksen arviointineuvoston julkaisuja 31, pp. 79-98.
Menzies, T.V. & Paradi, J.C. (2003) “Entrepreneurship education and engineering students: Career path and business performance”. *The International Journal of Entrepreneurship and Innovation*, 4(2): 121-132.
Ministry of Education and Culture. (2009). *Guidelines for Entrepreneurship Education,* Publications of the Ministry of Education 2009:7. Helsinki: Yliopistopaino.
Organziation for Economic Co-operation and Development (OECD). (2009). *Evaluation of Programmes Concerning Education for Entrepreneurship.* OECD Report.
Pickford, R. & S. Brown (2006). *Assessing skills and practice.* London: Routledge.
Race, P., Brown, S. & Smith, B. (2005). *500 tips on assessment (2nd editio*n). London: Falmer.
Remes, L. (2001) “Yrittäjyyskasvatus pedagogisessa toimintatehtävässä”. *Kasvatus,* 32(4): 168-181.
Remes, L. (2003) *Yrittäjyyskasvatuksen kolme diskurssia.* Jyväskylän yliopisto. Jyväskylä: Jyväskylä University Printing House.
Remes, L. (2004). ”Yrittäjyys”. In M-L Loukola (ed.) *Aihekokonaisuudet perusopetuksen opetussuunnitelmassa.* Jyväskylä: Gummerus, pp. 89-90.
Ristimäki, Kari, 2004, *Yrittäjyyskasvatus.* Hamina: Oy Kotkan Kirjapaino Ab.
Seikkula-Leino, J., Ruskovaara, E., Ikävalko, M., Mattila, J. & Rytkölä, T. (2010) *Promoting entrepreneurship education: the role of the teacher?,* Education and Training, 52 (2), 117-127.
Steyaert, C. & Katz, J. (2004). *Reclaiming the space of entrepreneurship in society: geographical, discursive and social dimensions.* Entrepreneurship & Regional Development 16: 179-196.
Teacher Training School Strategy for 2020. [pdf] Available at: http://www.enorssi.fi/suho-harre/Harjoittelukoulujen_strategia_2020.pdf [Accessed 12.12.2011].
Tiikkala A., Ruskovaara E., Rytkölä, T., Seikkula-Leino, J., & Troberg E. (2010) “Evaluation and Values of Entrepreneurship Education.” Paper presented at *The 7th ESU Conference on Entrepreneurship 2010*. Conference proceedings of ESU Conference 2010, pp.45-67. Tartu, 2010.
Trade Union of Education in Finland. (2010). *Opettajankoulutus Suomessa.* pdf] Available at: http://www.oaj.fi/pls/portal/docs/PAGE/OAJ_INTERNET/01FI/05TIEDOTTEET/03JULKAISUT/OAJ_OPETTKOULUTUS_10_WEB.PDF [Accessed 12.12.2011].

This paper is produced within the YVI project. YVI (2010-2013) is a Finnish nation-wide multi-science development and research project and it aims at developing entrepreneurship education for teacher education, both vocational and general. The project is financed by ESF, the Finnish National Board of Education.

V

Evaluation of Academic Staff Performance

Examples of Academic Faculty Evaluation Systems from the Czech Republic and Finland

Mikael Collan: LUT School of Business, Lappeenranta University of Technology , Lappeenranta, Finland
Jan Stoklasa, Jana Talasova; Department of Mathematical Analysis and Applications of Mathematics, Palacky University, Olomouc, the Czech Republic

Abstract

Evaluation of the performance of academic faculty is a yearly recurring task at almost all universities, and it is an issue that is getting more and more attention as many universities are required to report on their efficiency to financing bodies. In this paper we present three instances of academic faculty performance measurement systems, two of the three examples are from Finland and one from the Czech Republic. We shortly discuss the different systems and find that they are very different, although they are used for the same purpose: academic faculty performance evaluation.

Keywords: University, Performance, Faculty Evaluation, Support System

Introduction

Development of faculty evaluation systems at universities has become a more pressing issue as the university systems are changing and focus of universities is shifting in many countries (Bana e Costa and Oliveira 2012). As funding of universities becomes more performance based in many countries, see e.g. (Hicks 2012), it is natural that also the faculty is evaluated according to their performance. National systems for evaluating academic faculty have been defined, or are in the process of being defined in many European countries (Minelli, Rebora et al. 2006; Elmore 2008). Academic faculty evaluation is not a simple problem, as evaluation systems should take into consideration the diverse fields of contribution of academics; teaching and research contribution, as well as any contributions within and outside the traditional university sphere. For some background on faculty evaluation systems see, e.g., (Bana e Costa and Oliveira 2012).

There are many approaches to academic faculty evaluation systems, ranging from manually operated simple rule-based approaches (or scorecards) to mathematically advanced multiple criteria evaluation software systems (Uzoka 2008). The purpose of this paper is to shortly describe three selected instances of actual academic faculty evaluation systems that are in place in Finland and in the Czech Republic, to illustrate the diversity of systems in place. The three instances were selected after going through various academic faculty evaluation models currently used in universities in the Czech Republic (Talasova and Stoklasa 2010; Jan Evangelista Purkyne University in Ústí nad Labem 2012; Jana Evangelista Purkyne University 2012; Masaryk University 2012; Tomas Bata University in Zlín 2012), in Finland (Board of the Turku School of Economics 2004; Lappeenranta University of Technology 2011) and elsewhere (University of Technology Sydney 2009; Wayne State University 2009; McGill University 2010; Flinders University 2012). These models were subjected to a detailed analysis regarding their practical and mathematical aspects. The selection of the three instances to be presented in this paper was based on the authors' personal experience and deep understanding of the selected systems.

The three selected systems are shortly illustrated, discussed, and some observations about the three example cases are made. The cases build on the following structure: "short

presentation of the university" in question, presentation of the "evaluation system", "research evaluation", "teaching evaluation", "other academic tasks evaluation", information about the "collection of the data used in the system", and "observations", which is a generic section where issues of particular interest in the system are brought up.

The motivation for this descriptive approach is that there are many different ways of academic faculty evaluation and the diversity of the systems seems to be great; it is not the goal of this paper to perform a thorough comparative analysis between systems, nor is it the goal to suggest best practices – the goal of this research in progress is to present three real world cases and highlight the diversity of faculty evaluation in universities and to observe some relevant issues for academic staff evaluation that are revealed. This paper excludes the detailed presentation of the national context of these systems.

Instance 1: University of Turku (FIN)

Introduction of the university: University of Turku (UTU) is a multidisciplinary scientific university located in the city of Turku on the Southwestern coast of Finland. UTU is one of the largest universities in Finland. In the beginning of 2010 The University of Turku and the Turku School of Economics (TSE) were merged into one university, called the University of Turku.

The evaluation system: The evaluation system in UTU is based on a "performance points system" that originates from the TSE and has been used there since 2005. The system has been taken into use in the whole UTU for the year 2011. The performance points system is a performance measure of research and research related activities. The system cannot be used as a holistic system in the evaluation of academic faculty, that is to say, that teaching or other pedagogical activity is not included in a standardized system.

Teaching evaluation: For teaching performance an ad-hoc system is in use: the employer representative does a heuristic case by case evaluation of teaching performance.

Research evaluation: The system consists of four types of activities that accrue performance points. These are divided into the following categories (A-D): A Publications, B Scientific expert assignments, C International teaching and research mobility, and D Research funding gathered. For each category there are a number of sub categories, for example, for the category A Publications, there are six sub-categories: A1 Monographs and sections in monographs, A2 Refereed journal publications, A3 Conference publications, A4 Theses, A5 Publications in scientific outlets without a referee practice, and A6 Citations (SSCI+SCI). Each one of the sub-categories is further divided into publication types, for which research points are awarded separately and depending on if the publication is international or national. An example of the division into and on the level of publication types is shown in Figure 1, similar division exists for almost all sub-categories.

	Domestic	International
A 2.1. High quality journal article • blind peer review • more than one reviewer • highly rated journal	6	12
Points deducted if elements of quality are missing:		
- no blind review	-1	-1
- only one reviewer	-1	-1
- not a highly rated journal	-1	-1 to -4
Points at minimum	3	6

Figure 1: A part of the University of Turku research point guidelines (w. translation)

Other academic tasks evaluation: Research points do not only accrue from publication activities, but also activity within the scientific community is rewarded by research points. Activity in editorial boards of journals, in organizing conferences, and within leading positions of scientific organizations are taken into consideration and acting as an expert in, for example, committees of the Finnish Academy of Sciences are considered meritorious. Also acting as review for journals and conferences are rewarded by points. Acting as a faculty appointed reviewer or opponent to a dissertation or as an expert with regards to selection of faculty positions, and acting in other expert tasks for, e.g., the Finnish Parliament, the EU Commission, or such, yield points. Also being rewarded a prize for scientific achievements will yield research points. Faculty mobility and collected research funding are also considered.

Collection of evaluation data: The collection of research points is done variably, unit by unit, usually by ad-hoc excel sheets maintained by a nominated person (usually the department secretary) that collects the information from the academic faculty; most often by circulating points submission requests by email. As the points' collection is not standardized and there are colorful practices within different organizations regarding the collection of points, some research merits that would generate research points may possibly end up never being reported. The excel sheets are then sent to "central administration" where the information is aggregated. It is not unusual that the same (research related) information is collected at UTU even three times by different organizations within the university. The reported research points are approved by a research board.

Observations: Some observations that can be made about the system include the fact that impact factor of journals does not have a direct effect on the points given to publications in international journals and that the points that a single international journal article can fetch are at maximum twelve points, while a refereed conference proceedings article in an international conference fetches four points. Also it is notable that citations by others of the researcher's work in articles in SSCI and/or SCI databases will yield three research points each.

The research points are not used directly in determining the remuneration of academic faculty, but a high annual research point accumulation is considered as a clearly positive indication of research activities. The research points are also used in ranking departments (at least in the TSE) and perhaps in the evaluation of faculties (in the new UTU).

The research points are calculated in the same way for all academic faculty members, from junior to senior.

Instance 2: Lappeenranta University of Technology (FIN)

Introduction of the university: Lappeenranta University of Technology (LUT) is a medium size university located in the South-East of Finland, specializing in the nexus of technology and business.

The evaluation system: Academic faculty performance is evaluated with a points system that awards a maximum of 255 points for yearly performance, as an average of a two-year observation period. Points are awarded for teaching and supervision, publication, and raising research funding. Also research visits abroad and pedagogical studies are rewarded.

In the Finnish (national) university remuneration system academic faculty is divided into eleven levels depending on how demanding the task is, according to a nationally agreed upon "demand level chart". In the LUT evaluation system the academic faculty is divided into three sub categories, determined by seniority and the level of their tasks. "Junior level" is the levels 1-4 of the national system, "middle level" is the levels 5-7, and senior level is the levels 8-11. The "junior" level includes, e.g., doctoral students, the middle level includes

academic faculty up to junior professors with fixed term contracts. “Senior” level includes professors with fixed term, or permanent position obtained through a (faculty appointed external) expert assessment of competence.

LUT uses the same system for the evaluation for all categories, but for each of the three categories the amount of awarded points for different types of research merits is different. That is, senior faculty receives a lower amount of points for the same merits than junior faculty.

Teaching evaluation: Evaluation points are given for teaching as an average of two years’ teaching performance and the feedback from the courses is taken into consideration. This is done in a way that the number of credit points given for the courses given during two years are multiplied by three (a weight for “normalizing” the teaching score) and then multiplied by average student feedback (scale 1-5) divided by 3 (“half way” feedback score). That is if there are on average 12 credits per year, and average feedback is 3,5, then the evaluation points received are 12*3*3,5/3=42. If there are more than 200 or more than 400 enrolled students on the course the points are multiplied by 1,2 or 1,5.

Points are given also for supervised completed bachelor degrees (max 10 points) and for supervised completed master’s degrees (max 15 points).

Research evaluation: We take the evaluation points awarded for publication activities as a more detailed example of the system. For the “middle level” International level refereed (journal) publications are rewarded 20*TSC points, so that:

TSC = R*RIM*RCD*RR*RRM*RI, where
TSC = the total score of the publication
R = for a refereed article (1), non refereed (0)
RIM = when the publication outlet has an impact factor (1,25), no impact factor (1)
RCD = for multi-disciplinary publication within the LUT (1,1) otherwise (1). With multidisciplinary is meant the collaboration between authors from different faculties and between different departments within the technical & natural sciences.
RR = for publications with Russian universities (1,15), otherwise (1)
RRM = for publications having to do with Russian markets (1,15), otherwise (1)
RI = for publications with other international universities or organizations (1,1), otherwise (1)

This means that refereed article with impact factor done by authors from two faculties at the LUT in collaboration with a researcher from a Russian and an Italian university about Russian markets will yield a multiplicator of ~2,00 while the minimum multiplicator for a refereed publication is 1,00.

National level refereed (journal) publications and international conference publications yield four points per publication (with a maximum of five conference publications counted). Scientific monographies will fetch at maximum 40 points (depending on the level & quality), book sections in refereed books account for 12 points. The system gives a push to publish in refereed journals with an impact factor. The maximum number of evaluation points that can be awarded for publication activities is capped at 75.

Other academic tasks evaluation: Organizing research financing for the university results in maximum 20 points, the maximum is awarded for 200000€ of annual funding gathered. Pedagogical studies and internationalization (long term staff mobility) and the whole university meeting the set goals (university level goals) also contribute to the evaluation points (together max 60 points).

Collection of evaluation data: The instrument for collecting the information is an excel sheet that automatically calculates the points accumulation after the inputs are fed into the sheet, the sheet includes both the research and the teaching merits. The academic faculty

members are responsible for reporting their own evaluation points and the corresponding reference information etc. to back it up; if they do not report they will be evaluated based on zero points accumulation. There is a very clear incentive to report all merits that accrue evaluation points.

Observations: The points-accumulation is directly connected to academic faculty remuneration level for the period after the evaluation. The actual salary level is set according to discussions with the faculty member's superior, but in case there are no "special circumstances" the points accumulation is a very strong indicator of the remuneration level.

The remuneration matrix is agreed in collective negotiations between the university employers and the Finnish academic employees union AKAVA. The level of personal achievement can account for a maximum of 46% of a staff member's salary.

	Points required for each acchievement level								
Demand level	1	2	3	4	5	6	7	8	9
5	<50	50	60	70	80	90	105	120	135
6	<55	55	70	85	100	115	130	145	165
7	<60	60	80	95	110	130	150	170	200

Teaching- and researchpersonnel

Jobdemand level	Workperformance level								
	1	2	3	4	5	6	7	8	9
1	1 747,01 €	1 816,89 €	1 923,46 €	2 028,28 €	2 133,10 €	2 239,67 €	2 344,49 €	2 451,06 €	2 555,88 €
2	1 921,63 €	1 998,50 €	2 115,71 €	2 231,01 €	2 346,31 €	2 463,53 €	2 578,83 €	2 696,05 €	2 811,34 €
3	2 114,13 €	2 198,70 €	2 327,66 €	2 454,50 €	2 581,35 €	2 710,31 €	2 837,16 €	2 966,12 €	3 092,97 €
4	2 403,36 €	2 499,49 €	2 646,10 €	2 790,30 €	2 934,50 €	3 081,11 €	3 225,31 €	3 371,91 €	3 516,12 €
5	2 787,19 €	2 898,68 €	3 068,70 €	3 235,93 €	3 403,16 €	3 573,18 €	3 740,41 €	3 910,43 €	4 077,66 €
6	3 254,17 €	3 384,34 €	3 582,84 €	3 778,09 €	3 973,34 €	4 171,85 €	4 367,10 €	4 565,60 €	4 760,85 €
7	3 754,51 €	3 904,69 €	4 133,72 €	4 358,99 €	4 584,26 €	4 813,28 €	5 038,55 €	5 267,58 €	5 492,85 €
8	4 543,23 €	4 724,96 €	5 002,10 €	5 274,69 €	5 547,28 €	5 824,42 €	6 097,01 €	6 374,15 €	6 646,75 €
9	5 119,79 €	5 324,58 €	5 636,89 €	5 944,08 €	6 251,26 €	6 563,57 €	6 870,76 €	7 183,07 €	7 490,25 €
10	5 796,43 €	6 028,29 €	6 381,87 €	6 729,66 €	7 077,44 €	7 431,02 €	7 778,81 €	8 132,39 €	8 480,18 €
11	6 703,08 €	6 971,20 €	7 380,09 €	7 782,28 €	8 184,46 €	8 593,35 €	8 995,53 €	9 404,42 €	9 806,61 €

Figure 2: The LUT faculty evaluation result matrix and the direct connection to the remuneration matrix (salary matrix in force from 1.3.2012).

To the best of our knowledge, LUT is the only university in Finland that has a system that directly and systematically connects the academic faculty evaluation result to the remuneration matrix, see figure 2. The system is well documented and information about it is available to the employees in the intranet of the university. There is also a yearly cycle to develop and enhance the system continuously.

Instance 3: Palacky University in Olomouc, Faculty of Science (CZE)

Introduction of the university: Palacky University in Olomouc is one of the oldest universities in Central Europe and with almost 23,000 undergraduate students on eight faculties is one of the largest in the Czech Republic.

The evaluation system: Faculty of Science at the Palacky University has created and uses an information system for the evaluation of academic faculty, the system is called "IS HAP". The evaluation by the system includes almost every aspect of academic faculty activity; performance of each member of the academic faculty is evaluated in pedagogical, research and development (R&D), as well as other areas of activities. The system uses only easy to verify and objective data, and is designed to be easy to work with for the evaluator and the academic faculty being evaluated. The evaluation system is designed to reflect the performance of a given academic faculty member as well as possible. This is achieved by not

just calculating a simple average of partial evaluations in separate areas of activity, but by using intelligent (soft) aggregation. This type of aggregation (by a linguistic fuzzy rule base) is transparent and comprehensible even to a layman as it is described verbally and provides verbal outputs.

The IS HAP system, after several years of its development, provides a sophisticated mathematical background of the evaluation mechanism, yet still well understood by the evaluators, an intuitive on-line interface for gathering input data, and clear way of presentation of evaluation outputs. For more details concerning the development of the system see (Stoklasa, Talasova et al. 2011).

Teaching evaluation: Three areas of activities are taken into consideration for pedagogical performance evaluation: lecturing, student supervision, and work associated with the development of the fields of study. Each particular activity is assigned a score, mainly based on the time used for the task.

Research evaluation: The evaluation of research and development activities is based on the national Czech guidelines for R&D evaluation (Government office of the Czech Republic 2010), but also other activities, like project management, editorial board memberships and the like are included. The most important role in the evaluation of R&D outcomes in the Czech national system is played by journals with non-zero impact factor and issued patents.

The Czech national system uses formula (1) for the evaluation of scientific papers published in journals with non-zero impact factor (IF). This formula uses the rank of the journal in a decreasing sequence of all journals in the current field ordered according to the IF. It is meant to minimize the differences between the evaluations of various scientific fields regarding the IF of journals. Each paper is assigned a certain score (J_{imp}) depending on the journal it was published in. High evaluation is assigned to papers published in the journals with the highest IF in their field, whereas papers published in journals with low IF (relative to the current field) are assigned a lower evaluation (1).

$$J_{imp} = 10 + 295 \cdot factor,$$

where

$$factor = \frac{1 - N}{1 + \frac{N}{0.057}},$$

and

$$N = \frac{P - 1}{P_{\max} - 1}$$

and

N is the normalized rank of the journal in the respective field according to IF,

P is the rank of the journal in a decreasing sequence ordered according to the IF (according to Journal Citation Report)

$P_{\max}$ is the total number of journals in the current field according to Journal Citation Report.

Such evaluation results in a score from 10 to 305 for each paper. Two high impact factor journals “Nature” and “Science” open to all fields are treated separately, each paper published in these journals is awarded 500 evaluation points (the same amount of points is awarded for a European, American or Japanese patent). The score assigned by this method to a paper is then divided among the authors of the paper based on their relative contribution to the paper.

Other academic tasks evaluation: The system takes into account the secretarial and managerial activities performed by each member of the academic faculty (understood as activities that drain time away from and thus reduce the performance in teaching and

research).

Collection of evaluation data: The IS HAP system is currently being used in the form of a web based application, which is accessible through the Internet. Each academic faculty member fills in an on-line form summarizing his/her activities in the previous twelve months. All items in the form are divided into categories and subcategories. A brief help for understanding and inputting any of the items is also available. The current version of the software is able to communicate with the main information system of the Palacký University in Olomouc and to draw information directly from this system. This reduces the time necessary to complete the form. All the filled in forms of a particular department are accessible to the head of the department and all the forms within a faculty are accessible to the dean.

Observations: Even though the mathematical apparatus used to calculate the final evaluation is relatively complicated, the results obtained are presented in way comprehensible even to a layman - that is by using linguistic terms and graphical presentation (see Figure 3). The output of the evaluation (obtained by a fuzzy rule based system) provides a rough piece of information, which still gives the evaluator a sufficient idea concerning the overall performance of the faculty member. The main advantage of using a fuzzy rule based aggregation is that it allows to set-up the shape of the aggregation function used in the evaluation of academic faculty members completely in line with the evaluator's requirements; for example, giving the evaluator the possibility to appreciate excellence achieved in one specific area more than in other areas. The IS HAP system provides an easy to understand overall view of the academic faculty performance, to enable the identification of possible problems and discrepancies. A more thorough analysis of all the evaluation data (partial evaluations and even single items from the forms) is readily available through the system, and allows a deeper understanding of the possible reasons for a given evaluation in detail.

Name	Pedagogical activities	Research	Overall evaluation	Academic functions	Overall workload
Academic staff 1 Professor (1.00)	High (100%) Pedagogical activities 1200.00 a) lecturing 410.00 b) supervising students 610.00 c) development of fields of study. 180.00	Extreme (100%) Research and development 408.17 a) scored results 55.67 b) other results 67.50 c) administration 285.00	Excellent (100%) 2		Extreme (100%) Overall workload 2
Academic staff 2 Assistant (1.00)	Extreme (100%) Pedagogical activities 1750.00 a) lecturing 1580.00 b) supervising students 15.00 c) development of fields of study. 155.00	Standard (43%), High (57%) Research and development 22.00 a) scored results 10.00 b) other results 0 c) administration 12.00	Excellent (100%) 2		Extreme (100%) Overall workload 2
Academic staff 3 Assistant professor (1.00)	Extreme (100%) Pedagogical activities 2273.00 a) lecturing 1705.50 b) supervising students 412.50 c) development of fields of study. 155.00	Very low (38%), Low (62%) Research and development 6.50 a) scored results 0 b) other results 6.50 c) administration 0	Very good (38%), Excellent (62%) 1.81		High (38%), Extreme (62%) Overall workload 1.81

Figure 3. A sample output of IS HAP – overview of evaluations of all academic faculty members.

From the human resource management perspective, the most important part of the evaluation are perhaps the filled-in forms and the partial evaluations in the areas of interest, the overall evaluations are easy to understand and also useful for quick orientation in large numbers of evaluation outputs. The evaluation results are given in a verbal form on all aggregation levels. Both the pedagogical and R&D areas are evaluated by a standard scoring system. There are different standard scores set up for academic faculty members of different seniority. The standard scores are set up to reflect the characteristics of the faculty on which the evaluation

is performed. The evaluation representing a partial evaluation of a member of the academic faculty in a certain evaluated area of activities (PA, RD) is determined as a multiple of the respective standard, or expectation, for the faculty member's position. For better clarity and easier interpretation, these numbers are transformed into verbal evaluation using linguistic scales, see (Talasova and Stoklasa 2010). The IS HAP system can also be used as a HR management tool on all levels of management.

The evaluation by the IS HAP system is not directly connected with academic faculty remuneration – it is intended primarily for human resource management purposes. But the outputs it provides can be easily compared with the remuneration of academic faculty and discrepancies can be identified and eliminated.

Summary and Discussion

The three presented instances of academic faculty evaluation systems are different from each other. The formal part of the University of Turku (UTU) system is a research activities focused scorecard that lists and rewards for a wide number of research related academic tasks, also other than publications, but teaching achievement is not evaluated in a structured way. The Lappeenranta University of Technology (LUT) system is also a scorecard based system that includes teaching and publication activities; for the part of research the system mostly concentrates on publication merits and fund raising. The Palacky University (PU) system, the IS HAP, is a web based software system that considers both research and pedagogical performance and provides output also in a linguistic form that is enhanced with color coding. Of the three systems the PU system is the most advanced by its usability design (a working software), the UTU system takes only research related performance systematically into consideration, and the LUT system includes a direct connection between the performance and the salary level.

	UTU	LUT	PU
Research evaluation	YES	YES	YES
Pedagogical evaluation	NO	YES	YES
Averaging over more than one year	NO	YES	NO
Publication rewarded	YES	YES	YES
Fund raising rewarded	YES	YES	YES
Other academic tasks rewarded	YES	NO	YES
Fuzzy / linguistic approach	NO	NO	YES
Seniority consideration	NO	YES	YES
Software supported	NO	NO	YES
Direct connection to salary	NO	YES	NO

Table 1: Selected characteristics of the evaluation systems

The only one of the systems that uses averages over more than a year is the LUT system; this is interesting as it is often not in the hands of the academic faculty member when, for example an article is published, and thus they have limited capability in affecting their research evaluation for one single year. Publication is rewarded in all systems, this is no surprise, what is less obvious is that all three systems also reward for research fund raising. Other academic tasks, such as participation in editorial boards or other academic positions of trust are rewarded with performance points in the UTU and PU systems. Seniority of the academic faculty, when assessing performance is taken into consideration in the LUT and the

PU systems.

PU system is software supported and gives linguistic outputs in the academic faculty evaluation (approximate reasoning, fuzzy outputs) and also automatically calculates an overall evaluation for the academic faculty members, according to an intelligent aggregation method. The UTU and the PU systems are not primarily intended for the purpose, but can be used as indicators in the determination of academic faculty salary, while the output from the LUT system is more than just indicative; it offers a clear relationship with the performance and an exact salary level from the Finnish national academic faculty remuneration system. Table 1 presents some selected characteristics of these three systems and shows that there are similarities between the systems, but equally there are differences.

Each of these systems represents a real world case of academic faculty performance evaluation and as such offers insights into how university management views academic faculty evaluation. It has been the goal of this paper to present the three cases and shortly discuss them; further work will include putting effort into collecting data about a number of systems in place internationally, to be able to draw conclusions about the different types of systems in place and in the hopes of learning about what in them can be characterized as being “best practice” and thus supporting universities in their efforts to design and perhaps to even harmonize academic faculty performance evaluating systems.

References

Bana e Costa, A. and M. Oliveira (2012). "A multicriteria decision analysis model for faculty evaluation." Omega **40**: 424-436.

Board of the Turku School of Economics (2004). Decision regarding the research points evaluation system. Turku, Turku School of Economics.

Elmore, H. (2008). "Toward objectivity in faculty evaluation." Academe(94): 38-40.

Flinders University (2012). Performance Management http://www.flinders.edu.au/ppmanual/review.html Accessed 14.2.2012.

Government office of the Czech Republic (2010). Methodology for evaluation of research organisations and outcomes of completed programmes http://www.vyzkum.cz/storage/att/5591F655709A8A76C4778C8E14BEB413/Metodika%20hodnoceni%20vysledku%20vyzkumnych%20organizaci%20a%20vysledku%20uko ncenych%20programu%20ve%20zneni%20pro%20rok%202011.pdf Accessed 14.2.2012.

Hicks, D. (2012). "Performance-based university research funding systems." Research Policy **41**: 251-261.

Jan Evangelista Purkyne University in Ústí nad Labem (2012). Academic Staff Evaluation Criteria for Personal Extra Pay Distribution F. o. Environment, http://fzp.ujep.cz/dokumenty/kritosoh.pdf Accessed 14.2.2012.

Jana Evangelista Purkyne University (2012). Academic Staff Evaluation Criteria for Personal Extra Pay Distribution F. o. Environment, http://fzp.ujep.cz/dokumenty/kritosoh.pdf Accessed 14.2.2012.

Lappeenranta University of Technology (2011). Sisäinen ohje (Internal instructions regarding the faculty evaluation), dated 24.3.2011. Lappeenranta, LUT.

Masaryk University (2012). Determination of Criteria for Pedagogical and Other Activities Evaluation. Masaryk University, Faculty of Law, http://is.muni.cz/do/law/ud/predp/archiv_predpisy/4862802/Pokyn_dek._c._7-2009_urceni_kriterii_pro_ hodnoceni_ped._a_j.cin..pdf Accessed 14.2.2012.

McGill University (2010). Academic Performance Evaluation http://www.mcgill.ca/medicine-academic/performance/ Accessed 14.2.2012.

Minelli, E., G. Rebora, et al. (2006). "The Impact of Research and Teaching Evaluation in Universities: Comparing an Italian and a Dutch case." Quality in Higher Education **12**(2): 109-124.

Stoklasa, J., J. Talasova, et al. (2011). "Academic staff performance evaluation – variants of models." Acta Polytechnica Hungarica **8**(3): 91-111.

Talasova, J. and J. Stoklasa (2010). Assessing Academic Staff Performance Using Multiple Criteria Evaluation Models. 2nd International Conference on Applied Operational Research, Turku, Finland, Uniprint.

Tomas Bata University in Zlín (2012). Pedagogical and Creative Activities Evaluation http://web.fai.utb.cz/cs/docs/SD_09_09.pdf Accessed 14.2.2012.

University of Technology Sydney (2009). Performance Management, http://www.hru.uts.edu.au/performance /reviewing/rating.html Accessed 14.2.2012.

Uzoka, F.-M. (2008). "A fuzzy-enhanced multicriteria decision analysis model for evaluating university academics' research output." InformationKnowledgeSystemsManagement **7**: 273-299.

Wayne State University (2009). Guidelines for Evaluation of Academic Staff http://www.aaupaft.org/pdf/AcStaffguidelines_2009-10.pdf Accessed 14.2.2012.

A holistic approach to academic staff performance evaluation – a way to the fuzzy logic based evaluation

Jan Stoklasa (jan.stoklasa@upol.cz), Department of Mathematical Analysis and Applications of Mathematics, Palacky University, Olomouc, the Czech Republic
Pavel Holecek, Jana Talasova; Department of Mathematical Analysis and Applications of Mathematics, Palacky University, Olomouc, the Czech Republic

Abstract

To evaluate the academic staff member performance, many aspects should be taken into account. If the evaluation is seen from the perspective of a university department, at least the research activities, pedagogical activities and administration (management of projects, development of fields of study etc.) need to be considered to ensure the departments welfare. Once a multiple criteria evaluation is needed, the question of partial evaluations proper aggregation arises. In this paper we present the academic staff performance evaluation system (IS HAP) that is currently being used at Palacký University in the Czech Republic. We describe the process of its development and summarize the advantages of linguistic fuzzy modeling for staff evaluation. We show how the knowledge of the desired performance of staff members can be easily and comprehensibly represented by a base of linguistic rules, regardless of its complexity. Possible uses of the presented model in university human resource management and development are also discussed. Outputs and findings from the first run of the described system are also presented.

Keywords: evaluation, fuzzy set, linguistic modeling, multiple criteria evaluation, academic staff, universities

Introduction

The human factor is one the most valuable assets any organization can possess. To fully unlock its potential, it must be managed appropriately. A proper management requires evaluation tools to be in place to identify the strong sides and chances for improvement of staff members. The evaluation mechanism should be devised so that not only assessment, but also motivation and development of the staff are enabled (Matheson, Van Dyk and Millar, 1995); remuneration, promotions and outsourcing are often also based on evaluation results. The evaluation should be performance-focused, comprehensible to the evaluators and the staff and should reflect the organization's (and its staff members') goals. Evaluator biases should be avoided.

Among all institutions and organizations, universities (tertiary education institutions) have a special place thanks to "academic freedom". Academic freedom in the context of the Czech Republic ensures that in these institutions the staff members are free to choose their area of interest and their research focus; in fact they are encouraged to specialize. However, to be considered academic staff members, employees of these institutions are expected to participate in at least two main areas of activities – pedagogical activities (PA) and research and developmental activities (RD). Academic freedom makes the HR management at these institutions a challenging task, considering that the universities strive to fulfill multiple goals. Matějů et al. (2009) define the goals for the Czech Universities: i) to provide quality education; ii) to perform high level research and development and connect it with educational activities; iii) to be beneficial to the economic, social and cultural environment of the region or even broader area. Within the evaluation process the academic freedom of choice (and the

respective outcomes and their quality) have to be combined with the demands of the superiors of the current work unit (the unit's goals and the whole university goals) and interpreted in the context of the whole unit and/or institution.

In this paper we strive to show what modern methods of multiple criteria evaluation can offer in this area and how linguistic modeling of expert knowledge can prove useful in the evaluation process and to share our experience in designing such models. We will do so using as a case study the information system for academic staff performance evaluation (IS HAP), which has been developed for Palacký University in Olomouc (Czech Republic) as a novel performance assessment and HR management tool. Palacký University in Olomouc is one of the oldest universities in Central Europe. With 8 faculties covering all the main areas of scientific interest, it is a good example of a Central European university. The development of the evaluation model presented in this paper started in 2006. At first it was intended for Palacký University, Faculty of Science (which is a strongly research oriented faculty). Recently the described evaluation system has been incorporated into the Individual National Project "Maintaining and assessing quality in tertiary education" and financed by the European Social Fund and the state budget of the Czech Republic. It is therefore realistic to assume that the Ministry of education of the Czech Republic will recommend this system to be used by other Czech universities.

General requirements of Palacký University on the evaluation model were as follows: It should 1) include, if possible, every aspect of academic staff activity; 2) use only easy to verify and objective data; and 3) be easy to work with. Other requirements were for the final evaluation: 4) to maximally reflect staff benefit to the faculty; and 5) to be able to flexibly respond to management needs. The desired output of the model was not to arrange members of academic staff in order of their performance, nor to obtain a single number interpretable only with difficulty. A basic piece of information on both focus and performance of the academic staff was considered sufficient. The previously mentioned requirements and the need of complete comprehensibility for academic staff members and their superiors implied in the end the use of linguistic fuzzy modeling – linguistic variables, rule bases, and approximate reasoning (i.e. of fuzzy expert systems).

Before the development of the system, various academic staff evaluation models currently used on universities both in the Czech Republic and abroad were subjected to a detailed analysis regarding their practical and mathematical aspects. The analysis resulted in the design of several academic staff evaluation models, differing in how members of academic staff are evaluated in separate areas of their activity and in the aggregation method for these partial evaluations (weighted average, OWA, WOWA – see Yager (1988) Torra (1997) or Torra and Narukawa (2007) for more details). The final choice of fuzzy methodology was also justified by the analysis of performance of the aggregation methods for partial evaluations. The weighted average (as the most commonly used aggregation operator in the staff evaluation models; weights are fixed for specific areas of activities) is unable to accept staff specialization, penalize unsatisfactory performance and promote excellent performance. The linguistic fuzzy modeling provides a solution to all the mentioned shortcomings while maintaining a high level of comprehensibility of the model (see Stoklasa, Talašová and Holeček, 2011, for a detailed analysis of the OWA and WOWA operators).

Various universities worldwide deal with the task of evaluation differently; see e.g. Masaryk University Brno (2009), University of Technology Sydney (2009) or Wayne State University (2009). In all the analyzed cases, the WA was the only aggregation operator used for aggregating partial evaluations.

Methodology

Fuzzy set theory and linguistic modeling

The fuzzy set theory and linguistic variables introduced by Zadeh (1965 and 1975 respectively) provide a valuable tool for modeling human experience and expert knowledge of systems. Using this framework, we can deal with linguistically described variables (see Figure 1), goals and restrictions, linguistically described relationships and we are able to get linguistic outputs. The idea behind linguistic fuzzy models can be, for the purpose of this paper, roughly simplified into the following statement: anything that can be described in everyday language can be represented mathematically and computations (evaluation procedures) can be carried out and still remain completely understandable. What is even more attractive, the description of complex relationships can remain linguistic (see Figures 2 and 4) and still the corresponding functions (i.e. evaluation functions) can be derived and used (see Figure 3).

Before we proceed further, let us briefly describe some basic notions of fuzzy set theory. Dubois, Prade (2000) provide formal and more thorough definitions of the notions used in this paper. Talašová (2003), Mamdani and Assilian (1975), Sugeno (1985) and Ruspini (1969) also provide valuable insights into this topic. A fuzzy set can be seen as a generalization of a regular (crisp) set. If we consider crisp sets, we can only decide for any element of the universe that it either fully belongs to the set or does not belong to that set at all. However, such approach is not well suited for dealing with sets where partial belonging to a set is conceivable. These situation occur frequently when we describe a characteristic feature of a set linguistically (consider the set of all *intelligent* people). For such a set it may not be easy to decide whether a current person belongs to the set or not, needless to say that such thinking would be rather counterintuitive (we can imagine someone that is “rather intelligent” or “extremely intelligent” and we usually distinguish between these categories in real life). For any model that should represent such reality we need a different tool than crisp sets.

Fuzzy sets allow for the elements of the universe to partially belong to the set. The strength with which an element x belongs to a fuzzy set A is typically expressed as a real number $A(x)$ from [0,1], where $A(x)=1$ means the element x fully belongs to the fuzzy set A and $A(x)=0$ means that x does not belong to the fuzzy set A at all. This way $A(x)=0,7$ can be interpreted in terms of the previous example that a person x is 70% intelligent, or in other words that the linguistic term “intelligent” is 70% accurate for the description of the person x. $A(x)$ is called a membership degree of x to the fuzzy set A. Fuzzy sets make linguistic description of various features and characteristics of people, their performance etc. mathematically manageable.

If we choose to describe the values of a variable only linguistically, we get a linguistic variable (ie. “*performance*”, with values {*poor, acceptable, excellent*}). We can assign a fuzzy set to every linguistic value of such variable to represent its meaning mathematically (see Figure 1). Linguistic variables for which the meanings of the linguistic terms are ordered and for which it holds that the belonging of any element of the universe (1 or in other words its full membership) can be completely divided among the values of the linguistic variable, are called linguistic scales. With such linguistic variables we can perform similar operations to those we are used to with classical (i.e. real-valued) variables. Figure 1 presents examples of linguistic scales (used in the presented system to describe performance in PA and RD).

We can even construct linguistic rules and perform deduction formally (deriving outputs from a linguistic fuzzy rule base is called approximate reasoning).

In the presented application, we use a modification of the Sugeno-Yasukawa approach to

approximate reasoning (Sugeno and Yasukawa, 1993) described in more details in (Stoklasa et al., 2011). Let us consider two linguistic variables A and B (input variables) with the sets of linguistic values $\{a_1, a_2,\ldots,a_n\}$ and $\{b_1, b_2,\ldots,b_m\}$ respectively. The linguistic values (linguistic terms) of these variables are assigned meanings in the form of fuzzy sets $\{A_1, A_2, \ldots, A_n\}$ and $\{B_1, B_2, \ldots, B_m\}$ respectively. We assume one output variable O with the set of linguistic values $\{o_1, o_2,\ldots,o_s\}$, their meanings modelled by fuzzy sets $\{O_1, O_2,\ldots,O_s\}$. Let us assume a description of the relationship between the input variables and output variable in the form of a set of IF-THEN rules:

$$\begin{gathered}\text{IF } A \text{ is } a_1 \text{ AND } B \text{ is } b_1 \text{ THEN } O \text{ is } o_1\\ \text{IF } A \text{ is } a_2 \text{ AND } B \text{ is } b_1 \text{ THEN } O \text{ is } o_2\\ \ldots\\ \text{IF } A \text{ is } a_n \text{ AND } B \text{ is } b_m \text{ THEN } O \text{ is } o_s.\end{gathered} \tag{1}$$

If just one of these rules fires for a particular input, then the result of the deduction is the linguistic term on the right hand side of the respective rule, and the strength of such an output is based on the firing strength of the rule (on how well the linguistic terms a_i and b_j on the left hand side of the rule describe the particular input pair (a', b') that can be computed as $A_i(a')*B_j(b')$). For more possibilities of computing the rule firing strength see (Dubois and Prade, 2000).

Should more rules fire for an input pair (a', b'), the output is computed using the firing strengths of the rules and the respective right hand side linguistic term's meaning (Talašová (2003), Mamdani and Assilian (1975), Sugeno (1985) and Stoklasa et al. (2011)). We can also replace the set of linguistic terms $\{o_1, o_2,\ldots,o_s\}$ by a set of numerical labels for the output linguistic terms, that in the case of evaluation can be considered to form a cardinal scale. We can then compute a numerical result (and thus construct an evaluation function – see Figure 3) and then use the linguistic terms $\{o_1, o_2,\ldots,o_s\}$ to describe this output (Stoklasa et al., 2011).

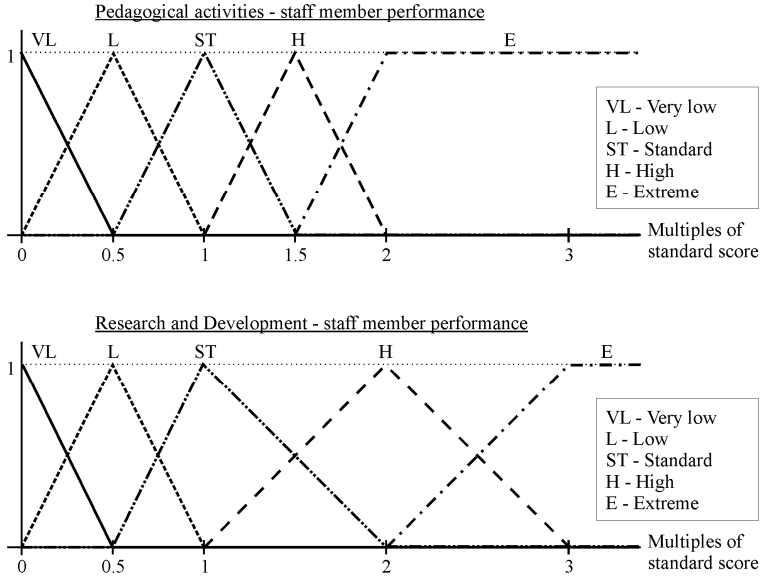


Figure 1. Meanings of the linguistic terms used to describe staff members' performance in PA (upper) and RD (lower).

An illustration of such approach will be presented further in the paper. Such evaluation methodology enables us to aggregate partial evaluations of different types (numerical, linguistic, fuzzy) and on different scales. Linguistic fuzzy modelling thus provides an easy to adjust evaluation tool, which can be effectively used even by laymen. The software and software packages for fuzzy modelling such as Fuzzytech (Von Altrock, 1995) or Matlab (Mathworks, 2011) can be utilized. Specialized tools developed primarily for the purpose of

multiple criteria (fuzzy) evaluation such as FuzzME (see http://fuzzme.wz.cz/ or Holeček and Talašová, 2010,) are also available.

Evaluation methodology

The performance of each member of academic staff is evaluated in both pedagogical (PA), and research and development (RD) areas of activities. Input data are acquired from a form, filled in by the staff, where particular activities are assigned a score according to their importance and time-consumption. Three areas are taken into consideration for pedagogical performance evaluation: a) lecturing, b) supervising students, and c) work associated with the development of fields of study.

The evaluation of RD activities is based on the methodology of evaluation of research and development outcomes valid in the Czech Republic, but other important activities (grant project management, editorial board memberships etc.) is also included. Both pedagogical and RD areas are assigned a standard score – different for senior assistant professors, associate professors, and professors.

The number representing a partial evaluation of a member of academic staff in a certain area (PA or RD) is determined as a multiple of the respective standard for his or her position. For better clarity and easier interpretation, these numbers are transformed into verbal evaluations using linguistic scales (see Figure 1). We may observe that the meaning of "extreme" performance differs in PA and in RD, which is the result of different evaluation approach to these two areas of interest. The evaluation of PA reflects mainly time-consumption of the activities and performance better than twice the standard score (twice the standard performance) is considered to be well captured by the linguistic term "excellent". Whereas the RD area is evaluated in accordance with the current methodology for R&D outcomes evaluation in the Czech Republic (see Government office of the Czech Republic, 2010), where the quality of the outcome (particularly in the case of scientific papers the quality measure is based on the rank of the journal in a decreasing sequence of journals in the current field ordered according to the impact factor of the journal) plays an important role. This way "excellent" is a 100% accurate description for performance that is evaluated better than 3 times the standard score.

Overall performance of a current staff member in PA and RD		**Research and development performance**				
		Very low	**Low**	**Standard**	**High**	**Extreme**
Pedagogical activities performance	**Very low**	**Unsatisfactory**	**Unsatisfactory**	**Substandard**	**Standard**	**Very good**
	Low	**Unsatisfactory**	**Unsatisfactory**	**Substandard**	**Very good**	**Excellent**
	Standard	**Substandard**	**Substandard**	**Standard**	**Very good**	**Excellent**
	High	**Standard**	**Very good**	**Very good**	**Excellent**	**Excellent**
	Extreme	**Very good**	**Excellent**	**Excellent**	**Excellent**	**Excellent**

Figure 2. Linguistic fuzzy rule base describing the aggregation of partial evaluations in two areas of activities – PA and RD. Grey tinted cells describe the overall performance.

As the partial evaluations differ in their nature, linguistic fuzzy expert system is used to aggregate both partial evaluations – for pedagogical and R&D areas of activities. One of the main advantages of this type of aggregation is that it allows to set-up the shape of the aggregation function completely in line with the evaluator's requirements (e.g. to appreciate excellence achieved in one of the areas). This type of aggregation is transparent and comprehensible even to a layman as it is described in linguistic terms (see Figure 2). The evaluation function described by the linguistic fuzzy rule base in Figure 2 can be represented graphically (see Figure 3).

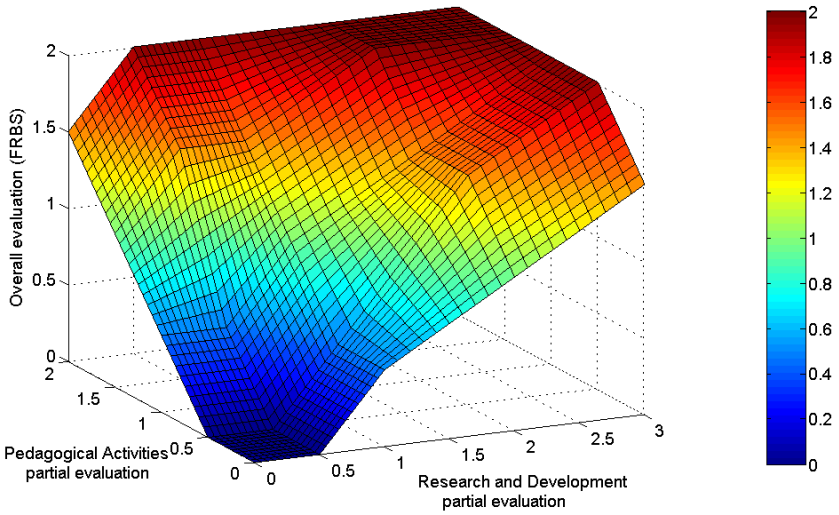


Figure 3. Plot of the function used for aggregating partial evaluations in PA and RD (FRBS). This function is described by the linguistic fuzzy rule base (see Figure 2).

We can see that the linguistic description of the process of aggregation of partial evaluations is consistent in the sense that the evaluation function is increasing in both variables and that it reflects well the linguistic description in Figure 2. We may also observe how a nontrivial evaluation function can be "elegantly" described linguistically and all the important information thus remain clear to both the evaluator and to the people that are being evaluated. The overall aggregated evaluation is also available as a linguistic expression.

A formal description of the aggregation of partial evaluations in PA and RD as described linguistically in Figure 2 can be formulated in the following way - see (Stoklasa et al., 2011) for more details.

For any pair of real inputs $pa \in [0,BB]$ and $rd \in [0,CC]$ we can now compute the output (real) value $eval(pa, rd)$ using formula (2).

$$eval(pa, rd) = \frac{\sum_{j=1}^{k} A_j(pa) \cdot B_j(rd) \cdot ev_j}{\sum_{j=1}^{k} A_j(pa) \cdot B_j(rd)} = \sum_{j=1}^{k} A_j(pa) \cdot B_j(rd) \cdot ev_j , \qquad (2)$$

where

- *pa* (or *rd*) is the partial evaluation of a particular academic staff member in the area of PA (or RD) in terms of standard scores multiples
- where *BB* and *CC* are sufficiently high real numbers not to be exceeded by any actual PA and RD partial evaluations respectively
- A_j is the fuzzy set representing the meaning of the linguistic term describing *PA* in rule *j*, *j=1, ..., k;*
- B_j is the fuzzy set representing the meaning of the linguistic term describing *RD* in rule *j*, *j=1, ..., k;*
- ev_j is the numerical label corresponding with the linguistic term on the right side of the rule *j*, *j=1, ..., k.*

Although the formula (2) can be used to determine the overall evaluation, the same process described linguistically offers the possibility of using the evaluation function to motivate people in the desired direction and to communicate what performance is desirable and what is undesirable.

Our model also takes into account the load of secretarial and managerial activities with each member of academic staff (understood here as activities draining away from his or her time and thus reducing the performance in each of the two areas of evaluation mentioned above). Another fuzzy expert system is used to adjust the evaluation according to the managerial activities load of the particular academic staff member. The overall work load of members of academic staff is thus described in words.

Practical implementation of the evaluation

The IS HAP is a web based application, which is accessible through Internet. Each of members of the academic staff fills in a form summarizing his or her activities in the last year. The form was implemented so that it would be as easy to fill it in as possible. All items of the form are divided into categories and subcategories. A brief help for any of the items is also available. The current version of the software is able to communicate with the information system of Palacký University in Olomouc. This reduces the time necessary to complete the form.

Each of the academic staff members can see his or her own evaluation before the form is sent. After the form is filled in, the head of the respective department and the dean can see the overview of the resulting evaluations of the academic staff members (Figure 4). All the filled in forms are also available to the heads of departments (they can access the forms of the academic staff members of his/her department) and to the dean (he/she can access the forms of all academic staff members of the faculty). All these features make the resulting evaluations very easy to interpret and compare.

The head of the department is provided with the following data about members of his or her department: his/her name and position, evaluation in PA including the information on the type of the activities for which it was acquired, evaluation in RD, which is again supplemented by an additional information on type of activities, the list of managerial and secretarial functions and the final evaluation of all the activities (PA and RD) which also takes into the account the working time drained by the previously mentioned functions. The heads of the departments can view a table summarizing the results of the evaluation (overall evaluations of all the staff members are expressed graphically using colour-bars - each of the performance classes is represented by a different colour), complete forms of the members of their departments and see all of their activities. In the form, the academic staff members also specify their plans in pedagogical area and RD for the next evaluation period. The heads of their departments can easily check if their subordinates managed to realize their plans. Academic staff members can access only their own evaluation forms and results.

IS HAP represents a unique software tool that implements novel approach to performance evaluation. It is easy to use for both academic staff members and their subordinates. The sophisticated mathematical background provides evaluations highly valued for the HR management, which are presented in a well-arranged and comprehensible way.

Results

After its pilot testing on one department in 2009, the evaluation methodology described in this paper was presented to the academic senate of the Faculty of Science, which confirmed the meaningfulness of the presented evaluation approach and accepted it well. The system has been set up to meet not only the requirements of the management, but to be a valuable tool for the academic staff members themselves (providing a list of all the activities performed by the current staff member e.g. for writing professional CV, providing clear information what activities are important for the department, what performance is expected). This has been

achieved by discussions with the management and academic staff members, psychologists and HR specialists. An extensive pilot testing on 14 departments followed in 2010. The results of the pilot testing confirmed that the model meets all the requirements. It was therefore decided to implement the designed methodology of evaluation into an information system for academic staff performance evaluation (IS HAP). Since 2011 IS HAP has become the annual performance evaluation tool for academic staff members of the Faculty of Science.

The presented evaluation methodology (and its software implementation) is a product of interdisciplinary cooperation (mathematicians, computer scientists, a psychologist, a HR practitioner, and the management of the Faculty of Science participated during the development of the presented tool). This makes it a tool reflecting the needs of correctness and soundness of the used mathematical methods, but also the needs of easy interpretability and comprehensiveness. This way the proposed methodology is an effective support tool for human resource management.

The outputs of the evaluation are available on different levels of aggregation (the filled in forms, the evaluations in each area of interest and its subareas and the overall evaluation) and in the form of linguistic expressions. This makes it possible to further analyse the outputs and obtain much more information concerning the units (people or even departments) that are being evaluated. Having the aggregation function of partial evaluations described linguistically (by a fuzzy rule base) enables the employees to understand the process of evaluation. It is easy to see what performance in PA and R&D will result in what evaluation. The evaluator provides a list of activities that benefit to the faculty and are therefore "worth doing" (activities that are assigned scores). The score reflects the importance or time consumption of a particular activity. Evaluator sets up standards and describes how the performance in all areas of interest should be balanced in order to get a positive evaluation (by setting up a linguistic fuzzy rule base for aggregation of partial evaluations). This way the evaluator can express his/her preferences and motivate academic staff members to engage in activities that are important for the well-being of the university, faculty or department.

The system provides information concerning the overall performance of each academic staff member; more detailed partial evaluations in each area of interest are also available, as well as the filled in forms. This way a comprehensive set of information concerning the performance of each academic staff member is obtained. The outputs of evaluation are used by the management of the University for HR development (chances for improvement are identified), for setting up personal goals, for recruitment and outplacement purposes. The overall evaluations can also be compared with academic staff member's actual remuneration, discrepancies can be identified and corrected. The graphical representation of results enables quick orientation in a vast number of evaluation protocols, linguistically described evaluations are easy to interpret for the evaluator and the academic staff.

Based on the outputs of the evaluation it is for example also possible to determine the type of the worker. Let us assume that we have to distinguish among several types of units (in this case academic staff members) and that we are able to characterize the units by their performance in the evaluation areas of interest. We can allow this characterisation to be in the linguistic form – represented by linguistic rules. Figure 4 provides an example of a linguistic fuzzy rule base that can be used for the purpose of such classification.

Type of the academic staff member		Research and development performance				
		Very low	Low	Standard	High	Extreme
Pedagogical activities performance	Very low	Nonspecific	Nonspecific	Nonspecific	Researcher	Researcher
	Low	Nonspecific	Nonspecific	Nonspecific	Researcher	Researcher
	Standard	Nonspecific	Nonspecific	Nonspecific	Nonspecific	Researcher
	High	Teacher	Teacher	Nonspecific	Nonspecific	Nonspecific
	Extreme	Teacher	Teacher	Teacher	Nonspecific	Nonspecific

Figure 4. Linguistic rule base for determining of the type of a particular staff member based on his/her performance in PA and RD. Grey tinted cells describe the resulting type of the academic staff member.

In this case we assume there are three types of academic staff members. *Researchers* can be roughly described as focusing on R&D ("high" to "extreme" performance is typical for them) while their typical performance in PA is "standard" or lower. This rough description is transformed into a set of linguistic rules depicted in Figure 4 (all the rules that result in "*researcher*"). On the other hand *teachers'* performance in PA is expected to be "high" to "extreme" while maintaining a "standard" or lower performance in R&D. Academic staff members, that can be characterized neither as researchers nor as teachers will be labelled as *nonspecific*. Based on the classifier described by the fuzzy rule base in Figure 4 and using the single winner method of fuzzy classification (see Ishibuchi, Nakashima and Morisawa (1999) for more details) an academic staff member whose evaluation is 2.1 in PA ("extreme" is 100% appropriate linguistic label) and 0.5 in RD ("low" is 100% appropriate linguistic label) will be classified as a *teacher*. Another example might be a staff member with 1.3 in PA ("high" is a 60% appropriate linguistic description, the performance is somewhere between "standard" and "high") and 2.2 in RD ("high" is an 80% appropriate linguistic label, performance is between "high" and "extreme") who will be classified as *nonspecific*.

Discussion

The paper presents an evaluation methodology that is based on multiple evaluation criteria enables the evaluator to describe the evaluation function linguistically and obtain linguistic outputs. The mathematical apparatus used in the methodology provides a sound mathematical background on which the evaluation methodology can be used, tuned, transformed and even new functions added (such as the proposed tool for determining type of the worker) on linguistic level. The fact that the mathematical level remains still in contact with the linguistic (interpretation) level makes the evaluation easy to set up - Figure 3 is a nice example of a complicated function described comprehensibly by linguistic rules (Figure 2). What is more important, the evaluation and its principles is described in a way comprehensible to people that are being evaluated. This way the evaluator can express his/her intentions and the staff can react accordingly. We do not only present an evaluation methodology, we also present a management tool for communicating goals and setting up boundaries. The use of such evaluation methodology can prevent misinterpretations of numerical outputs of classical evaluation methodologies and provide means for more creative and still theoretically sound approaches to evaluation. The presented evaluation methodology can be easily adapted to be used in various application fields. Evaluation and mathematical methods for evaluation support are a topic still open to research. Possible directions of enhancing the presented methodology include the use of additional evaluation criteria, finding an optimal way of presenting outputs to the evaluators on various levels of aggregation and processing the outputs of the evaluation process so that more information could be gained from the inputs. The development of linguistic fuzzy mathematical methods for evaluation processes is also an interesting and promising direction for future research.

References

Dubois, D. & Prade, H. (Eds.) (2000). *Fundamentals of Fuzzy Sets*. The Handbook of Fuzzy Sets Series. Kluwer Academic Publishers, Boston-London-Dordrecht. ISBN 0-7923-7732-X.

Government office of the Czech Republic (2010). Methodology for evaluation of research organisations and outcomes of completed programmes (in Czech) http://www.vyzkum.cz/storage/att/5591F655709A8A76C4778C8E14BEB413/Metodika%20hodnoceni%20vysledku%20vyzkumnych%20organizaci%20a%20vysledku%20ukoncenych%20programu%20ve%20zneni%20pro%20rok%202011.pdf Accessed 4.3.2012.

Holeček, P. & Talašová (2010). Designing Fuzzy Models of Multiple-Criteria Evaluation in FuzzME Software. *Proceedings of the 28th International Conference on Mathematical Methods in Economics*, **1**, pp. 250-256

Ishibuchi, H., Nakashima, T. & Morisawa, T. (1999). Voting in fuzzy rule-based systems for pattern classification problems. *Fuzzy Sets and Systems* **103** (2), 223-238.

Mamdani, E. H. & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller, *Int. J. Man-machine Studies*, **7**, pp. 1-13

Masaryk University Brno (2009). Determination of criteria for pedagogical and other activities evaluation (in Czech). [online]. Masaryk University, Faculty of Law, Brno Accessed 2. 3. 2012.http://is.muni.cz/do/law/ud/predp/archiv_predpisy/4862802/Pokyn_dek._c._7-2009_urceni_kriterii_pro_ hodnoceni_ped._a_j.cin..pdf .

Matějů et al. (2009). The white book of tertiary education. (in Czech) Ministry of Education, Youth and Sports, Czech republic.

Matheson, W., Van Dyk, C. & Millar, K. I. (1995). *Performance evaluation in the human services.* The Haworth Press, New York-London. ISBN 1-56024-379-1.

Mathworks Inc. (2011). Fuzzy logic toolbox 2.2.13 The Mathworks Inc.

Ruspini, E. (1969). A new approach to clustering. *Inform. Control*, **15**, pp. 22-32.

Stoklasa, J., Talašová, J. & Holeček, P. (2011). Academic staff performance evaluation – variants of models, *Acta Polytechnica Hungarica* **8** (3), p. 91 – 111, ISSN 1785-8860, available also online <http://www.uni-obuda.hu/journal/Stoklasa_Talasova_Holecek_29.pdf>.

Sugeno, M. (1985). An introductory survey on fuzzy control. *Information Sciences*, 36, pp. 59-83.

Sugeno, M. & Yasukawa, T. (1993). A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on fuzzy systems*, **1** (1), pp. 7-31.

Talašová, J. (2003). Fuzzy methods of multiple criteria evaluation and decision making. (in Czech). Palacky University, Olomouc, ISBN 80-244-0614-4.

Torra, V. (1997). The weighted OWA operator. *International Journal of Intelligent Systems*. **12** (2), pp. 153-166.

Torra, V. & Narukawa, Y. (2007). *Modeling Decisions*. Springer, Heidelberg. ISBN 978-3-540-68789-4.

University of Technology Sydney (2009). Performance Management [online]. Accessed 2. 3. 2012. <http://www.hru.uts.edu.au/performance/reviewing/rating.html >.

Von Altrock, C. (1995). *Fuzzy logic and neurofuzzy applications explained*. Prentice-Hall, New York.

Wayne State University (2009). 2009-10 Guidelines for Evaluation of Academic Staff [online]. Accessed 2. 3. 2012. <http://www.aaupaft.org/pdf/AcStaffguidelines_2009-10.pdf>.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria

decision making. *IEEE Trans. On Systems, Man and Cyberneics*, **1** (3), pp. 183–190.
Zadeh, L. A. (1975). The concept of linguistic variable and its application to approximate reasoning. *Information sciences*, Part 1: 8, pp. 199-249, Part 2: 8, pp. 301-357, Part 3: 9, pp. 43-80.
Zadeh, L. A. (1965). Fuzzy Sets. *Inform. Control*, **8**, 1965, pp. 338-353.

VI

Different Aspects of Evaluating Social Care and Health Services

From Improving to Proving: How a Program Evaluation Developed into a Research Project

Dr Susan Fletcher; Department of Social Work, Monash University, Melbourne Australia

Acknowledgements

The author wishes to thank Senior Lecturer Mollie Burley for her assistance in the Healthy Heart evaluation

Abstract

This paper describes how an evaluation of a cardiac rehabilitation program offered in rural Australia lead an agency to consider a new model of service delivery and stimulated a research project. The sequence of impacts will be reported in three phases. Firstly, the findings of the cardiac rehabilitation program evaluation. Secondly, the results of this evaluation lead to changes in delivery of this and similar chronic disease programs across the agency, and, thirdly, the findings inspired two local academics to commence a larger research project to explore some of the issues that arose from the evaluation.

Keywords: Program Evaluation; Cardiac Rehabilitation; Rural Health; Collaborative Research.

Introduction

The objective of this paper is to report on how the evaluation of a cardiac rehabilitation (CR) program, aimed at improving outcomes for participants, lead to a review of the way chronic disease programs were offered in this rural agency and stimulated a research project that will explore how we can better support clients post a cardiac event.

Evidence supports the benefits of Cardiac Rehabilitation (CR) as a means of secondary prevention of coronary heart disease (Sandararajan et al, 2004; Clarke et al, 2005; Grupta et al, 2007). However, the impacts of attending CR on the client are less well understood, especially for those living in rural settings (Thornhill and Stevens, 1998).

Phase 1: Program Evaluation

The cardiac rehabilitation program described in this paper is conducted at Latrobe Community Health Service (LCHS) in rural Victoria. It was an 8-week program, based on Heart Foundation Australia guidelines, and consisted of exercise sessions, health education about diet, stress, smoking cessation and behavioural change strategies. An evaluation of this program was requested by LCHS to assess whether the program was of benefit to participants and to suggest ways of improving the program.. Previous studies had focused on outcome measures, such as weight loss or reducing blood pressure, and while these are important elements of good health, it was decided to obtain feedback on the program from the participants' perspective.

Study Design

Data Collection. We were interested in hearing the voices of the program participants and so it was decided that a focus group design would be used for data collection. Focus groups allow participant perspectives to be examined in-depth and for interaction pertaining to a particular topic to take place within the group setting. Focus groups are commonly used in healthcare to explore health issues and to test ideas about and acceptance of new programs (Liamputtong and Ezzy, 2005).

Sample. The participant sample consisted of 22 clients who had attended the CR program in 2010. There were 19 males and 3 females with an average age of 71.7 years (range from 58-88 years). The study received Ethics approval from Monash University and invitations were sent out seeking volunteers to attend a focus group. Two hour long focus groups were arranged and the discussion was audio taped with consent. In addition to general questions about program content, participants were asked about their experience of maintaining lifestyle change after the program finished.

Data Analysis. Inductive content analysis was chosen because there was not enough knowledge about participant experiences of CR in a rural setting (Lauri and Kyngas, 2005). Two researchers independently read through the text several time to become completely familiar with the data. The data was divided into Focus group 1 and Focus group 2 and read again to reveal any differences between the groups As many headings as possible were written down to describe all aspects of the content (See Figure 1) and categories were generated.

Data Extract	**Coding for**
I have been organising my own motivation checklist. It's more important now that I have to take care of my grandson who can't live at home at the moment.	1. Talked about motivation strategy 2. Care role helped keep on track

Each category was named and participant statements with similar themes were grouped together as sub-categories and sub-categories were then grouped together as main categories.

Results

Eight of the eligible participants (seven men and one woman) volunteered to take part in the focus groups. This gender profile is consistent with the literature on attendance at CR (Sundararjan et al, 1998). Responses received to the questions raised in the two groups were very similar and so will be reported as one group. Analysis of the focus group data identified three main categories that illustrated the success of the program as well as the vulnerability still experienced by the participants at program completion.

Three main categories and six sub-categories were identified to describe the participants' experience of taking part in the program

The first main category, *'recovering confidence'* referred to the sessions and what did or didn't work for the participants in their experience of the program. There were two sub-

categories in this category, *‘ increase in self- confidence’* and *‘ supportive environment’*. The sub theme, ‘ increase in self-confidence’ illustrated the participants’ experience of feeling that their attendance at the program had increased their ability to exercise.Participants spoke of their vulnerability post their cardiac event and of feeling uncertain about what level of excerise would be safe.Information about diet and self-management was also highly regarded, although the participants reported difficulties in transferring this learning to their home lives.The sub-category *‘supportive environment’* illustrated the participants’ description of how the program structure helped them to increase their confidence.The group format was important as it provided a protective space for them to moniter how much effort they could put into the exercises. The encouragement provided by the clinician was very important in this sub theme.The participants reported experiencing a high level of uncertainity but that the *‘ good design of the program and the high level of supervision’ (* P3 –Group 1), enabled a restoration of self-confidence. The value of this level of support was highlighted by P3 - Group 1 who said, *‘ I can’t do it on my own’*.

The second main category *‘ putting it into practice’* referred to the fact that the program can only offer options, the individual has to be open to select and maintain motivation post the program. The sub-category, *‘ maintaining motivation’* provided mixed results from the participants. Maintanence of regular walking was the most reported common activity post program (n=3).A second sub-category was *‘ co-morbidity’*.Three participants reported pre-existing back problems and that these had compromised their ability to exercise regularly. P3-Group1 said that attendance at the program had, *‘flicked a switch on’* regarding the need to exercise for health maintance, but made the comment later that he,’ *struggled to get out of his chair at home’*. Diet changes were reported on by 3 participants. P2-Group 2, said that the advice, *‘ changed my life’* .P3-Group 2 was, *‘watching food, less coffee, eating less and types of food’*. and P3- Group 2 had taken on board the need to *‘ change eating habits’*.

The third sub-category ,‘ family support’ refered to the need for participants to receive on-going support post program,in order to maintain the lifestyle changes needed to reduce risk factors. Three participants highlighted the need for family and partners to be supportive. P4-Group 1’s partner had accompanied him to all of his CR sessions but he found his adult childrens’ attitudes were deterring him from further exercise opportunities. P1-group 1 reflected that it was harder *‘ being single, as opposed to having a partner.* P2-Group 1 said that it was a *‘struggle’* in all aspects of maintaining motivation post program because they lacked, *‘ some-one to hold your hand, no-one to talk to at home’*.

The third main category *‘ feeling abandoned’* referred to participant comments about how difficult it was to move from a structured , supportive environment to their normal homelife. The sub-category *‘ feeling alone and isolated’* was dominated by stressors that were created outside the Healthy Heart Program., such as a lack of family support or where co-morbities came into the foreground. The most concening comments, however, related to a perceived lack of support available after the program finished. P2-Group 1 summed it up with the comment, *‘ no support, no call back, no follow-up’*. Others suggested that they had needed more time to restore confidence and that follow-up was necessary to maintain their motivation. It was acknowledged that it would be different for each person. The sub category *’ linking back into the community’* illustrated how the participants struggled to move from the CR program to other community programs, such as a gym. Even though part of the CR program content was specifically aimed at addressing this need by identifying resources and visiting facilities so as to facilitate the transition, participants still reported difficulty in accessing community options. Comments included, *‘ cost of sessions at gym’* (P2-Group 1) and two participants (P3-Group 1 and P4 –Group 1) mentioned feeling less safe in an unstructured exercise environment.

Summary of Phase 1

Findings from this evaluation showed that the participants had a positive association between the opportunity to learn about diet, self-management , the social aspect of the group and the reassurance of exercising within a safe environment. These aspects of the findings are consistant with previous studies(Clark et al,2011). Recovery, however, is not a time limited process, as the lifestyle adjustments needed once a cardiac diagnosis is made are life long (King et al,2001).Participants in the focus groups described their difficulties in maintaining motivation for change post program. According to the National Heart Foundation Cardiac Rehabilitation Framework, "ongonig maintenance of behaviour change beyound the period of inpatient and outpatient rehabilitation is critical if long-term health benefits are to be realised" (NHFCRF,2004 :4).Participants' accounts in this study indicted that the content of the CR program was not strongly linked to longer-term health benefit change. There were individual efforts to continue exercise and moniter diet but the majority spoke of the struggle this involved once they were at home. Comments about feeling abandoned and left to their own devices without sufficent regular professional input surfaced in several participant responses.

While the ability to self-manage is a highly desireable goal of the program, it was made clear to the researchers that the participants felt ill-prepared to maintain lifestyle change..This was not as a result of the program content or clinicians, all of whom were commended by the participants, but because the participants felt unable to to take full responsibility at this point in time. Echoing other research (Kielmann et al,2010), participants expressed the need for more support to promote behaviour change after the completion of the program.The participants in this study have experienced a traumatic health event and are then asked to change their lifestyle and adhere to a self-management approach.All of these significant events came together in a relatively short period of time and provided many challenges for the participants..It also raised a challenge for me as a researcher. How could we develop the cardiac program to meet the ongoing support needs of the participants?

Phase 2: Evaluation Recommendations and Agency Adaptation

One way that health professionals can assist clients with a chronic disease is by identifying barriers to self-management and working with them to overcome these barriers. Gately (2007) encourages change in service organizations and professional practices at this interface with clients.

Having completed the evaluation a number of recommendations were prepared for the organization. The recommendation that addressed the issue of clients 'feeling abandoned' was the one immediately focused on by the Community Health Service. This recommendation suggested the development of a **Review** and **Revise** strategy. In the review process, clients would be asked to complete a goal setting exercise towards the end of the CR program. This exercise would ask them to identify change/s they wanted to make over the next month, what steps they would take to reach the goal/s, what difficulties they could foresee, how they might overcome these and to rate their level of confidence in achieving their goal/s. One month after the program, the Community Health Service clinician would follow-up (via phone) to discuss how the client was going with achieving their goal/s and maintaining behaviour change. If goals were being achieved then new goals could be set and an arrangement would be made to phone again in three months.

If goals were not being met and the client was concerned about their ability to maintain motivation, then the process would move into the Revise stage. In this event the client would be invited to join an established exercise group or if the issues were more complex then

referral to a health coach could be considered. This recommendation is currently moving through the agency's' internal processes and has potential application to other chronic disease groups.

Phase 3: Research Project Development: Readiness to Rehabilitate

The evaluation finding ' feeling abandoned' raised the important question of whether the current cardiac rehabilitation program design adequately supported the client's need to establish and maintain self-managed behavioural change. Previous feedback about attendees' participation at CR had been positive. Participants expressed high levels of satisfaction and clinicians observed increased exercise tolerance and attitude change. It wasn't until participants were invited to discuss their individual experience that deeper issues emerged. This became the stimulus for this researcher to propose a research project aimed at exploring client decision-making about their attendance or non-attendance at cardiac rehabilitation opportunities and their capacity to self-manage.

Cardiac rehabilitation services appear to have been conceptualised in a simplistic manner, whereby programs are presumed to provide services in a fixed, uniform manner to passive inter-changeable subjects (Clark et al, 2004). While the efficacy of cardiac rehabilitation is ultimately determined by changes in cardiovascular risk in the long-term, a more sophisticated approach to understanding participant's perspectives on the dynamics and sustainability of behaviour change associated with cardiac risk is needed. Despite the prevalence of research exploring factors associated with low attendance rates at CR programs (Cooper et al, 2002; De Angelis et al, 2008), there has been limited examination of the decision-making processes clients go through about whether to attend or not attend CR and how this decision-making process supports or hinders the maintenance of behaviour change.

Previous Australian studies about attendance at cardiac rehabilitation have focused on attendance rates and on enablers or barriers to attendance, such as transport, gender or co-morbidities (Bunker et al, 1998;Schultz and McBurney, 2000; Sundararajan et al, 2004). There are few studies that have considered the client as being on a continuum from deciding that they will or will not alter their risk factor management behaviour through deciding to take lifelong responsibility for maintaining behaviour change.

The proposed study is collaboration, between researchers from Monash University, Latrobe Regional Hospital (the local public hospital) and Latrobe Community Health Service. After treatment for a cardiac event, clients are referred to an outpatient's rehabilitation program at the hospital and then to a community-based program at the local health service. By understanding the decisions clients make about their participation in the rehabilitation offered, these agencies will be better able to tailor referral pathways and programs to better address the wellbeing of these clients. This study is only beginning. Ethics approval has been sought and a grant application has been successful. What follows is the research plan we have for the project.

Research method

This study will follow the usual care of cardiac clients in the Latrobe Valley. The cardiac rehabilitation co-ordinator (CRC) at Latrobe Regional Hospital meets new cardiac patients, either at the time of admission for an acute event, or by appointment if they are referred from another facility. At this visit the CRC will ask the client's permission to provide their contact details for follow-up by one of the research team at various times over the next six to twelve months.

Clients who do not attend the hospital's outpatient's cardiac rehabilitation program after

their discharge from hospital will be asked to join a focus group to explore and discuss the factors that influenced their decision and any factors that might alter their decision. The focus group will use a semi-structured format to ensure a broad ranging discussion but allow for individual perspectives. Clients who attend the cardiac rehabilitation program at Latrobe Regional hospital (LRH) will be asked, both at the start and completion of the program, to complete questionnaires about their readiness and self-efficacy for cardiac risk factor management. Questionnaires to explore the participants' readiness to change their behaviour have been developed and tested in Victorian cardiac rehabilitation programs (McBurney et al, 1999). These are based on the stages of the trans-theoretical model of behaviour change as proposed by Prochaska et al (1988).

Some of these clients will go on to attend the rehabilitation program offered dry Latrobe Community Health Service (LCHS). These clients will again be asked, both at the start and completion of the program, to complete questionnaires about their readiness and self-efficacy for cardiac risk management.

Those clients who do not opt to go to the LCHS program will be asked to participate in a focus group to explore and discuss the factors that influenced their decision, and any factors that might have altered their decision.

A six-month follow up of clients will be conducted to review their progress. Clients will be randomly assigned to follow up, either by phone or by face-to-face interview. At this stage clients will again be asked about their ongoing readiness and self-efficacy for cardiac risk factor management.

Data Analysis

Quantitative data will be analysed using repeated measures analysis of variance by ranks to assess change in readiness to act on risk factors and self-efficacy for making or maintaining the appropriate behaviours.

Focus group discussions will be recorded and transcribed. In order to increase the rigor of the findings, transcripts will be analysed by two of the research team independently to identify barriers to attendance and factors that would facilitate attendance at programs. The final results of the analysis will be an agreed set of themes that describe the barriers to and facilitators of participation at different times in the client journey through rehabilitation.

Conclusion

We anticipate that the outcomes of this project will allow local agencies to provide cardiac rehabilitation programs that better meet the needs of participants. By understanding the individual decision drivers around whether to attend or not attend cardiac rehabilitation, we hope to identify strategies that will encourage more people to attend and complete rehabilitation. Studying the enablers and barriers to clients maintaining the lifestyle changes that are recommended if they are to lower their risk of sustaining another cardiac event may lead to the redesign of programs that are more flexible and responsive to the needs of these clients and should have application to a wide range of clients with a chronic disease.

From little things, big things grow. Findings from the evaluation of a community-based cardiac rehabilitation program lead to the agency reviewing the longer-term support needs of their client group. The same findings lead a local academic researcher to develop a study aiming to explore the decision-making process that clients use to attend or not to attend the rehabilitation programs on offer. " The purpose of evaluation is to *improve*, not prove" (Shufflebeam, 2007). The goal of the Healthy Heart evaluation was to improve the particular program. The planned research project aims to produce new knowledge in the field but they

are directly related by their desire to improve services for clients.

References

Australian Cardiovascular Health and Rehabilitation Association. 2008 *Practitioner's Guide to Cardiac Rehabilitation.* Sydney

Bunker, S., McBurney, H., Cox, H and Jelinek, V. 1998. Identifying participation rates at outpatient cardiac rehabilitation programs in Victoria, Australia. *Journal of Cardiopulmonary Rehabilitation and Prevention.* 19: 334-338.

Clarke, A., Barbour, R., White, L and MacIntyre. 2004 Promoting participation in cardiac rehabilitation: patient choices and experiences. *Issues and Innovations in Nursing Practice.* 47 (1): 5-14.

Clarke, A., Wheeler, H., Barbour, R and MacIntyre, P.2011. A realist study of the mechanism of cardiac rehabilitation. *Journal of Advanced Nursing.* 52(4): 362-371.

Cooper, A., Jackson, G., Weinman, J and Horne, R. 2002. Factors associated with cardiac rehabilitation attendance: a systematic review of the literature. *Cardiac Rehabilitation* 16: 541-552.

De Angelis, C., Bunker, S and Shoo, A. 2008. Exploring the barriers and enablers to attendance at rural cardiac rehabilitation programs. *Australian Journal of Rural Health.* 16: 541-552.

Gately, C. 2007. Re-thinking the relationship between long-tern condition self-management education and the utilisation of health services. *Social Science Medicine* 65:935-945.

Gupta, R., Sanderson, B and Bittner, V. 2007. Outcomes at one-year follow-up of woman and men with coronary artery disease discharged from cardiac rehabilitation: what benefits are maintained? *Journal of Cardiopulmonary Rehabilitation and Prevention* 27: 11-18.

Kielmann, T., Huby, G., Powell, A., Sheikh, A., et. al. 2010. From support to boundary: a qualitative study of the border between self-care and professional care. *Patient Education and Counselling.* 79: 55-61.

King. K., Humen, D., Smith, H., et al 2001. Psychosocial components of cardiac recovery and rehabilitation attendance. *Heart* 85: 290-294.

Lauri, S and Kyngas, H. 2005. *Developing Nursing Theories* Werner Soderstrom, Dark Oy, Vantaa.

Liamputtong, P and Ezzy, D. 2005. *Qualitative Research Methods: a health focus.* 2nd Edition. Australia: Oxford University Press.

McBurney, H., Reid, J and Bunker, S. 1999. " Stage of Physical Activity" and Actual Exercise Levels: Are We Doing Enough? Proceedings of the 9th Annual Scientific meeting of the Australian Cardiac Rehabilitation Association; Penrith, New South Wales.

National Heart Foundation of Australia and the Australian Cardiac Rehabilitation Association. 2004. *Recommended Framework for Cardiac Rehabilitation.* Sydney: NHFA and ACRA.

Schultz, D and McBurney, H. 2000. Factors which influence attendance at a rural cardiac rehabilitation program. *Coronary Health Care.* 4(3): 135-141.

Shufflebeam,D. 2007. *CIPP Evaluation Model Checklist.* Retrieved March 30th, 2012 from http://www.wmich.edu.evalctr/archive

Sundararajan, V., Bunker, S., Begg, S., et al 2004. Attendance rates and outcomes of cardiac rehabilitation in Victoria, 1998. *Medical Journal of Australia.* 180: 268-271.

Thornhill, M and Stevens, J. 1998. Client perceptions of a rural-based cardiac rehabilitation program: a grounded theory approach. *Australian Journal of Rural Health.* 52(4): 362-371.

Yoga for vulnerable adults with cancer

Jodi Constantine Brown, California State University, Northridge

Abstract

Objective: Previous research shows that yoga is associated with improvements in the overall quality of life of cancer patients including improved emotional well-being and physical outcomes such as sleep quality, mood, and stress. The purpose of this research is to describe and evaluate an agency-based yoga program for low-income adults with cancer. *Method:* Study participants included a convenience sample of adult patients seeking cancer treatment at a county hospital (n=26) and a separate group participating in a community-based yoga class (n=12). *Results:* Hospital and community class participants report that they are more relaxed and manage stress better, but results of a Mann-Whitney U test revealed no significant difference in quality of life between hospital yoga class participants (Md=63.91, n=14) and non-participants (Md=65.01, n=12), U=67.50, z=-.26, p=.79, r=-.05. *Conclusions:* The hospital yoga classes evaluated in this study differ from more traditional classes, and perhaps one yoga class is not enough to effect significant improvement in quality of life. Despite the limitations of this study, it serves as an excellent example of the challenges inherent in translational research.

Introduction

Almost 14 million Americans are living with cancer and have associated medical costs estimated over $124 billion annually (Mariotto, Yabroff, Shao, Feuer & Brown, 2011). Individuals diagnosed with cancer are faced with life-changing decisions that must be made in the face of fear, uncertainty, and high medical costs. Coupled with the disease itself are the side effects of treatment including fatigue, depression, anxiety, nausea, and other symptoms that negatively affect a patient's quality of life (Dorval, Maunsell, Seschenes, Brisson & Masse, 1998).

Complementary alternative medicine (CAM) treatments are rapidly becoming popular supplements to traditional cancer treatment (Lancaster, 2008). CAM therapy during and after traditional western medical treatments are often less costly, alleviate symptoms and engender feelings of empowerment over patient's bodies and course of treatment. A growing body of literature about CAM includes a wide variety of therapies such as herbs (Elkins, Rajab, & Marcus, 2005), yoga (Smith & Pukall, 2009), diet (Yates et al, 2005), mindfulness meditation (Kvillemo & Branstrom, 2011) and acupuncture (Brauer, El Sehamy, Metz & Mao, 2010). Elkins et al. (2005) found that CAM therapies were used by patients for a variety of reasons, while Buettner et al, (2006) found that factors associated with alternative medicine use varied according to the type of therapy. In a study of different types of complementary alternative medicine used by breast cancer survivors, Buettner et al (2006) found that users of most types of CAM had worse quality of life outcomes than nonusers, with the exception of cancer survivors who reported practicing yoga.

Research shows that yoga and exercise can help reduce the side effects of chemotherapy including fatigue (Banasik, Williams, Haberman, Blank & Bendel, 2011; Danhauer et al, 2009), stress (Pritchard & Birdsall, 2010), distress (Carlson et al, 2004), and depression (Danhauer et al, 2008). Previous research shows that yoga is associated with improvements in the overall quality of life of cancer patients (Culos-Reed, Carlson, Daroux & Hately-Aldous, 2006; Danhauer et al, 2008; Moadel et al, 2007) including improved emotional well-being and physical outcomes such as sleep quality, mood, and stress (Bower, 2008).

Randomized controlled trials of yoga programs revealed the positive impact of yoga on psychological outcomes (Smith & Pukall, 2008). Specifically, Cohen, Varneke, Fouladi, Rodriguez and Chaoul-Reich (2004) found significant improvement in sleep measures, and Culos-Reed et al (2006) reported significant improvement in quality of life and decreased levels of stress. Findings from non-controlled trials support evidence that practicing yoga during cancer treatment results in increased energy (Carson et al, 2007) and quality of life (Carlson, Speca, Patel & Goodey, 2004).

Despite positive findings, many Hispanic women are not using CAM therapy, including yoga, as readily as non-Hispanic Caucasian women (Fouladbakhsh & Stommel, 2010). Numerous barriers to accessing health care exist for diverse populations including cultural beliefs, language, and lack of knowledge about the importance of cancer screening (del Carmen, 2009; Medina, 2010; Watts et al., 2009). According to Medina (2010), Latina women are "likely to feel disempowered during an encounter with their provider" (p. 75) resulting in delays in seeking medical care. Often, Latina women do not seek appointments with a doctor due to lack of money, time, childcare, or because they do not have symptoms requiring health care (Byrd, Peterson, Chavez & Heckert, 2004; McMullin, De Alba, Chavez & Hubbell, 2005; Zambrana, Breen, Fox & Gutierrez-Mohamed, 1999). Barriers to seeking treatment, coupled with unfamiliarity with additional resources such as yoga, results in a decrease in the likelihood of vulnerable populations using CAM therapies during cancer treatment. In a recent study, Foulabakhsh and Stommel (2010) report "African American and Hispanic cancer survivors have substantially lower odds of engaging in a CAM practice than non-Hispanic Caucasians" (p. E10).

A recent review of the empirical literature on the health-related quality of life of Latina breast cancer survivors reveals a dearth of information across intrapersonal, community, and institutional levels (Lopez-Class, Gomez-Duarte, Graves and Ashing-Giwa, 2011). When ethnic breakdown is included in studies of the effects of yoga on cancer (e.g., Carson et al, 2007; Danhauer et al, 2008; Moadel et al, 2007; Warner, 2007) the majority of participants are typically Caucasian (e.g., Carson et al, 2007; Danhauer et al, 2008; Warner, 2007). Published multiethnic studies of yoga and cancer survivors reveal beneficial effects of yoga (Moadel et al, 2007) and call for additional research with vulnerable populations.

Increasingly, studies are examining the effects of yoga with diverse cancer populations (e.g., Moadel et al, 2007), but none explore the effects of yoga on patients while they wait for cancer treatment. The purpose of this research is to describe and evaluate an existing yoga program for vulnerable adults with cancer. The people participating in this on-going program do not usually have access to yoga classes as they are low-income, typically live in communities where yoga is not readily available or offered in their first language, and are in treatment for cancer with all of the time constraints and side effects inherent in dealing with disease treatment. Taking advantage of the time people have already set aside for cancer treatment and offering bilingual yoga classes at their cancer treatment site reduces numerous barriers to participation including time, language, and childcare. The program makes yoga accessible to vulnerable populations and consists of two components: 1) on-site hospital yoga classes for low-income adults currently receiving treatment for cancer, and 2) a community yoga class offered by the same provider focused on serving cancer survivors. Hospital and community yoga classes are bilingual (Spanish/English) serving a primarily Hispanic population.

The specific objectives of this evaluation are to 1) explore participants' reasons for attending yoga classes and their thoughts and feelings about the yoga classes, and 2) compare differences in the reported quality of life between hospital patients who participate in a yoga class, hospital patients who do not participate in a yoga class, and community yoga class members. Specifically, patients attending a one-hour yoga class while waiting for treatment

at the hospital are hypothesized to report a higher quality of life than hospital patients who choose not to participate in a yoga class. Further, community yoga class members with cancer in remission are hypothesized to report a higher quality of life than hospital patients currently undergoing treatment for cancer.

Methods

Description of Agency/Intervention

The Agency is a nonprofit 501(c)(3) organization whose mission is "to provide free exercise and fitness opportunities for adults living with cancer …many of our programs address the needs of medically underserved or low-income women" (accessed from agency website, January 10, 2012). Grants and foundations are the main source of support for this small agency serving approximately 300 adults annually in Southern California. The Agency provides two weekly yoga classes to cancer patients waiting for treatment at a County Hospital run by the California Department of Health Services serving the needs of society's most vulnerable. Patients waiting to see the doctor for their cancer treatment are invited to participate in a yoga class adjacent to the waiting room.

The yoga classes provided by the Agency are held in the nurses' break room across the hall from the treatment waiting room. Two yoga instructors personally invite patients to attend the class. The yoga instructors are Hispanic women in their mid-30s, bilingual (Spanish/English), and certified with more than 200-hours of yoga-specific training. One yoga instructor teaches the class while the other remains in the waiting room listening for the names of class attendees called by nurses for their appointment. In this manner, patients do not miss their appointment, treatment, or paperwork, and are still able to participate in the yoga class. The class is chair-based and focuses on breathing and gentle stretching. Participants do not need special clothing or equipment and are encouraged to participate to the fullest extent possible with modifications offered for activities deemed too difficult by the patient.

In addition to the two weekly County Hospital yoga classes, the Agency provides one weekly community yoga class in collaboration with another not-for-profit cancer organization. Unlike the hospital class(es), the community class is a much more traditional yoga class in that participants are regular attendees, yoga occurs on mats on the floor as opposed to being chair-based, and participants have a connection to cancer as either a survivor or caretaker. At the time of the evaluation none of the community class participants were actively in treatment for cancer.

Participants

Study participants (n=38) included a convenience sample of adult patients seeking cancer treatment at a County Hospital in November 2011 and January 2012 (n=26) and all class participants who attended a community-based yoga class on January 11, 2012 (n=12). The hospital is located on the outskirts of a large urban area and "serves the needs of low income and indigent patients as well as the surrounding middle class community" (accessed from the hospital website, February 26, 2012) with approximately 50% of the patients identifying as Hispanic. Patients who chose to attend a yoga class (n=14) were compared to a group of patients waiting for their treatment appointment who opted not to participate in a yoga class (n=12). The community-based class is taught by the same instructor who teaches the hospital class(es). The community class is located in a suburb of a large urban area. Residents are primarily Hispanic (40%) or Caucasian (36%) and 14.3% live below the poverty level.

Design and Procedure

For hospital patients, a posttest-only design with nonequivalent groups was used to assess the effects of yoga on their quality of life (Rubin & Babbie, 2011). The quality of life of patients who chose to attend the yoga class was compared to patients who did not attend the class as measured by the FACT-G, and yoga class attendees were also asked to complete an Agency-specific outcome questionnaire. Surveys were completed by hospital class attendees at the end of class, and data were collected from the comparison group while they waited for their treatment appointment. Informed consent was obtained from all patients, and participants received a $5 gift card in appreciation for their time completing the questionnaires. Like Hospital class attendees, community class members were asked to complete the FACT-G and the Agency outcome questionnaire after their class. All data were collected anonymously. Prior to data collection the study received approval from the California State University, Northridge Institutional Review Board.

Measures

Quality of Life. All participants completed the Functional Assessment of Cancer Therapy – General (FACT-G), a widely used health-related quality of life survey. The FACT-G is a 28-item self-report measure of quality of life in cancer patients measuring four dimensions: physical well-being, social/family well-being, emotional well-being, and functional well-being (Cella, Tulsky, & Gray et al, 1993) and has established reliability and validity. The FACT measures are available for multiple conditions and illnesses. The option existed to use illness-specific FACT scales (e.g. the FACT-B, a breast cancer focused quality of life measure), but a decision was made to use only the FACT-G to 1) stay true to the mission of the Agency to serve all types of cancers, and 2) reduce the amount of time patients spent filling out questionnaires.

Demographics. Age, ethnicity, and gender were assessed to describe the participants. Patients were not asked about their type of cancer, and all hospital participants were receiving individualized cancer treatment.

Agency Survey. Four years ago the Agency designed and implemented an annual agency-specific satisfaction survey that focused on collecting data used to report client outcomes to funders. The survey includes three sections: Attendance, Thoughts and Feelings about the Yoga Class, and Demographics. The survey is comprised of 11 questions: four attendance questions, five thoughts and feelings questions, and two demographic questions (age and ethnicity). Attendance questions included 1) how did you hear about the yoga class?, 2) how many times have you attended?, 3) would you come back?, and 4) why did you decide to attend the class? Thoughts and Feelings Questions included 1) how has this yoga class affected your treatment and recovery?, 2) how important is this yoga class to you?, 3) what skills have you learned from this yoga class?, 4) would you recommend this class to your friends and family?, and 5) what does attending this class give you that you haven't found anywhere else? The agency survey has not been tested for reliability or validity.

Data Analysis

Data were analyzed using the Statistical Package for the Social Sciences (SPSS) Version 19. Data analyses include descriptive statistics and Chi-Square bivariate analyses to explore the

Agency's program outcomes specific to client's reasons for attending the yoga class(es) and their thoughts and feelings about the yoga class(es). A Mann-Whitney U Test for differences between two independent groups was used to test the hypothesis that patients attending a one-hour yoga class while waiting for treatment in the hospital will report a higher quality of life than patients receiving their usual course of treatment with no yoga class. A Kruskal-Wallis Test for differences between three independent groups was used to compare the quality of life scores of hospital patients who took a yoga class, hospital patients who did not take a yoga class, and community yoga class participants not currently in treatment for cancer.

Results

Demographic information was self-reported by respondents. There was no significant difference between the hospital class participants, non-participants and community class members on age, ethnicity, or gender. The mean age of the sample was 49.8 (SD=14.58) years and the majority of study participants are Hispanic women (57%).

Exploring Agency Outcomes

Chi-square tests were utilized to compare differences between the hospital class attendees and community class attendees on agency outcome variables. The majority of hospital class attendees heard about the class from the instructor/recruiter (χ^2 (1, n = 26) = 15.80, p = .00, phi = -.86) while the community class participants were more likely to report hearing about the class from a friend. As expected, hospital class participants are more likely to be attending class for the first time or only one other time whereas community class members report attending two or more times. 100% of the hospital respondents said they would return to take another yoga class if they have a doctor's appointment scheduled. As seen in Table 1, relaxation is the biggest reason participants decided to attend the yoga class.

Participants were asked how the yoga class affected their cancer treatment and recovery. As seen in Table 1, hospital and community participants report that they are more relaxed and manage stress better. The majority of participants reported breathing as a skill they learned that they could use at home. 100% of respondents said they would recommend the class to their friends and family.

Table 1. Frequency (%) and Mean (SD) of Demographics and Agency Outcome Variables by Group

	Hospital Class Participant (n=14)	Hospital Non-Participant (n=12)	Community Class Participant (n=12)	p
Ethnicity				
African American	0	2 (5%)	0	
Asian American	1 (3%)	2 (5%)	1 (3%)	
Caucasian	1 (3%)	3 (7%)	2 (5%)	
Hispanic	12 (32%)	5 (13%)	9 (24%)	
Gender				
Male	1 (4%)	4 (14%)	1 (4%)	
Female	13 (46%)	8 (28%)	1 (4%)	
Age	45.8(13.8)	53.1 (9.9)	50.8 (18.9)	.38
Quality of Life Score	64.9 (14.2)	65.1 (15.6)	85.2 (25.5)	.07
Heard About Class?*				
Recruiter	12 (86%)		0	.00
Friend	0		5 (42%)	
Flyer	0		2 (17%)	
Doctor	0		1 (8%)	

Hospital	2 (14%)		0	
Newspaper	0		3 (25%)	
Times Attended				
First Time	3 (21%)		3 (25%)	
1 Time	7 (50%)		0	
2-3 Times	3 (21%)		3 (25%)	
3+ Times	1 (8%)		6 (50%)	
Why Attend?*				
Something to do	4 (31%)		1 (8%)	
For Relaxation	9 (69%)		9 (75%)	
Help with Cancer	4 (31%)		3 (25%)	
To Feel Better	5 (39%)		2 (25%)	
My Friends Come	1 (8%)		0	
Current Ability	2 (23%)		4 (33%)	
Affordable	1 (8%)		4 (33%)	
How Affected Tx?*				
Physically Stronger	1 (8%)		5 (42%)	
Less Tired	2 (15%)		7 (58%)	
More Flexible	4 (31%)		7 (58%)	.32
Mentally Stronger	5 (39%)		4 (33%)	
More Relaxed	10 (77%)		9 (75%)	
Manage Stress	5 (39%)		6 (50%)	.85
Skills Learned?*				
Breathing	13 (100%)		11 (92%)	
Meditation	8 (62%)		8 (67%)	
Stretching	9 (69%)		7 (58%)	
Exercise	4 (31%)		8 (67%)	.16

Note: p value reported only if assumptions of Chi-Square were met (minimum expected cell count greater than 5). *respondents were asked to mark all that apply.

Comparing Quality of Life Scores

Results of a Mann-Whitney U test revealed no significant difference in quality of life between yoga class participants (Md=63.91, n=14) and non-participants (Md=65.01, n=12), U=67.50, z=-.26, p=.79, r=-.05 in the hospital. Hospital yoga participants do not report a higher quality of life than non-participants. Despite community class participants' markedly higher mean quality of life scores than hospital participant and non-participant scores, a Kruskal-Wallis Test revealed no statistically significant difference in quality of life between the hospital non-participants, hospital class participants, and community class members (Gp1, n=12: non-participants, Gp2, n=12: hospital class participants, Gp3, n=8: community class), $\chi^2(2, n=32) = 5.35$, p=.07. The median FACT-G score for hospital non-participants is 65.0, for hospital class participants is 63.9, and for community class members is 86.6.

Discussion

Although class participants reported learning new skills and high levels of satisfaction with the services they received, there was no significant difference found in the quality of life between yoga class participants and non-participants in the hospital. Previous research reveals support for the effectiveness of yoga in improving quality of life of cancer patients (Danhauer et al, 2008; Moadel et al, 2007; Rosenbaum, Gautier, & Fobair et al, 2004). However, previous studies of the effectiveness of yoga in reducing symptoms detailed interventions lasting from 30 minutes to two hours with home practice encouraged (Banerjee et al, 2007; Carson et al, 2007; Culos-Reed et al, 2006; Raghavendra et al, 2007) with the

most rigorous studies carefully implementing an on-going 6-8 week group (Carson et al, 2007; Culos-Reed et al, 2006). Although evidence exists for the positive effects for as few as two classes per week (e.g., Rosenbaum et al, 2004), it is likely that a single yoga class is simply not enough time to effect improvement in a cancer patients' quality of life.

Fawzy & Fawzy (1998) found that group interventions with cancer patients result in positive outcomes including improved quality of life. Previous literature reports evidence of the effectiveness of psychotherapy for cancer patients in groups (Boynton & Thyer, 1994; Kissane et al, 2003; Trijusburg, van Knippenberg & Rijpma, 1993), but published studies do not detail differences between on-going groups and one-time classes that may affect patient outcomes. The Agency's hospital yoga classes differ from traditional yoga classes in that participants vary widely from week to week and thus have not bonded the way participants of a more traditional support group or yoga class might. Future research should explore potential group effects of an active, experiential group such as yoga versus more traditional, on-going support groups.

Previous research found that yoga can effectively reduce the negative symptoms and side effects of cancer treatment, but little research has been done with vulnerable populations who usually do not have access to yoga classes. The results of this study show that diverse populations appreciate the experience of the yoga classes, but results do not conclusively show that yoga positively affects quality of life.

Limitations

There is a marked difference in reported quality of life between the community class members and hospital patients. Although not statistically significant, community members report better quality of life than do hospital patients in treatment for cancer. Proponents of yoga might argue that the community class having a significantly higher quality of life than hospital patients is a direct result of the yoga class those participants attend weekly. The design of this study does not allow for that argument; indeed a greater likelihood is that community class attendees have been out of treatment longer and thus are not as symptomatic as are cancer patients still receiving treatment. Participants delight in reporting the benefits of yoga, and while it is possible that regular attendance in a yoga class helped these community class members recuperate from their cancer, the design of the study does not allow for a causal inference regarding the effectiveness of the Agency yoga class on cancer symptom reduction or quality of life improvement.

Since an experimental design is not employed it cannot be stated unequivocally that the Agency hospital yoga program is effective. However, previous Agency program evaluations examined only participant's views of the program and their self-assessed outcomes, so the current evaluation improves upon the previous design by surveying non-participants' quality of life to compare between groups. While including a standardized measure of quality of life is also an improvement on previous evaluations, using an agency-designed outcome measure that has not been tested for reliability and validity limits the credibility of the results. Previous Agency evaluations report excellent outcomes and high satisfaction, and these results concur with previous findings. Although participants happily report their satisfaction, the measure used may be designed to elicit just such a positive response.

Conclusion

Despite, or perhaps because of, the limitations of this evaluation, it serves as an excellent example of real-world research and the challenges inherent in implementing, translating, and evaluating evidence-based practices in regular agency settings (Proctor & Rosen, 2008). The

Agency measure of outcome (i.e., number of skills learned) directly corresponds with the goals set in response to funding requests. By reporting that at least 80% of participants learned skills and are satisfied with their service, the Agency continues to secure grant funding and foundation support to maintain and expand their program. In this manner, the organization continues to provide yoga classes for vulnerable adults with cancer; classes that are welcomed and appreciated, if not measurably effective.

Acknowledgements

This work was supported in part by a grant from the College of Behavioral and Social Sciences at California State University, Northridge. The author would like to thank the Agency where the research occurred and two anonymous reviewers for their helpful feedback.

References

Banasik, J., Williams, H., Haberman, M., Blank, S.E., & Bendel, R. (2011). Effect of Iyengar yoga practice on fatigue and diurnal salivary cortisol concentration in breast cancer survivors. *Journal of the American Academy of Nurse Practitioners,* 23, 135-142.

Banerjee, B., Vadiraj, H.S., Ram, A.,…& Hande, M.P. (2007). Effects of an integrated yoga program in modulating psychological stress and radiation-induced genotoxic stress in breast cancer patients undergoing radiotherapy. *Integrative Cancer Therapies,* 6, 242-250.

Bower, J.E. (2008). Behavioral symptoms in breast cancer patients and survivors: fatigue, insomnia, depression and cognitive disturbance. *Journal of Clinical Oncology, 26*(5), 768-777.

Boynton, K.E. & Thyer, B.A. (1994). Behavioral social work in the field of oncology. *The Journal of Applied Social Sciences, 18*(2), 189-197.

Brauer, J.A., El Sehamy, A., Metz, J.M, & Mao, J.J. (2010). Complementary and alternative medicine and supportive care at leading cancer centers: A systematic analysis of websites. *The Journal of Alternative and Complementary Medicine, 16*(2), 183-186.

Buettner, C., Kroenke, C.H., Phillips, R.S., Davis, R.B., Eisenberg, D.M., Holmes, M.D. (2006). Correlates of use of different types of complementary and alternative medicine by breast cancer survivors in the nurses' health study. *Breast Cancer Research Treatment*, 100, 219-227. DOI 10.1007/s10549-006-9239-3.

Byrd, T. L., Peterson, S. K., Chavez, R., & Heckert, A. (2004). Cervical cancer screening beliefs among young Hispanic women. *Preventive Medicine, 38*(2), 192-197.

Carlson, LE., Speca, M., Patel, K.D. & Goodey, E. (2004). Mindfulness-based stress reduction in relation to quality of life, mood, symptoms of stress, and levels of cortisol, dehydroepiandrosteronesulfate (DHEAS) and melatonin in breast andprostate cancer outpatients. *Psychoneuroendocrinology,* 29, 448-474.

Carlson, L.E., Angen, M., Cullum, J., Goodey, E., Koopmans, J…& Bultz, B.D. (2004). High levels of untreated distress and fatigue in cancer patients. *British Journal of Cancer,* 90, 2297-2304. DOI: 10.1038/sj.bjc.6601887.

Carson, J.W., Carson, K.M., Porter, L.S., Keefe, F.J., Shaw, H., & Miller, J.M. (2007). Yoga for women with metastatic breast cancer: results from a pilot study. *Journal of Pain Symptom Management,* 33, 331-341.

Cella D, Tulsky D, Gray G, et al. (1993). The functional assessment of cancer therapy (FACT) scale: Development and validation of the general version. *Journal of Clinical Oncology, 11*(3), 570–579.

Cohen, L. Varneke, C. Fouladi, R.T., Rodriguez, M.A., & Chaoul-Reich, A. (2004). Psychological adjustment and sleep quality in a randomized trial of the effects of a Tibetan Yoga intervention in patients with lymphoma. *Cancer*, 100, 2253-2260.

Culos-Reed, S.N., Carlson, L.E., Daroux, L.M., and Hately-Aldous, S. (2006). A pilot study of yoga for breast cancer survivors: Physical and psychological benefits. *Psycho-Oncology*, 15, 891-897.

Danhauer, S.C., Mihalko, S.L., Russell, G.B., Campbell, C.R., Felder, L., Daley, K., & Levine, E.A. (2009). Restorative yoga for women with breast cancer: Findings from a randomized pilot study. *Psycho-Oncology*, 18, 360-368.

del Carmen, M. G. (2009). The burden of cervical cancer in minority populations: Effective strategies in reducing disparity. *The Internet Journal of Gynecology and Obstetrics*. Retrieved September 10, 2011 from http://go.galegroup.com/ps /i.do?action=interpret&id=GALE|A197106122&v=2.1&u=csunorthridge&it=r&p=ITOF &sw=w&authCount=1

Dorval, M., Maunsell, E., Deschenes, L., Brisson, J., Masse, B. (1998). Long-term quality of life after breast cancer: Comparison of 8-year survivors with population controls. Journal of Clinical Oncology, 16, 487-494.

Elkins, G., Rajab, M.H., & Marcus, J. (2005). Complementary and alternative medicine use by psychiatric inpatients. *Psychological Reports*, 96, 163-166.

Fawzy, F. & Fawzy, N. W. (1998). Group therapy in the cancer setting. *Journal of Psychosomatic Research,* 45, 191-200.

Fouladbakhsh, J.M. & Stommel, M. (2008). Comparative analysis of CAM use in the U.S. cancer and noncancer populations. *Journal of Complementary & Integrative Medicine,* 5, Article 19. doi:10.2202/1553-3840.1140.

Fouladbakhsh, J.M. & Stommel, M. (2010). Gender, symptom experience, and use of complementary and alternative medicine practices among cancer survivors in the U.S. cancer population. *Oncology Nursing Forum, 37*(1), E7-E15.

Kissane, D.W., Bloch, S., Smith, G.C., Miach, P., Clarke, D.M., Ikin, H., et al (2003). Cognitive-existential group psychotherapy for women with primary breast cancer: A randomized controlled trial. *Psycho-Oncology*, 12, 532-546.

Kvillemo, P. & Branstrom, R. (2011). Experiences of a mindfulness-based stress-reduction intervention among patients with cancer. *Cancer Nursing, 34*(1), 24-31.

Lancaster, J. (2008). From the editor. Family & Community Health, 31(3), 187.

Lopez-Class, M., Gomez-Duarte, J., Graves, K., & Ashing-Giwa, K., (2011). A contextual approach to understanding breast cancer survivorship among Latinas. *Pyscho-Oncology*, DOI: 10.1002/pon.1998.

Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, and Brown ML. Projections of the Cost of Cancer Care in the United States: 2010-2020. Jan 19, 2011, JNCI, Vol. 103, No. 2.

Medina, R. (2010). Cervical cancer and Latinas: A preventable disease. *Harvard Journal of Hispanic Policy, 22*, 73-78.

McMullin, J. M., De Alba, I., Chavez, L. R., & Hubbell, F. A. (2005). Influence of beliefs about cervical cancer etiology on Pap smear use among Latina immigrants. *Ethnicity & Health, 10*(1), 3-18.

Moadel, A.B., Shah, C., Wylie-Rosett, J., Harris, M.S., Patel, S.R., Hall, C.B., & Sparano, J.A. (2007). Randomized controlled trial of yoga among a multiethnic sample of breast cancer patients: Effects on quality of life. *Journal of Clinical Oncology, 25*(28), 4387-4395.

Pritchard, M. Elison-Bowers, P., & Birdsall, B. (2010). Impact of integrative restoration (iRest) meditation on perceived stress levels in multiple sclerosis and cancer outpatients. *Stress and Health*, 26, 233-237.

Proctor, E. & Rosen, A. (2008). From knowledge production to implementation: Research challenges and imperatives. *Research on Social Work Practice, 18*(4), 285-291.

Raghavendra, R.M., Nagarathna, R., Nagendra, H.R…& Nalini, R. (2007). Effects of an integrated yoga programme on chemotherapy-induced nausea and emesis in breast cancer patients. *European Journal of Cancer Care*, 16, 462-474.

Rosenbaum, E., Gautier, H., Fobair, P., Neri, E., Festa, B., Hawn, M., Andrews, A., Hirshberger, N., Selim, S. & Spiegel, D. (2004). Cancer supportive care, improving the quality of life for cancer patients. A program evaluation report. *Support Care Cancer*, 12, 293-301.

Rubin, A. & Babbie, E. (2011). Research methods for social work. New York, NY: Brooks/Cole.

Smith, K.B. & Pukall, C.F. (2009). An evidence-based review of yoga as a complementary intervention for patients with cancer. *Psycho-Oncology*, 18, 465-475.

Trijsburg, R., van Knippenberg, F.C., & Rijpma, S.E. (1992). Effect of psychological treatment on cancer patients: A critical review. *Psychosomatic Medicine*, 54, 489-517.

Warner, A.S. (2006). Exploration of psychological and spiritual well-being of women with breast cancer participating in the Art of Living program. (Doctoral dissertation pub #: 3227466).

Watts, L., Joseph, N., Velazquez, A., Gonzalez, M., Munro, E., Muzikansky, A.,…del Carmen, M. G. (2009). Understanding barriers to cervical cancer screening among Hispanic women. *American Journal of Obstetrics & Gynecology, 201*(2), 199.e1-e8.

Yates, J.S., Mustian, K.M., Morrow, G.R., Gillies, L.J., Padmanabah, D., Atkins, J.N., Issell, B., Kirshner, J.J.,& Colman, L.K. (2005). Prevalence of complementary and alternative medicine in use in cancer patients during treatment. *Supportive Care in Cancer*, 13, 806-811.

Zambrana, R. E., Breen, N., Fox, S. A., & Gutierrez-Mohamed, M. L. (1999). Use of cancer screening practices by Hispanic women: Analyses by subgroup. *Preventive Medicine 29*(6), 466-477.

Evidence based development of clinical learning environment in Finnish health care services

Riitta Meretoja, RN, PhD, Adjunct Professor, Hospital District of Helsinki and Uusimaa, University of Turku, Finland, riitta.meretoja@hus.fi
Mikko Saarikoski, RN, PhD, Adjunct Professor, Turku University of Applied Sciences, University of Turku, Finland, mikko.saarikoski@turkuamk.fi

Abstract

Studying during clinical practice is an essential part of professional health care education. Systematic quality evaluation of clinical learning environments started in Finland from Hospital District of Helsinki and Uusimaa in 2007, and the used evaluation model was soon adopted for national use in order to have a national benchmarking data. The evaluation tool includes background variables and Clinical Learning Environment, Supervision and Nurse Teacher scale (Saarikoski & Leino-Kilpi 2002, Saarikoski et al. 2008). The national data was collected in 2010 from 17 hospital districts and community units consisting of 10342 students. Findings revealed that the respondents were generally satisfied with their clinical placements. Although the overall level of quality of clinical supervision was assessed to be at good levels there were differences between organisations. Results offer good initiatives to the units to increase the quality elements of their evaluation process.

Keywords: *learning environment, nursing student, health care services, evaluation studies*

Introduction

Nurse educational systems have undergone a remarkable transformation in Finland as the locus for educational programmes moved from the vocational college systems to provisions in higher education institutions. It was not until the end of 1990's that European Ministers of Education (1999) agreed in Bologna the principles required to underpin a universal approach to higher education and research across Europe. There are still many differences between the countries in executing these educational reforms both in theoretical and practical studies of nurse education (Salminen et al. 2010).

Studying during clinical practice is still an essential part of professional health care education. It is also one of the most important cooperation forms between educational organisations and health care services. (Barrett 2007.) There is research evidence that a high quality learning environment and an individualized supervision system are the most important single quality factors in students' clinical learning experiences (Saarikoski & Leino-Kilpi 2002, Hosoda 2006, Midgley 2006, Johansson et al. 2010).

Origins for the evidence based research of clinical learning environment

The educational outcomes for student nurses depend, in part, upon the quality of the teaching and learning environment provided in clinical settings. The educational practice requires a process for monitoring and evaluating the quality of student nurses' clinical placements. A literature search for instruments evaluating student nurses' clinical learning experiences identified a limited number of instruments developed in Europe. They are mainly developed in a cultural environment locating e.g. in Austral-Asian or in Northern America areas (Chan

2002, 2003, Dunn & Burnett 1995, Hosoda 2006). The only widely reported instrument developed in Europe is the Clinical Learning Environment, Supervision and Nurse Teacher (CLES+T) scale (Saarikoski et al. 2008).

In the critical review of the instrument development articles, the CLES+T scale was culturally suitable to Finnish health care environments. The articles (Saarikoski & Leino-Kilpi 2002, Saarikoski et al. 2008, Saarikoski et al. 2009) reported the adequate features of the whole validation process: sample size, robust internal reliability, rationales for decisions made during psychometric analysis, international comparisons to validate the tool beyond the initial sample, and concurrent validity. Additionally, the CLES+T scale has been validated also in New Zealand (Sims et al. 2010, Watson et al. 2012), Sweden (Johansson et al. 2010, Bos et al. 2012), Belgium (De Witte et al. 2011), Norway (Skaalvik et al. 2011) and in Italy (Tomietto et al. 2012).

Systematic, evidence-based evaluation of the quality of clinical learning environment by using the CLES scale (Saarikoski & Leino-Kilpi 2002) and its later CLES+T version (Saarikoski et al. 2008), started in Finland in the Hospital District of Helsinki and Uusimaa in 2007. The scale was firstly adapted for national use in order to have a national benchmarking data for quality evaluation and its development. The aim of this quality assessment is to improve the quality of clinical learning environment and clinical supervision of nursing students during their clinical placements. The data can be used in decision making for the purposes of developing the outcomes of students' learning and clinical units' supervisory activities. The quality of learning environment is also identified as a crucial element when recruiting new staff to health care organisations.

Methods

Study design and research questions

The survey type study explored how nursing students studying in 14 Finnish hospital districts' hospitals and in three community unit hospitals perceived their clinical learning environments, the supervisory relationship with the personal supervising staff nurse and the level of intervention with the nurse teacher during their clinical placements. In Finland a personal supervising nurse, working in a nursing team, is named to be responsible for supervision during the clinical placement period. The overall aim of the study was to provide a view how Finnish health care services facilitated students' clinical learning during their clinical placements.

The research questions were:
(1) How nursing students experience their clinical learning environment,
(2) the supervision provided by their personal supervisor nurses and
(3) the level of intervention with their nurse teacher.

Data collection and ethical issues

The national data for this study was collected between the 1st of January and 31st of December 2010 from 14 hospital districts' and three community unit hospitals. The data was collected by using Internet based data collecting tools at the end of the nursing students' clinical placements. The cover letter included information about the purpose of the survey, issues related to research ethics, and how to access the electronic questionnaire. This cover letter acts as an information letter to participants and contained enough detailed information to allow students to make an informed choice over giving consent to participate in the study or not.

Research instrument and data analyses

The adopted benchmarking version of the CLES+T evaluation scale is basing on an extensive review of empirical studies (n=87) and learning environment audit instruments (n=6) published between 1980 and 2006 (Saarikoski 2002, Saarikoski et al. 2009). The national consensus group of experts from hospital districts and community units made minor revisions and edited some terms of the internationally validated CLES+T evaluation scale so that the scale was applicable in all health care education programs (nursing, physiotherapy, radiography etc.) and usable as benchmarking data for quality evaluation and development. There are 35 items in the CLES+T scale divided to five sub-dimensions: Atmosphere on the ward/ 7 items; (2) Premises of learning on the ward/ 7 items; (3) Premises of nursing care on the ward/ 4 items; (4) Supervisory relationship/ 8 items and (5) Role of nurse teacher/ 9 items.

The students assessed the quality of supervision with an electronic questionnaire by using 10-point scale (totally disagree – totally agree). The national data base is gathered yearly for national comparisons (e.g. differences between the hospitals). Every participating organization can also use their own data for their own analyses in developing the quality of learning environment and supervision system in their units.

The data was analysed using descriptive statistics (frequency, mean and standard deviation). The statistical analyses have been undertaken mainly using the results of whole sample, not by the hospitals. The internal consistency reliability of CLES+T scale has been analysed using Cronbach's alpha coefficient. The alpha coefficients varied from 0.83 to 0.96 by the sub-dimensions.

Results

Sample characteristics

A total of 10 342 students returned a completed questionnaire. Most of the respondents were studying in the five Finnish university hospital district' hospitals (n=7030/ 68.0 %), the rest in nine non-university hospital districts' hospitals (n=2264/ 21.9 %) or three community unit hospitals (n=1048/ 10.1 %). The majority of the respondents were aged 20-29 years (71.5 %). Three quarters of the respondents were students of second or third term (77.4 %). In addition, a total of 6747 students studying in four university hospital districts' hospitals, six non-university hospital districts' hospitals and three community units completed the teacher subscale.

The quality of clinical learning environments

Majority of the respondents were very satisfied with the achievement of their own learning goals and felt that supervision supported their professional development. However, the respondents were quite critical how their earlier theoretical nursing studies supported their learning during their clinical placements (Table 1). Majority of the students would very positively or positively recommend their clinical placement unit to his/her study mates (85.7 %).

Table 1. Satisfaction to own learning during clinical placements (n= 10342)

Item	Very well %	Fairly well %	Moderately %	Quite poorly %	Very poorly %
How well you achieved your learning goals during the	44.8	48.5	5.7	0.9	0.2

placement?					
How well supervision you received supported your professional development?	54.4	34.6	8.4	2.1	0.7
How well the theoretical studies supported your learning in clinical practice?	16.2	45.9	28.0	8.8	1.1

The results revealed that the respondents were generally satisfied with their clinical learning environment and they evaluated it to be at a good level. Although the overall quality of clinical supervision was assessed to be at a good level, there were differences between hospital districts and community units. Supervisory relationship and Premises of learning were assessed to be at the highest level. On the other hand Premises of nursing and Atmosphere in the work place were assessed to be at a little lower level. According to the respondents, the highest rated items were "The mentor showed a positive attitude towards supervision", "Mutual respect and approval prevailed in the supervision relationship", "I felt that I received individual supervision". The lowest rated items were "Feedback from ward manager could easily be considered as a learning situation" and "The staffs were generally interested in student supervision". (Table 2)

Table 2. Assessments of the quality of clinical learning environment

CLES SUBSCALES	Mean n = 10342	Range between organizations n=17
ATMOSPHERE SUBSCALE	**8.2**	**7.5 – 8.7**
The staffs were easy to approach	8.4	7.9 – 9.0
During staff meetings I felt comfortable taking part in the discussions	8.0	6.0 – 8.8
I felt comfortable going to the ward at the start of my shift	8.7	8.2 – 9.4
There was a good positive atmosphere on the ward	8.3	7.8 – 8.7
The staff was regarded as a key resource on the ward	8.3	7.8 – 8.7
The effort of individual employees was appreciated	8.2	7.7 – 8.5
Feedback from ward manager could easily be considered as a learning situation	7.2	6.6 – 7.7
PREMISES OF NURSING SUBSCALE	**8.3**	**8.0 – 9.5**
The value base of patient care was clearly defined	8.2	7.7 – 9.4
Patients received individual care	8.7	8.3 – 9.8
Documentation of patient care was clear	8.3	7.8 – 9.5
There was no problems in the information flow related to patient care	7.9	7.5 – 9.2
PREMISES OF LEARNING SUBSCALE	**8.3**	**7.9 – 8.7**
Basic familiarization was well organized	8.1	7.4 – 8.8
The staffs were generally interested in student supervision	7.4	6.7 – 8.1
The staff learned to know the student by their personal name	8.1	7.3 – 8.4
Patient cases were used in my supervisory process.	8.6	8.1 – 9.0
There were sufficient meaningful learning situations on the ward	8.6	8.1 – 8.9
The learning situations were multi-dimensional in terms of content	8.5	8.3 – 9.0
My supervisors supervision skills supported my learning	8.7	8.2 – 9.3
SUPERVISORY RELATIONSHIP SUBSCALE	**8.5**	**8.0 – 9.2**
The mentor showed a positive attitude towards supervision	8.9	8.2 – 9.6
I felt that I received individual supervision	8.8	8.2 – 9.5

I continuously received feedback from my mentor	7.7	7.0 – 8.4
Overall I am satisfied with the supervision that I received	8.7	8.1 – 9.5
The supervision was based on a relationship of equality and promoted my learning	8.6	8.0 – 9.3
There was a mutual interaction in the supervision relationship	8.7	8.3 – 9.4
Mutual respect and approval prevailed in the supervision relationship	8.8	8.1 – 9.6
The supervisory relationship was characterized by a sense of trust	8.2	7.6 – 8.9
Overall mean of items	**8.3**	**7.9 – 8.8**

The respondents evaluated their nurse teachers' role during their clinical placements mainly positively but clearly at a lower level (overall mean 7.0) than the learning environment subscales (overall mean 8.3). Highest rated items were related to categories of "Relationship among student, mentor and teacher" and "Teacher enabling integration of theory and practice". The respondents were most critical on items related to the cooperation between clinical placement and nurse teacher. (Table 3)

Table 3. Role of the teacher during clinical placements

CLES - TEACHER SUBSCALE	Mean n=6747	Range between organisations n=13
Teacher enabling integration of theory and practice	**7.3**	**6.9 – 7.7**
In my opinion, the teacher was capable of integrating theoretical knowledge and everyday practice of patient care	7.4	7.1 – 7.8
The teacher was capable of operationalising the learning goals of this placement	7.5	7.0 – 7.9
The teacher helped me to reduce the theory-practice cap	7.1	6.5 – 7.5
Cooperation between clinical placement and teacher	**5.9**	**5.5 – 6.7**
The teacher was like a member of the nursing team	5.2	4.6 – 6.2
The teacher was able to give his or her expertise to the clinical team	5.5	4.7 – 6.2
The teacher and the clinical team worked in supporting my learning	7.1	6.6 – 7.7
Relationship among student, mentor and teacher	**7.6**	**7.1 – 8.1**
The common meetings between myself, my supervisor and the teacher were comfortable experience	7.5	7.0 – 8.1
In our common meetings I felt that we are colleagues	7.4	6.6 – 7.9
Focus on the meetings was in my learning needs	7.9	7.3 – 8.3
Overall mean of items	**7.0**	**6.5 – 8.0**

Discussion

The CLES+T evaluation instrument has proven to be a good instrument at measuring the quality of clinical training during the clinical placements and the students have been very motivated to give feedback with this instrument. The findings of this study revealed that the respondents were generally satisfied with their clinical placements. Even though the differences between the hospitals have been rather small, still the results offer good initiatives to organizations to increase the quality elements of their supervision and learning environments. The organizational commitment of using this instrument has been at a high level in Finland. In 2010, 70 percent of Finnish hospital districts took part in the national benchmarking. This data can be used in decision making for the purposes of developing the outcomes of students' learning and units' supervisory activities.

One of the advantages of using this instrument is the possibility for reliable international comparisons. In the critical review of the international instrument development articles, the CLES+T scale has reported adequate features of cultural validity (Sims et al. 2010, Johansson et al. 2010, De Witte et al. 2011, Skaalvik et al. 2011, Bos et al. 2012, Tomietto et al. 2012,Watson et al. 2012). This may provide evidence for the decision making process used in the development of the education system within health care and education.

There is clear research evidence that the frequency of meetings between students and their teachers has decreased during last decades (Wills 1997, Barrett 2007). It can be seen an indicator of development process which has occurred in many European countries during the transition from hospital or vocational college based school system to Higher Education Institutions. The Finnish respondents evaluated clearly lower the subscale exploring teacher's share than then quality of learning environment or supervisory relationship. This result supports the findings of few earlier studies (Johansson et al. 2010, Warne et al. 2010). In this process the role of a teacher has changed from a clinical skilled practitioner to a cooperation person between the education institutions and health care service organizations (Warne et al. 2010). This matter can be seen as a reason to retest the content validity of the Teacher subscale items in the future.

The quality of learning environment has been identified as a crucial element when recruiting new staff to health care organizations and this evaluation instrument has proven to be a useful tool to benchmark the attractiveness of organizations among graduating students. However, the major disadvantage in using this single instrument model is that we measure the only the input for clinical supervision. More evidence is needed of the learning outcomes during clinical placements. The cooperation between teachers, students and supervising clinical staff should be more carefully explored.

References

Barrett, D. (2007). The clinical role of nurse lecturers: past, present and future. *Nurse Education Today* 27; 367-374.

Bos, E., Alinaghizadeh-Mollasaraie, F., Saarikoski, M. & Kaila, P. (2012). Development and validation of the instrument Clinical Learning Environment, Supervision and Nurse Teacher (CLES+T) in the context of primary health care in Sweden. *Journal of Clinical Nursing*. (accepted)

Chan, D. (2002). Development of the clinical learning environment inventory: Using the theoretical framework of learning environment studies to assess nursing students' perceptions of the hospital as a learning environment. *Journal of Nursing Education*, 41; 69-75.

De Witte, N., Labeau, S. & De Keyzer, W. (2011). The clinical learning environment and supervision instrument (CLES): validity and reliability of the Dutch version (CLES+NL). I*nternational Journal of Nursing Studies* 48; 568-572.

Dunn, S.V. & Burnett P. (1995). The development of a clinical learning environment scale. *Journal of Advanced Nursing*, 22(6); 1166-1173.

European Ministers of Education. (1999). Bologna Declaration. The European Higher Education Area,
Bologna: The National Unions of Students in Europe. Retrieved 15.11.2011. Available: <http://www.esib.org/BPC/docs/Archives/CoP007bolognadeclaration.pdf>.

Hosoda, Y. (2006). Development and testing of a Clinical Learning Environment Diagnostic Inventory for baccalaureate nursing students. *Journal of Advanced Nursing* 56(5); 480-490.

Johansson, U. B., Kaila, P., Ahlner-Elmqvist, M., Leksell, J., Isoaho, H. & Saarikoski, M. (2010). Psychometric evaluation of the Swedish version of Clinical Learning Environment, Supervision and Nurse Teacher evaluation scale. *Journal of Advanced Nursing* 66; 2085-2093.
Midgley, K. (2006). Pre-registration student nurses perception of the hospital-learning environment during clinical placements. *Nurse Education Today* 26; 338–345.

Saarikoski, M. & Leino-Kilpi, H. (2002). The clinical learning environment and supervision by staff nurses: developing the instrument. *International Journal of Nursing Studies* 39; 259-267

Saarikoski, M., Isoaho, H., Warne, T. & Leino-Kilpi, H. (2008). The Nurse Teacher in clinical practice: developing the new sub-dimension to Clinical Learning Environment and Supervision (CLES) scale. *International Journal of Nursing Studies* 45; 1233-1237.

Saarikoski, M., Warne, T., Kaila, P. & Leino-Kilpi, H. (2009). The role of nurse teacher in clinical practice; an empirical study of Finnish student nurse experiences. *Nurse Education Today* 29; 595-600.

Salminen, L., Stolt, M., Saarikoski, M., Suikkala, A., Vaartio, H. & Leino-Kilpi, H., (2010). Future challenges for nursing education - A European perspective. *Nurse Education Today* 30, 233-238.

Sims, D., Watson, P., Seaton, P., Whittle, R., Jamieson, I., Saarikoski, M. & Mountier, J. (2010). Evaluating the quality of workplace learning for nursing students in community settings. Research report of Ako Aotearoa/ NZ.

Skaalvik, M.W., Normann, H.K. & Henriksen, N. (2011). Clinical learning environment and supervision: experiences of Norwegian nursing students – a questionnaire survey. *Journal of Clinical Nursing* 20, 2294–2304.

Tomietto, M., Palese, A., Saiani, L., Cicolini, G., Watson, P. & Saarikoski, M. (2012). Clinical Learning Environment and Supervision plus Nurse Teacher (CLES+T) scale: testing the psychometric characteristics of the Italian version. (submitted)

Warne, T., Johansson, U-B., Papastavrou, E., Tichelaar, E., Tomietto, M., Van den Bossche, K.,Vizcaya-Moreno, M. F. & Saarikoski, M. (2010). An exploration of the clinical learning experience of nursing students in nine European countries. *Nurse Education Today* 30; 809-815.

Watson, P., Seaton, P., Whittle, R., Sims, D., Jamieson, I., Saarikoski, M. & Mountiesr, J. (2012). Exploratory Factor Analysis of the Clinical Learning Environment, Supervision and Nurse Teacher Scale (CLES+T). *Journal of Nursing Measurement*. (accepted)

Wills, M. E. (1997). Link teacher behaviours: student nurses' perceptions. *Nurse Education Today* 17, 232-246.

Practice Evaluation in Child Welfare: Methodological Considerations

Janissa Miettinen, MSc, Department of Social Sciences, University of Eastern Finland

Abstract

The question of how and under what conditions child welfare services can achieve the desired outcomes is crucial in child welfare research. In Finland, after a guarded beginning, the possibility of evaluating child welfare practices seems promising, yet evaluation research in social work, especially in child welfare, is rather scant. Finnish child welfare researchers face the challenge of building a sufficient knowledge-base to help the children and families in need of child welfare and *open care* services. This paper thus discusses the research methodology in practice evaluation. It analyses the scientific logic of the realist evaluation paradigm, and shows how abduction, a third form of making scientific inferences, could guide a research that utilises the ideas of realist evaluation.

Keywords: research methodology, realist evaluation, abduction, retroduction, child welfare

1. Introduction

1.1 Guarded orientation to evaluation of child welfare: a promising beginning in Finland

In Finland in 2010, over 70, 000 children under 18 years old received community-based child welfare interventions (Lastensuojelu 2010), or open care. [1] In order to be able to help these children, we still need more knowledge of "*how* to help" (my emphasis) as this could inform the child welfare social workers in their practical work in future (Munro, 2004; Raunio, 2009, p. 150). Cheetham, Fuller, McIvor and Petch (1992, see pp. 3−5) presented different reasons, both external and internal, for the evaluation of social work. Such research should support the aims of social work practice and be driven by its primary goal, characterised, for example, by Raunio (2009, pp. 58−62) as intervening so as to generate positive change in a client's life, and by the International Federation of Social Workers (IFSW, 2000) as follows:

> The social work profession promotes social change, problem solving in human relationships and the empowerment and liberation of people to enhance well-being. Utilising theories of human behaviour and social systems, social work intervenes at the points where people interact with their environments. Principles of human rights and social justice are fundamental to social work. … Professional social work is focused on problem solving and change … (IFSW, 2012).

The demand to prove the effectiveness of social work interventions in Finland arose in the 1980s and had grown strong by the end of 1990s, alongside the ideas of New Public Management (e.g. Niiranen et al., 2005). Subsequently, Rajavaara (2007) even launched the

---

1 According to the Child Welfare Act, "the municipal body responsible for social services must provide support in open care … without delay if" either: "the circumstances in which the children are being brought up are endangering or failing to safeguard their health or development; or the children's behaviour is endangering their health or development" (Lastensuojelulaki 417/2007, 34§, p. 15). Thus, *open care* refers to provision of various services to a child and his/her family (see also Taskinen, 2007, pp. 41−44) due to the above defined criteria in order to preserve the family (see also Miettinen 2011); provision of open care is based on the willingness of the family. If the support given through open care is insufficient, and the best interest of a child demands it, s/he will be taken into care (see L 417/2007, 40§, p. 18).

term "outcome-oriented society" (*vaikuttavuusyhteiskunta*), to show the prevalence of the concept of *effectiveness* in public governance and discourse. However, the challenging mission of evaluating practice and supplementing the *knowledge-base* of child welfare social work is still in its initial stages in Finland. Several scholars have noticed this gap and recommended further research, particularly evaluative research on the "core activities" or "practice" of child welfare social work in the country (e.g. Eronen, 2007; Pekkarinen, 2012, pp. 53, 55).

The reason for the scarcity of social work evaluation research in Finland might be that, as Suikkanen (2008, p. 101) indicated concerning the premise of evaluation in social sciences, in social work the demand for evaluation has come from outside the discipline. As several scholars have remarked on, the external demands for effectiveness evaluation diverges from the premise of social work, and this contradiction may have resulted in the avoidance of evaluation (for more details, see e.g. Raunio, 2009, pp. 146−149; Pohjola, 2012, pp. 26−28). Fortunately, e.g. the first compilation of Finnish research on the effectiveness of social work (Pohjola, Kemppainen & Väyrynen, 2012), shows that scholars in the area are developing the discipline's own premises in the evaluation of social work practice.

The present study explores the contents and prerequisites of effectiveness for open care in child welfare (*lastensuojelun avohuolto*), which falls under the domain of *child- and family-specific child welfare* in the Finnish child welfare system.[2] This is an area of much dispute since the definition of effectiveness in social work is continually debated (see e.g. Pohjola et al. 2012). It is in this context, therefore, that this study considers the recognition and modelling of the mechanisms that might promote or inhibit the desired outcomes of open care (see also Miettinen, 2011), assuming that a realist methodology is suitable for the explanation of complex child welfare practice. This methodological paper aims at illustrating possible ways of utilising ideas of realist evaluation in respect of child welfare practice.

In the following section I briefly review Finnish child welfare evaluation research. After this I consider some philosophical aspects of the origins of realist evaluation. Then, I analyse the possibilities for abduction as a basis for a research design that utilises ideas of realist evaluation. Next, I illustrate the use of abduction and retroduction in planning a realist research design and data analysis, and report on the data gathering procedure. Finally, I discuss the possible relevance of this suggested research methodology to child welfare.

1.2 Overview of evaluative studies in Finnish child welfare

Since a comprehensive review of Finnish child welfare research is not possible here (for more information see e.g. Eronen, 2007; Kivipelto, 2010),[3] I will limit myself to a few notable publications. Probably the first work to consider the effects of child welfare for children form social work perspective was Reino Salo's 1956 dissertation. A search made on the Linda data

---

2

Taskinen (2007) divided the Finnish child welfare services into three domains: a) universal, b) preventive and c) child- and family-specific child welfare. The last of these is the highest level of safeguarding, and includes "the investigation of the need for child welfare measures, a client plan and the provision of support in open care", and "… emergency placement of a child and taking a child into care, as well as substitute care and after-care related to these" (see L 417/2007, 3§, p. 1).

3

Kivipelto's unofficial report, a summary of Finnish social services effectiveness research, includes 48 publications, also those e.g. of adults, elderly, substance abuse, rehabilitation and theoretical research (See Haverinen's (2012, pp. 74−75) summary of the report; She also stated, knowledge of the effectiveness of social work in Finland is scant (p. 78).)

base for relevant literature found between 11 and 72 references, depending on the search term.[4] Plenty of the later references are not evaluation studies. These references show that one of the first experimental evaluations in open care in child welfare was done in 1990, by the social bureau of Helsinki (Vaikuttavuuden arviointikokeilu...1991).

Recently, Rousu's (2007) dissertation considered the (organisational) factors associated with the effectiveness of child welfare. She found five groups of critical success factors, three of which are prerequisites of effectiveness in child welfare: 1) the organisation's client orientation; 2) adequate competence among personnel; and 3) processes that aims at empowering the client. A follow-up-study in child welfare open care by Huuskonen and Korpinen (2009), and one by Huuskonen, Korpinen, Pösö, Ritala-Koskinen and Vakkari (2011), both continued the work of Tarja Heino, who had asked in 2007, "Who are the new child welfare [i.e. open care] clients?" The National Institute for Health and Welfare has also published relevant studies lately, e.g. Perälä, Salonen, Halme and Nykänen's (2011) study.

Several Finnish scholars have regarded realist evaluation as perhaps the most promising methodology for evaluating social work effectiveness (e.g. Julkunen, Lindqvist & Kainulainen, 2004; cf. see Suikkanen, 2008). Mäntysaari (2005, p. 88) has even proposed "realism as foundation for social work knowledge". Probably the first Finnish social work scholar to utilise the *realist science philosophy* was Mikko Mäntysaari in the research process started in 1983 and which led to his dissertation in 1991 (Mäntysaari, 2006, p. 88). *Realist evaluation* was, supposedly, used for the first time in Finnish social work research in 1998, in the evaluation of the Monet-project (Rostila, 2001, cited in Kazi, Blom, Morén, Perdal & Rostila, 2002). Subsequently, the realist evaluation of Nuotta-project focused on 17-24 year olds with low motivation in education and work (Karjalainen & Blomgren, 2004), and Alpo Heikkinen's (2007) realist ethnography on a child welfare boy group. A search made on the Linda data base[5] found 12 Finnish studies on realist evaluation from 2000-2010.

2. Methodological considerations in realist evaluation

2.1 The background of realist evaluation

Realism includes many orientations but, as Mäntysaari (2005) indicates, *realist evaluation* developed on the basis of critical realist science philosophy, which has its origins in Roy Bhaskar's (1975, 1979) work on transcendental realism, later renamed *critical realism* by its followers (Mäntysaari, 2005, pp. 89−90). The central assumptions in critical realist science philosophy concern the essence of reality (ontology) and knowledge (epistemology). Bhaskar (1975) noted that both reality and knowledge are stratified: reality is seen to be "an open system", referring to the fact there are no "constant conjunctions of events" (p. 13). Furthermore, reality is also seen as independent of the researcher's consciousness[6] and is

4

A search on the Linda data base, which contains the publications of Finnish universities, found 11 references, 1 duplicate, for the (Finnish) search words *(all*) *vaikuttavuus* (effectiveness) and *lastensuojelu* (child welfare) and 72 references with a wider search (with the word for effectiveness shortened, as *vaikut*) (12.4.2012).

5

Linda data base found a total of 16 references, including four duplicates, from 2000-2010 with the (Finnish) search words (*all*) *realistinen* (realist) and *arviointi* (evaluation), on 15.4.2011.

6

Within the science philosophy of realism there is a dispute as to whether reality is totally independent of our consciousness (e.g. Mäntysaari 2005, p. 91). Niiniluoto (1999, pp. 26−27) differentiated between

stratified into three distinct levels: real, actual, and empirical; generative mechanisms are situated in the domain of the real (e.g. Bhaskar, 1975, table 0.1, pp.12−14, 25). The research perspective of *realistic evaluation* was introduced by sociologists Ray Pawson and Nick Tilley in 1997, as they combined the ideas of a "*realist* tradition in the philosophy of science" which "avoids the … poles of positivism and relativism" (p. 55) for the development of evaluation. They adapted the definition of *generative causality* (coined Harré in 1972, cited in Pawson & Tilley, 1997, p. 32), and proposed that evaluation should explain *why* and in *what conditions* a program is successful. This means recognising the *generative mechanisms*, which can trigger change, and the *conditions*, which activate these mechanisms.

The potential of *realist* [coined by Kazi, 2003b] *evaluation* is usually defined in Scriven's (1999) concepts as *from-black-box-towards-the-white* reasoning, (e.g. Kazi et al., 2002; Blom & Morén, 2010), which refers to the degree of explanation the research is able to offer. The "box" represents the inner workings of the program as a whole, which is very difficult to reveal entirely (Scriven, 1999, p. 523). There are different versions of realist evaluation. Kazi (2003a, 2003b), in particular, has continued to develop the paradigm as a practical research methodology for realist evaluation in the form of single-case-design-evaluation, qualitative, and quantitative methodologies. Blom and Morén (2010) have also developed Pawson and Tilley's (1997) ideas (on context-mechanism-outcome, or COM pattern configurations) further as CAIMeR-theory (context-actor-intervention-mechanism-result), which they describe as a model of "how social work in general works" (p. 117). The CAIMeR theory emphasises the workings of generative mechanisms that can induce human change. Houston's (2010) version of critical realist evaluation in the context of action research includes also temporal dimension [context+time+mechanism+human agency = outcomes] (p. 76) for the evaluation of whether change has occurred in outcome measures over time. In 2012, Oliver introduced the perspective of *critical realist grounded theory* for social work research, which offered quite a different view from its predecessors. By combining the ideas of grounded theory with critical realism, the perspective addresses (real) events, "approach(ing) data with the preconceived analytical concepts of emergence and generative mechanisms", yet "grounding findings ... in participants' experiences" (see Oliver, 2012, pp. 378, 384).

2.2 Deduction as the premise of realistic evaluation—is there another possibility?

Discussion of the scientific logic of realist evaluation paradigm might start with Stame's (2004) inclusion of Pawson and Tilley's (1997) realistic evaluation as one form of "theory-oriented approach" – along with Chen and Rossi's (1989) theory-driven evaluation and Weiss's (e.g. 1997) theory-based evaluation. Stame claims that although the three perspectives establish three different ways of addressing the above-mentioned problem of *opening-the-black-box*, their common feature is that they emphasise theory rather than disputing method (Stame, 2004, pp. 60−61). The creators of realistic evaluation, Pawson and Tilley, stated it as follows:

> When it comes to research methods and strategies, our realist injunctions do not constitute a plea for complete and permanent revolution. Thus, in broad terms,

---

ontological independence and causal independence, such that the human mind can be (at last partially) ontologically independent of the reality that is outside human mind, *and* in causal relationship with it. Ontological realists tend to accept at the very least the assumption that "at least part of reality is ontologically independent of human minds", while the definition of truth divides opinions. The major distinction here is between definitions of *truth as correspondence* and *surrogates of truth*. The latter is defended by *internal realists*, as Mäntysaari (2005, p. 89), and the former by, for example, *critical realists* (see Niiniluoto, 1999, pp. 10−11; he advocates a version of scientific realism, while accepting a "mixed position between metaphysical and internal realism" (p. 226)).

> research designs for realist evaluation studies actually follow exactly the same basic logic of inquiry as that underpinning any other area of social science. … in all cases the 'wheel of science' is followed (Wallace, 1971[Reference original]) (Pawson & Tilley, 1997, p. 84).

On the next page, Pawson and Tilley (1997) introduce the *realist evaluation cycle* (see p. 85). This cycle seems to have a deductive-oriented logic for scientific inferences (see also Fig. 9.1), which suggests preferring the procedure of deriving testable hypothesis, "what might work for whom in what circumstances", testing the hypothesis with suitable methods, and, finally, confirming or specifying the candidate theories.[7] However, the emphasis could be on testing or developing the theory; if there is not a theory of how the program works in existence, the realist theory to be tested can be formulated by interviewing or observing the practitioners of the program (ibid. p. 107; see Kazi, 2003b; Houston, 2010), and further, supplemented by external theories and previous research (Pawson & Tilley, 1997, pp. 138, 147). The rules of realistic evaluation however, are quite abstract (Kazi 2003b), so I will discuss of abduction, the third logic of making scientific inferences, as a methodological tool.

3. Introducing the research design

3.1 Abduction and retroduction as methodological tools for building hypotheses

Reasoning in realist evaluation is essentially *retroductive* (e.g. Kazi 2003b). Charles Peirce is known as the developer of *abductive reasoning*, but he used also the concept of retroduction (McKaughan, 2008), and scholars have different opinions of what the pragmatist had in mind when he talked about abduction as a distinct form of scientific inference, apart, that is, from deduction and induction. McKaughan (2008, pp. 452−453), for example, differentiated three ways of interpreting Peirce's ideas of abduction. First, there is abduction as a path to scientific discovery; second is "reasoning toward the justification of the truth of theories", or "inference to the best explanation"; and finally, in the interpretation he found most correct, there is abduction as a way to determine the "*pursuitworthiness* of theories" before they are tested.

Chiasson (2001) defined retroduction and abduction as separated concepts: retroduction as a method for "engendering theories … by the interplay of abduction, deduction, and induction", and abduction as a distinct form of scientific inference for formulating hypotheses; abduction is necessary stage of retroduction. Kazi (2003a) defined retroduction as the "process that enables the realist inquirer to investigate the causal mechanisms and other conditions under which certain outcomes will or will not be realized" (p. 805). Since Houston (2010, p. 83) also illustrated retroduction as a five-step procedure, this supports Chiasson's (2001) conclusion of abduction as a narrower concept and a part of retroduction.

Table 1 illustrates with Peirce's syllogisms−in which I focus on−the model of scientific reasoning in deduction, induction, and abduction. As a method of scientific inference, abduction operates from *rule* and *result* to a "reasonable assumption", i.e., *case* (see Table 1),[8] and cannot be reduced to either deduction or induction (Bertillsson & Christianssen, 2001, p. 469−470). Abduction means *seeking* plausible explanations, or hypotheses, regarding a given

---

7 Realistic *theory* means propositions of how a program generates positive change in the pre-existing conditions.

8 Table 1 is adapted from Bertillsson & Christianssen (2001, p. 469) and interpreted from top to bottom, such that the two first clauses are premises and the bottom is the conclusion.

situation. Such hypotheses need to be likely in a pragmatic sense, and might have other impact in the future (ibid. 2001, p. 469−470), such as in child welfare practice. *Pursuitworthy* hypotheses also need to be potentially testable (McKaughan, 2008).

By analysing the role of theory in publications of some developers of realist evaluation, I was able to discern that the three proposed perspectives, deductive inductive and abductive orientation, already existed—implicitly, to some extent—in the previous research in realist evaluation. Accordingly, I have placed some developers of realist evaluation in line with this classification of the logic of scientific reasoning (Table 1). The classification of different authors into the three orientations might be somewhat simplifying, of course, but is offered in an attempt to show different possible emphases of research that utilise ideas of realist evaluation, especially the conception of generative causality, to explain how social work, or child welfare practice, operates to generate the desired change.

Table 1. Models of scientific inference in realist evaluation

Deduction (Pawson & Tilley, 1997)		**Induction** (Oliver, 2012)		**Abduction** (Kazi, 2003; Heikkinen, 2007; Houston, 2010)	
Rule	All beans in this bag are white.	Case	These beans are from this bag.	Rule	All beans in this bag are white.
Case	These beans are from this bag.	Result	These beans are white.	Result	These beans are white.
Result	These beans are white.	Rule	All beans in this bag are white.	Case	These beans are from this bag.

An abductive orientation to realist evaluation emphasises a procedure for *theory formulation* by abduction and retroduction. A deductive orientation, alternatively, offers the "realist evaluation cycle" procedure, which emphasises a more theory-driven orientation to *theory refinement*; while in an inductive orientation, hypotheses are developed mainly *without explicitly including existing theories* in the analysis. I place Pawson and Tilley (1997) as deductive orientation because of their emphasis on theory-driven analysis (see e.g. pp. 155, 161), and Oliver (2011) as an example of inductive orientation. However, Oliver's idea is to combine the idea of critical realism to grounded theory, not utilise realist evaluation, and her perspective could also be considered approaching the abduction-oriented one. Nevertheless, her orientation could represent a somewhat opposite position as related to theory in explanative research to that of Pawson and Tilley. Kazi (2003b), Heikkinen (2007) and Houston (2010) are seen as somewhat abduction-oriented scholars, yet Houston seems to emphasise theory-refinement though cyclic research procedure. In the following, I will show how abduction and retroduction can operate as a scientific method in realist research design.

3.2 Abduction in realist research design in practice

Grönfors (2011, pp. 17−20) underscores how abductive inference means that some "guiding principle leads the research and helps the researcher to focus on essential issues during data collection and analysis" − yet also surprising findings might result in new theoretical ideas (Grönfors 2011). Thus, a (realist) theory could be formulated by deriving inferences from the data and comparing this information to relevant theories or other relevant information. In this study abduction and retroduction are regarded as a scientific method of making inferences, creating testable theoretical hypotheses, and testing at least parts of the

hypotheses empirically. This is done by a two-phased research design.[9] The first phase aims at creating theoretical hypotheses with focus group interviews and abductive reasoning, and the second phase at testing at least parts of these hypotheses empirically (by retrospective longitudinal statistical data gathered from child welfare case files).

In the first phase, the prerequisites of effectiveness of open care in child welfare were, as much as possible, studied by searching for an explanation for the retroductive question of "why the outcomes developed as they did" (Kazi 2003a, p. 805). The explanation was sought by "translating" the idea of retroduction and the logic of generative causality through the analytical work of deconstructing the components of explanative model of realist evaluation within semi-structured focus group interview questions. In focus groups, the participants could convey their practical expertise to researcher, and the interview themes comprised, for example, the following issues: documentation, targets of open care, means to reach the targets (e.g. services), the situation of client families and the conditions which could promote or inhibit the desired outcomes in certain situations.

In data analysis, abduction and retroduction are synthesising 1) elements of expert knowledge (the data), 2) theories (e.g. Bronfenbrenner's [2005] bio-ecological model of human development), and 3) previous research findings (e.g. the risk, resilience and protective factors perspective (Pecora, 2006)) along the logic of generative causality into theoretical hypotheses. If we accept McKaughan's (2008) interpretation of Peirce's work (as described above), then the process of abduction in the first phase of research design will produce *pursuitworthy* hypotheses, which are at least partly, and/or in simplified form, testable with case files. As theory testing is preferable after the hypotheses are formulated (ibid, 2008, pp. 454−455), this is potentially possible in a two-phased research design.

Miles and Huberman (1994, p. 4) argued as transcendental realists that qualitative analysis can reveal causal mechanisms that are not just hypotheses by utilising a local, contextual and temporal analysis of which occurrence preceded another. In particular, the retrospective method, defined as retroduction here, may help to identify causal mechanisms (ibid. pp. 146−147). Kazi (2003b, pp. 30−31) suggested that "systematically tracking" changes in the outcomes, in the contents of the intervention, in client's life contexts, and, finally, in the mechanisms, all enables the formulation of explanations of why a program worked or did not. Applying this method, the components of explanation are first gathered into a matrix. Their operations and connections are then analysed in a given context(s) in relation to the theoretical framework (see Miles & Huberman, 1994) that can act as a "rule"; this process is defined as abduction. In hypotheses formulation, abduction then becomes a part of retroductive inference process (e.g. Chiasson, 2001). I conclude that if a causal mechanism exists in reality as it is identified in this qualitative analysis, then a causal mechanism is found. Causal mechanisms are likely partially observable (see Blom & Morén, 2010). However, some parts of the hypotheses might not be observable or testable, because they are situated in the domain of the real (Bhaskar, 1975, p. 13), or the information is not reached by interviews or documented in case files and the research design is mostly retrospective. Then, causal mechanisms might be captured by theorising (Blom & Morén, 2010). If several sources indicate the same inference of some causal mechanism, and/or if another source could plausibly explain a situation such that a mechanism becomes plausible, then the existence of the identified causal mechanism is *pursuitworthy*.

---

9 E.g. Pawson & Tilley, (1997) and Houston (2010) have also suggested a design with several linking studies. Currently (May 31, 2012), I have conducted the focus group interviews, and analysis is at a beginning stage.

3.3 Data gathering

Morgan (1988) indicates that focus groups can be an instrument in generating hypotheses that will not be derived solely from existing theories or research. This is suitable for abductive reasoning. For this research, semi-structured focus group interviews were conducted from December 7, 2011 to March 28, 2012, during working time. Two groups of child welfare social workers and one group of child welfare family workers were interviewed twice. The participants received the themes beforehand. At the beginning of the second interview, a summary of the first interview was given, on which respondents could comment. A group of child welfare managers of the municipality was interviewed once.

Based on these interviews, I will formulate a data-gathering instrument. Then, I apply for permission for the theory-testing phase with child welfare case-file data. A retrospective case-control study might reveal the factors related to the defined outcome criteria. In addition to the planned statistical path analysis, a deepening qualitative analysis from case files could reveal causal mechanisms and either confirm or disconfirm the prior qualitative data analysis.

4. Discussion

As a design with several linking studies and the logic of abduction and retroduction already exists in realist evaluation, the design proposed here is not particularly new. However, this paper has introduced a slightly new emphasis to a study that utilises ideas of realist evaluation by suggesting abductive reasoning as one premise of realist evaluation design. It has also tried to illustrate possible emphasises within realist evaluation. The abductive method in practice evaluation offers a tool to theory-building, considering also the practice expertise of how the desired outcomes might be obtained, and yet taking advantage of relevant theories and earlier research findings. So, the research methodology presented in this paper would ideally lead to a piece of practical theory of "*how* to help" in child welfare open care. To achieve practical implications the findings should be disseminated to practitioners and policy actors, so that child welfare practice can further develop via theory building.

Acknowledgements

I thank the anonymous referees for their constructive comments, and the National Post-Graduate School for Social Work and Social Services, and research project *Needs, processes and outcomes in child protection*, funded by the Academy of Finland, for enabling this study. I also want to thank the proofreaders of Scribendi Inc for their assistance.

References

Bhaskar, R. (1975). *A realist theory of science.* Leeds: Leeds Books.

Bhaskar, R. (1979). *The possibility of naturalism. A philosophical critique of the contemporary human sciences.* Third Edition. London and New York: Routledge.

Bertillsson, M. & Christianssen, P. V. (2001). Jälkisanat [Afterwords]. (Trans. from Swedish to Finnish: Markus Lång). In C.S.Peirce. *Johdatus tieteen logiikkaan ja muita kirjoituksia.* [Introduction to the logic of science and other thoughts]. (pp. 443-473).

Blom, B. & Morén, S. (2010). Explaining social work practice. The CAIMeR theory. *Journal of social work 2010 10*(1), 98-119. doi: 10.1177/1468017309350661

Bronfenbrenner, U. (Ed.). (2005). *Making human beings human. Bioecological perspectives on human development*. Thousand Oaks, CA: Sage.

Cheetham, J., Fuller, R., McIvor, G. & Petch, A. (1992). *Evaluating social work effectiveness.*

Buckingham, UK: Open University Press.
Chen, H. & Rossi, P. (1989). Issues in the theory-driven perspective. *Evaluation and Program Planning 12*(4), 299-306. doi: 10.1016/0149-7189(89)90046-3
Chiasson, P. (2001). Abduction as an aspect of retroduction. *Digital encyclopedia of Charles S. Peirce.* Retrieved March 5, 2012, from http://www.digitalpeirce.fee.unicamp.br/p-abachi.htm.
Eronen, T. (2007). *Katsaus 2000-luvulla julkaistuun suomalaiseen lastensuojelututkimukseen* [A review of the Finnish research on child protection], Sosiaalialan kehittämishanke, Lastensuojelun kehittämisohjelma, Retrieved March 5, 2012, from www.sosiaaliportti.fi/fi-FI/lastensuojelunkasikirja/tuke/lastensuojeluntutkimuskatsaus/.
Grönfors, M. (2011). *Laadullisen tutkimuksen kenttätyömenetelmät* [Fieldwork methods of qualitative research]. H. Vilkka (Ed.). Hämeenlinna: SoFia-Sosiologi-Filosofiapu Vilkka. Retrieved March 5, 2012, from http://vilkka.fi/books/Laadullisen_tutkimuksen.pdf (Original Work Published 1982).
Haverinen, R. (2012). Vaikuttavuus ja näyttö tavoitteena sekä sosiaalityön asiakastyön tutkimuksen kohteena [Effectiveness and evidence as target and as the object of social works client work research]. In A. Pohjola, T. Kemppainen, & S. Väyrynen (Eds.), *Sosiaalityön vaikuttavuus.* [Social work's effectiveness]. (pp. 65-85). Rovaniemi: Lapland University Press.
Heikkinen, A. (2007). Olenko mä sitä riskiryhmää? Murrosikäiset pojat kouluvaikeuksien metsäpolulla [Do I belong to the risk-group? Adolescent boys on the forest-path of school-difficulties]. In A. Heikkinen, P. Levamo, M. Parviainen & A. Savolainen (Eds.), *Näe minut Kuule minua. Kokemuksia ryhmistä.* Julkaisusarja 11. (pp. 15-68). Helsinki: Socca.
Heino, T. (2007). *Keitä ovat uudet lastensuojelun asiakkaat? Tutkimus lapsista ja perheistä tilastolukujen takana* [Who are the new clients of child welfare? A study of the children and families behind the statistics]. (Työpapereita, 30/2007). Helsinki: Stakes.
Houston, S. (2010). Prising open the black box. Critical realism, Action research and Social work. *Qualitative Social Work 9*(1), 73-91.doi: 10.1177/1473325009355622
Huuskonen, S. & Korpinen, J. (2009). *Runsas vuosi lastensuojelun avohuollon asiakkuuden alkamisesta: mitä lapsille kuuluu nyt? Lastensuojelun tieto -hankkeen loppuraportti* [Over one year since being registered in open care in child welfare: How are the children now? The final report on child welfare knowledge -project]. Tampere: Picassos Oy.
Huuskonen, S., Korpinen, J., Pösö, T., Ritala-Koskinen, A. & Vakkari, P. (2010). Kolme polkua lastensuojelun avohuollon organisatorisessa muistissa [Three paths in the organisational memory of child welfare open care]. *Yhteiskuntapolitiikka 75*(6), 650-658.
International Federation of Social Workers (ISFW). (2012). Definition of social work. (Allowed in July 2000 by International Federation of Social Workers General Meeting in Montréal, Canada). Retrieved March 5, 2012, from http://www.ifsw.org/f38000138.html.
Julkunen, I., Lindqvist, T. & Kainulainen, S. (Eds.). (2005). Realistisen arvioinnin ensimmäiset askeleet [First steps of realist evaluation]. (FinSoc työpapereita 3). Helsinki: Stakes.
Karjalainen, P. & Blomgren, S. (2004). *Oikorata vai mutkatie? Sosiaalista kuntoutusta ja työelämäpolkuja nuorille. Nuotta-projektin arvioinnin loppuraportti* [Short-cut or curve-road? Social rehabilitation and work-life paths to youngsters. Final report on the Nuotta-project's evaluation]. (FinSoc arviointiraportteja 2). Helsinki: Stakes.
Kazi, M. (2003a). Realist evaluation for practice. *Br J Soc Work, 33*(6), 803-818.
Kazi, M. (2003b). *Realist evaluation in practice. Health and social work.* London: Sage. Retrieved March 5, 2012, from University of Eastern Finland's Ebrary database.
Kazi, M., Blom, B., Morén, S., Perdal, A-L. & Rostila, I. (2002). Realist evaluation for

practice in Sweden, Finland and Britain. *Journal of Social Work Research and Evaluation, 3*(2), 171-186.
Kivipelto, M. (2010). Yhteenveto suomalaisista sosiaalipalvelujen vaikuttavuustutkimuksista, muistio sosiaali- ja terveysministerille [Summary of Finnish child welfare research]. Terveyden ja hyvinvoinnin laitos. Helsinki: Finsoc.
Lastensuojelu 2010. [Child Welfare 2010]. (2011). Statistical Report 29/2011. Helsinki: Stakes & Suomen virallinen tilasto. Retrieved May 29, 2012, from http://www.stakes.fi/tilastot/tilastotiedotteet/2011/Tr29_11.pdf.
Lastensuojelulaki (2007). [Child Welfare Act], Finland, 417/2007. Unofficial translation, Oikeusministeriö & Edita Publishing Oy. Retrieved May 29, 2012, from http://www.finlex.fi/fi/laki/kaannokset/2007/en20070417.pdf.
McKaughan, D. J. (2008). From ugly duckling to swan: C.S. Peirce, abduction, and the pursuit of scientific theories. *Transactions of the Charles S. Peirce society 44*(3), 446-468.
Miettinen, J. (2011). The prerequisites of success in child welfare open care. *Social Work and Society 9*(1). *Special issue "Practice Research"*, 142-145. urn:nbn:de:0009-11-29861.
Miles, M. & Huberman, A. (1994). *Qualitative data analysis: An Expanded Sourcebook.* Thousand Oaks, CA: Sage.
Morgan, D. L. (1988). *Focus groups as qualitative research.* Qualitative research methods series, Volume 16. Newbury Park: Sage.
Munro, E. (2004). The impact of audit on social work practice. *Br J Soc Work 34*(8), 1075-1095. doi: 10.1093/bjswtbch130
Mäntysaari, M. (2006). Realism as a foundation for social work knowledge. *Qualitative Social Work 4*(1), 87-98. doi: 10.1177/1473325005050202
Niiniluoto, I. (1999). *Critical scientific realism.* Clarendon Library of Logic and Philosophy. Oxford : Oxford University Press.
Niiranen, V., Stenvall, J., Lumijärvi, I., Meklin, P. & Varila, J. (2005). *Miten arvioida kuntapalvelujen tuloksellisuutta? Kartuke-tutkimuksen lähtökohdat, metodologiset sitoumukset ja tavoitteet* [How to evaluate the profitability of communal services? The premise, methodological engagements and objectives of Kartuke-research]. In V. Niiranen, V., Stenvall, J. & Lumijärvi, I. (Eds.), *Kuntapalvelujen tuloksellisuuden arviointi. Tasapainotettu mittaristo Kunnallisissa organisaatioissa.* (pp. 11-47). Jyväskylä: Ps-kustannus.
Oliver, C. (2012). Critical realist grounded theory: a new approach for social work research. *Br J Soc Work, 42*(2) 371-387. doi: 10.1093/bjsw/bcr064
Pawson, R. & Tilley, N. (1997). *Realistic evaluation.* London: Sage.
Pecora, P. (2006). Child welfare policies and programs. In J. Jenson & M. Fraser (Eds.), *Social policy for children & families. A risk and resilience perspective,* 19-66. Newbury Park, CA: Sage.
Pekkarinen, E. (2011). *Lastensuojelun tieto ja tutkimus – asiantuntijoiden näkökulma* [Knowledge and research in child welfare – the perspective of specialists]. Nuorisotutkimusverkosto/Nuorisotutkimusseura, Verkkojulkaisuja 51, 2011. Retrieved March 2, 2012 from http://www.nuorisotutkimusseura.fi/julkaisuja/lastensuojeluntieto.pdf.
Perälä, M-L., Salonen, A., Halme, N. & Nykänen, S. (2011). Miten lasten ja perheiden palvelut vastaavat tarpeita? Vanhempien näkökulma [How do services for families and children meet the needs? Parents' views]. Report 36/2001, Helsinki: THL.
Pohjola, A. (2012). Tutkimukseen perustuva vaikuttavuus [Research-based effectiveness]. In A. Pohjola, T. Kemppainen & S. Väyrynen (Eds.), *Sosiaalityön vaikuttavuus.* [Social work's effectiveness]. (pp. 19-42). Rovaniemi: Lapland University Press.

Pohjola, A., Kemppainen, T. & Väyrynen, S. (Eds.). (2012). *Sosiaalityön vaikuttavuus* [Social work's effectiveness] Rovaniemi: Lapland University Press.

Rajavaara, M. (2007). *Vaikuttavuusyhteiskunta: sosiaalisten olojen arvostelusta vaikutusten todentamiseen* [The outcome-oriented society: Styles of reasoning about the outcomes in social policy]. (Sosiaali- ja terveysturvan tutkimuksia 84). Helsinki: Kela.

Raunio, K. (2009). *Olennainen sosiaalityössä* [The essential in social work]. Helsinki: Gaudeamus.

Rousu, S. (2007). *Tuloksellisuuden arviointi organisaatiossa. Näkymätön tuloksellisuus näkyväksi* [Assessment of child welfare effectiveness in organizations. Making the invisible visible]. Acta Publications No 197. Helsinki: The Association of Finnish Local and Regional Authorities.

Salo, R. (1956). *Kunnallinen lastensuojelutyö sosiaalisen sopeutumisen kasvattajana: tutkimus Vaasan kaupungissa vuosina 1924-1952 suoritetun lakisääteisen lastensuojelutyön tuloksista English Summary. (Municipal child welfare work as promoter of social adjustment).* Vaasa: Vaasan Kirjapaino.

Scriven, M. (1999). The fine line between evaluation and explanation. *Research on Social Work Practice 9*(4), 521-524. doi: 10.1177/104973159900900407

Suikkanen, A. (2008). Arvioinnin avaimia kuntoutuksen lukkoihin [Keys of evaluation to the locks of rehabilitation]. In J. Mäkitalo, J. Turunen & I. Vilkkumaa (Eds.), *Vaikuttavuus muutoksessa.* [Effectiveness in change]. (pp. 99-111). Oulu: Verve.

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

Taskinen, S. (2007). *Lastensuojelulaki (417/2007). Soveltamisopas.* [Child welfare Act (417/2007). Application guidebook]. Oppaita 65. Helsinki: Stakes.

Vaikuttavuuden arviointikokeilu lastensuojelun avohuollossa vuonna 1990. (1991). [Experiment of effectiveness evaluation in open care in child welfare in 1990]. Helsingin sosiaaliviraston julkaisusarja A 2/1991.

Wallace, W. (1971). *The logic of Science in Sociology.* New York, NY: Aldine.

Weiss, C. (1997). Theory-based evaluation: Past, present and future. In D. J. Rog (ed.) Progress and Future Directions in Evaluation, 76. San Fransisco, CA: Jossey-Bass. Retrieved March 11, 2012, from http://onlinelibrary.wiley.com/doi/10.1002/ev.1086/pdf.

Evaluating welfare services amidst an ongoing reform

How to evaluate emergent changes and invisible effects?

Niiranen, Vuokko[1]*; Department of Health and Social Management, University of Eastern Finland*

Puustinen, Alisa[2]*; Department of Health and Social Management University of Eastern Finland*

Abstract

In our evaluation practice we face the problem of grasping a phenomenon that seems to be so emergent and fluctuating that it is difficult to keep track of all the changes taking place. This paper focuses on the theoretical and practical implications of complex interventions on the practice of realistic evaluation. The challenge of evaluating welfare service reforms is to identify the overlapping, paradoxical processes and interactions of several organizations and actors. We have turned our attention to complexity sciences to find new tools to describe the emergent phenomena. When we adopt a complexity theoretical evaluation frame it has implications both for the practice of evaluation and for the use of evaluation results. We claim that in order to succeed evaluators must address the issues of how traditional program evaluation methods fit the context of complex reforms and what kind of knowledge or tools are needed in these situations.

Keywords: evaluation, theory, welfare services, complexity, context

Introduction

The public sector as a whole and especially local governance and welfare services are facing great challenges all over Europe. In the Nordic Countries the welfare society model is also under great pressure. In Finnish local government there has been an extended period of gradual reforms, which has now lasted for almost twenty years. In recent years, the tempo of these reforms has accelerated, and the cycle of changes is probably shorter than ever before. This mainly rests upon demographic changes, particularly the aging population, and the recurring global and national economic downturn. The principles of the welfare society model bring further challenges to the implementation of local governance and service structure reforms in Finland. The model is based on a long and strong tradition of locally delivered and governed welfare services, and this has secured easily accessible and equal services for all citizens nationwide. In 2007, in order to anticipate the challenges, the Finnish government launched a broad reform to restructure the municipal and service structure (Act on Restructuring Local Government and Services, 169/ 2007). Public social and health care services amidst of the reform have been the focus of our longitudinal evaluation research program (ARTTU, 2008-2012[i]). We have been observing the services for children and

1 Vuokko Niiranen, P.O.Box 1627, 70211 Kuopio, e-mail: vuokko.niiranen@uef.fi

2 Alisa Puustinen, P.O.Box 1627, 70211 Kuopio, e-mail: alisa.puustinen@uef.fi

families, services for the elderly and primary health care, specifically the operation of local health centers. Our research has focused on three structural levels: strategic, organizational, and operational. Examples in this article are drawn from the evaluation of changes that have, or have not, taken place in the services for the elderly.

Changes that can be seen in the evaluation indicators of our research agenda are anything but clearly the cause of any one reform. If, for example, the number of elderly people in different kinds of institutional care seems to be rising in one particular municipality, the only reason is hardly the fact that the municipality has gone through a consolidation with another municipality or that there has been a change in the legislation. Causal relationships between reform and observed evaluation outcomes are complex and by no means clear (Raisio 2010). Realistically oriented evaluation already takes this into account by introducing the concept of generative causality, which recognizes the meaning of context and the variety of confusing variables and mechanisms that lead to different outcomes (Pawson 1997). Although realistic evaluation arrangements give us tools to identify the several interacting mechanisms present in one intervention, in our case a reform, the need to represent the complex reality remains somewhat unaddressed. Due to this, we are turning our attention to complexity sciences to find new ways to comprehend the phenomena we are facing in our evaluation practice. The aim is to broaden the evaluation horizon from a basic linear understanding of systems via an open systems model (Holland & Miller 1991) to even more evolving understanding of complex co-evolving systems (Mitleton-Kelly 2006, 225).

In this article, we will explore the possibilities of both realistic evaluation and the promise of complexity sciences in evaluation. First, we will describe our theoretical understanding of complexity related to the evaluation of complex interventions and combine this with the traditional realistic evaluation practices by pointing out some features of our own evaluation practice. Second, we continue by bringing together some guidelines identified in recent research on evaluating complex interventions and reflect these against our experiences of evaluating the municipal and service structure reform in Finland. Finally, in our discussion, we present a revised program theory model of our research subject and, given the restrictions of evaluation research and the complexity of any broad intervention or reform, we ask to what extent is it possible for the reform to work as anticipated. Nevertheless, there still remains the question of how to evaluate emergent changes that cannot, by definition, be predicted?

The meaning of complexity in realistically oriented evaluation

We have been observing the development of services for elderly people by analyzing the changes evident in several statistical evaluation indicators, such as the number of patients in long-term institutional care and the number of people cared for at home. Added to this, we follow up the restructuring of actual service delivery structures and organizational structures in our case municipalities and their strategic decision-making processes and governance strategies. Other modules in the larger ARTTU evaluation research program provide us with information on residential service satisfaction, personnel administration, democracy and policymaking, and on economic development.

When evaluating the changes observed in the local services, we make a clear distinction between *changes and reforms*. Changes may take place without target-oriented efforts to reform the course of actions or operational environment. Changes are positive, negative, or neutral, and they can derive from independent variables, such as demographic changes, political climate or economic fluctuations, as described in the introduction. Reforms, on the other hand, are a good launch pad for changes, and they always require intentional, often broad strategic interventions to shift in the desired direction. (Niiranen 2006.) In the case of

services for the elderly, intentional reforms are, for example, the aforementioned re-structuring of municipal and service structures and National Development Program for Social Welfare and Health Care (Kaste). At the same, there are plans to revise the legislation governing services for the elderly and the national framework for high-quality services for older people. National reforms usually run parallel to each other, but they may also counteract. One fact that we cannot escape is that the population is aging, and no legislation can change that. The service users are also in different positions due to their socio-economic and health backgrounds. It is evident that the mechanisms behind the observed outcomes of any intervention are very complex. They form emergent patterns, self-reinforcing cycles, and constantly co-evolve in many directions.

In this study, the basic orientation of realistic evaluation has been to track the changes and reforms, such as those described above: demographic changes, economic fluctuations, national governmental steering, local restructuring of service delivery and structures, legislative demands and indicators of service use, to name a few. From this diversified information we might begin to map context-mechanism-outcome configurations [CMO](Pawson & Tilley 1997; Kazi 2003). But, as described in the previous chapter, the mechanisms behind any outcome are so complex that basic CMO configurations only seem to scratch their surface.

We define complexity according to Simon Herbert (1962, 468) by referring to complex systems being 'made up of a large number of parts that interact in a nonsimple way'. Those parts are interwoven and hard to separate (Gershenson 2011). Complexity therefore consists of inter-relationships, inter-action and inter-connectivity of different elements inside the system, *as well as* in the links of any system to other systems and its environment (Mitleton-Kelly 2004, 292-293). From the perspective of an evaluator using complexity theoretical position, this means that changes, or reforms, in any one element of the board intervention affects all other elements and more than that, they co-evolve, not just react (Mitleton-Kelly 2006, 225).

In their evaluation of Health Action Zones in England, Marian Barnes et al. (2003) identified several dimensions of complexity. First are the *levels* or structural complexity of the intervention. In our case of evaluating services for elderly people, there are many vertical and horizontal levels present: broad demographic and economic changes, national government with the reform, sector ministries with their informational and legislative governance, local political decision-making, local service delivery structures and organizations, and finally the individual residents of any municipality. *Time* is important due to the prolonged period of gradual reforms public services have faced in Finland. Any outcomes in the indicators of service use also take a long time span to prove. There are multiple *players* on the field, on all the levels described earlier; not only public players, but also private and non-profit sectors play a crucial role in the delivery of welfare services. Several *strategies and models* interact at the same time, and this derives from the multiplicity of levels and players. *Rules and conditions* for service delivery are under continuous revision, both the legislation and informal rules of governing change. Finally, the *problem content* is complex. Social and policy problems that demand intervention from a range of agencies and bodies to address are often called 'wicked problems' (Barnes et al. 2003; Raisio 2010).

The services for the elderly consist of a mixture of social, economic, and health-related issues. To address these complex problems, complex interventions are put in place. Complex problems encompass a range of sub-problems, which may be simple, complicated, or complex. (Glouberman & Zimmerman 2002; Ling 2012.) From an evaluator's point of view, one interesting question is to define the nature of the intervention. Then we set out to observe if the tools used to put the intervention into operation match the nature of the intervention, e.g. are complex interventions put in place with tools that would be more appropriate for a

simple intervention. Or, what is even more important, is our own evaluation agenda designed to address a simple intervention, but in reality we are dealing with a complex reform. This has led us to assess our own evaluation methods and to look beyond the traditional realist evaluation to find more suitable theoretical models to address the complex reform we are studying.

Incorporating complexity thinking into program theory models

Quite often realistically oriented evaluations are accused of taking context – what works, for whom and in which context(s) - too far. If everything is bound to the context at hand, evaluation as such cannot predict the consequences of any other programs or interventions that are thought to be similar in nature, since the context is never exactly the same. Hence, you cannot transfer evaluation results to any other context than the one you are currently looking at. (Davis 2005.) Still, from the perspective of complex interventions context is of importance, and this proves to be one the main strengths of realistic evaluations. Peter Dahler-Larsen (2001, 336) introduces the meaning of moderators and independent variables in his schematic model of program theory. Moderators and independent variables intervene in the anticipated program theory of an intervention, and hence the actual change mechanisms are skewed, and direct causal relationships become impossible. Moderators and independent variables are context specific, and due to this the relationship between mechanisms and anticipated outcomes is complex and varies by context. (Davis 2005.) We might use the term program field instead of a program theory or CMO configuration to describe the conditions for interventions (Dahler-Larsen 2001).

Patricia Rogers (2008) and Lene Holm Pedersen & Olaf Rieper (2008) have explored the possibilities of realist evaluation, program theories, and complexity. Rogers (2008) starts from the differentiation of simple, complicated, and complex problems by Glouberman & Zimmerman (2002) and introduces logic models for simple, complicated, and complex interventions respectively. Holm Pedersen & Rieper (2008) tested the applicability of realistic CMO configurations to a broad public sector reform and came to the conclusion that it works as a general analytical framework, but has to be substantiated by two types of theories: theories at the level of each specific intervention of the reform and theories at the meso-level that help to identify the institutional patterns behind various interventions.

If we adopt the concepts of complexity sciences into evaluation research, the program field is understood as the space of possibilities for the intervention (Mitleton-Kelly 2003). The space of possibilities includes not just the context and mechanisms, but also the inter-relationships, interactions, and inter-connectivity of elements. This goes far beyond describing the context, the mechanisms, and the outcomes, which can be accused of being too simplified and inflexible in complex systems and interventions (Davis 2005, 292; Holm-Pedersen & Rieper 2008). In a space of possibilities where the system is constantly on the verge of chaos (i.e. it is far from equilibrium or stagnation) emergent outcomes appear. Emergence is a process that creates new properties, qualities, patterns, or structures that arise from the interaction of individual elements. Emergent outcomes, patterns, and properties are greater than the sum of their parts, and they cannot be predicted by studying only the individual elements. Emergent outcomes cannot be returned to their individual elements. (Mitleton-Kelly 2003 & 2006; see also Rogers 2008, 38-40.)

Emergent outcomes are obviously those that we did not anticipate in our program theory, but the expected outcomes of interventions may also be characterized as emergent. In complex interventions, such as restructuring the whole municipal and service structure, emergent outcomes are bound to appear. As evaluators, this is where we run into trouble. We might understand and even describe the context of the intervention at hand using the terms of

complexity, but how to evaluate emergent outcomes? By definition, they cannot be predicted by studying the individual elements. The outcomes of an intervention are the result of complex inter-relationships, inter-actions, and inter-connectivity of individual elements, not a linear input-output model. Besides mapping out the individual elements, we should be able to identify the interconnections of mechanisms, processes, and the interactions and co-evolution of several systems, sub-systems, and environments. One should be able to see the invisible – the patterns of emergence. If we only follow the declared program theory, the unexpected and unanticipated outcomes that are observable only outside our frame of reference remain hidden.

Evert Vedung (2000, 212) identifies six explanatory factors in the process of evaluation: historical background, intervention design, implementation, addressee response, other simultaneous interventions and issue networks and other environments. All these combined in the same program theory, together with the elements of complexity described in previous chapters, creates a challenging situation for evaluators. We have tried to map the context of our evaluation practice in Figure 1, which clearly shows how difficult it is to identify the mechanisms underlying an extensive reform. The revised program theory model becomes rather exhaustive and loses the original idea of simplifying the reality, but at the same time it visualizes the actual complexity of the reform at hand. If we as evaluators omit some aspects of the reality, we only look at the reform from a very narrow perspective and by so doing we may, for example, be explaining the outcomes with wrong inputs or totally ignoring some unanticipated but relevant outcomes. A complex intervention cannot be predicted or explained by studying only the individual elements.

Aiming at a moving target

The evaluation of complex interventions is like shooting constantly at a moving target. It depends on your weapons and your own contribution whether or not you hit the target. Working in a larger team may give some advantage, since big game is rarely caught alone. Ray Pawson (2003) and Tom Ling (2012) suggest several tips evaluators may consider when studying complex interventions. The tips are summarized in Table 1. Next, we reflect these tips against our evaluation practice.

Staring it in the face or understanding the intervention's theory of change and its related uncertainties forms the basis for any evaluation. This means mapping out the potential conjectures and influences that might shape the intervention under investigation (Pawson 2003, 486). But, as stated earlier, in the case of complex interventions, the key dependencies and inter-connections of systems, sub-systems and environments that lie outside the formal structure of the intervention must also be identified (Ling 2012, 87). This is what we have aimed to point out in our revised program theory of municipal and service structure reform in Finland (see Figure 1). The intervention's 'theory of change' includes other elements than just context, anticipated mechanisms and processes, implementation or other input. Confusing variables come in several forms, such as rival reforms and development programs, changes in policies and political climate, or demographic and economic changes, and it takes a lot of effort to incorporate all these into same evaluation program. Theory of change should also include the interconnections of all these, as stated above.

Pawson 2003	**Ling 2012**
1)Stare it in the face	1)Understand the intervention's Theory of Change and its related uncertainties

2)Concentrate your fire	2)Collect and analyze data focused on key uncertainties
-	3)Identify how reflexive learning takes place through the project, and plan data collection and analysis to support this
3)Go back to the future	4)Understand what would have happened in the absence of the intervention
4)Stand on others' shoulders	-
5)Criss and cross	5)Build a portfolio of activities and costs
6)Remember your job	6)The evaluation judgment should not aim to identify attribution, but rather to clarify contribution

Table 1: Tips for evaluators of complex interventions

When collecting data, concentrating on the key uncertainties pays back. You should concentrate on the empirical efforts on the linkages you consider vital to the effectiveness of the intervention (Pawson 2003, 486). In practice, this has proved to be a complicated task. Inter-connectivity, inter-actions, and inter-relationships do not show up in statistical evaluation indicators or strategic documents. They may be observable on the boundaries where different programs, interventions, and reforms meet, but having forty case municipalities and social and health services as the focus, it is practically impossible to be present at every interaction. Here criss-crossing between horizontal and vertical levels of intervention, standing on others' shoulders, and building a portfolio of activities and costs associated with the intervention are helpful practical tips. (Pawson 2003; Ling 2012.) In our evaluation research program, we have incorporated several sub-modules under the same umbrella, which enables us to share the expertise of several universities, faculties, and disciplines. Several ministries take part in the study, accompanied by experts from the Association of Finnish Local and Regional Authorities. If we are able to combine all the relevant information into a coherent whole, we might start to approach the complex nature of the reform under scrutiny.

Both Ling (2012) and Pawson (2003) stress the meaning of live evaluation, but also its retrospective and prospective elements. In being prospective, both understanding of what might have happened *without the intervention* and different scenarios of what might happen when the intervention is successfully implemented are needed. Our problem as evaluators is always that observing live changes depends on several indicators that can only be studied in retrospect; this is the case in our own practice of following up the usage of welfare services. We can map the strategic, structural, organizational, and operational reforms and even changes, but the effects of these only come into play afterwards. Long enough follow-up time is a crucial factor in evaluation, but sometimes it is impossible, since the speed of reforms is accelerating.

Finally both Pawson (2003) and Ling (2012) remind us to remember our job. Our job should be more to ascertain the contribution of interventions (how reasonable is it to believe that the intervention contributes to the intended goals effectively, and might there be still better ways of doing this) than to identify what proportion of the outcomes was produced by the intervention (Ling 2012, 88). We will return to this in our discussion below.

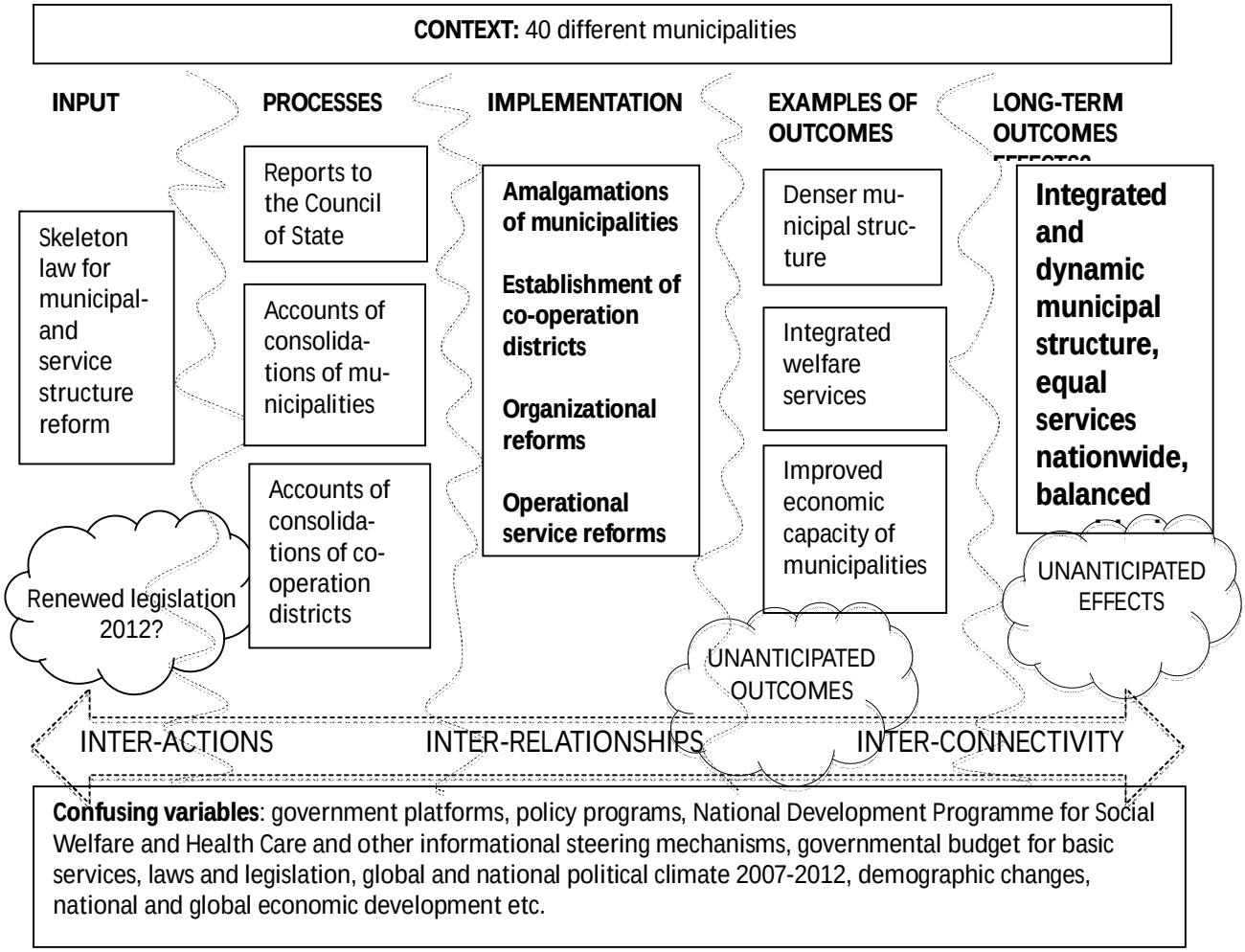
Discussion: to what extent we can say the reform 'works'?

In answering the question at the end of the previous section, 'how reasonable is it to believe that the intervention under study contributes to the intended goals effectively', we would like to introduce a somewhat revised version of the original program theory model (Figure 1). All reforms express a set of long-term outcomes or anticipated effects that comprise the goal for the reform. In the case of the Finnish restructuring of municipal and service structure, these include, for example, integrated and dynamic municipal structure, equal services to all citizens nationwide, and balanced municipal budgets, to name only a few. The implementation of the reform consists mainly of municipal amalgamations, the formation of co-operation districts for social and health service delivery, and a variety of organizational and operational reforms in single municipalities.

Processes that take place during the reform are of course more complicated than stated in the figure, but they mainly include different kinds of accounts of how municipalities plan to consolidate their structures. The context in our study consists of forty municipalities that differ in many aspects, such as size, population, geographic location, and the economic prospects for the future municipal budgets. The broken lines in Figure 1 demonstrate how contextual differences and several confusing variables may break the input-output chain. Rival reforms, such as reformed legislation, hover in the background. Unanticipated and unexpected outcomes and effects most certainly occur at some point. Finally, everything described above is made a little more difficult by the interaction, interrelationships, and interconnectivity of all the elements. This is the image we get while taking the complexity of interventions seriously.

When we combine the tips in Table 1 with the model in Figure 1, we start to extrapolate our evaluation practice. In practice, an evaluator often encounters criticism of his/her results. It is a common practice that the aims of interventions are expressed in a causal mechanism of input and desired outcomes: when X is put in place, Y happens. Nonetheless, the real logic embracing an intervention is far from linear causality, it is emergent. This is the dilemma or paradox inherent in the evaluation of complex interventions. The realizer of the intervention and/or the purchaser of the evaluation expect that after the evaluation has been done, it is proven whether the goals were achieved and the outcomes were caused by the intervention or not. In reality, the complexity of inter-connected elements and many simultaneous intervention mechanisms together make it (almost) impossible to prove any direct links between the intervention and outcomes, as illustrated in Figure 1.

Figure 1: A revised program theory model of the municipal and service structure reform in Finland (for the original model see e.g. Dahler-Larsen 2001).



As evaluators, we must ask ourselves whether we just monitor whether the desired outcomes were met, or whether we should try to understand the complex logic and processes of the intervention. A researcher needs to simultaneously analyze the context of operations, the mechanisms, the complex processes, and the interrelationships of broader targets with concrete measures, which do not necessarily present themselves in the outcomes (Hill & Hupe 2009, 163). In our evaluation research program, both the classical realist evaluation arrangement of 'what works, for whom, in which context(s)' (Pawson & Tilley 1997) and the painful fact that researchers cannot always reveal all necessary events or mechanisms occur in parallel. Our research on the municipal and service structure reform and social and health services shows that reforms first present themselves in organizational structures, strategies, and management systems. Long-term effects are detected in the operational arrangements, and in due course are reflected in the statistical evaluation indicators – but this takes several years. Social and health services are the arena for several simultaneous reforms, some of them even contradictory. Many paradoxical outcomes are present concurrently, and they cannot be traced back to any one of the reforms or interventions. This is what we have highlighted by taking the challenge of complexity seriously as part of our evaluation practice.

To some extent, the evaluation of complex reforms is a mission impossible. Complex

interventions contain many paradoxical elements. There are no unambiguous answers to give, but we might begin by mapping out and revealing the inherent complexity of reforms. Creating CMO configurations or program theories by mapping out the causal mechanisms of input-output chains is not possible when the context of evaluation is as broad and complex as described in Figure 1. An evaluator can detect only individual parts of the chain at a time and the broader picture may be lost. Similarly, by looking at the wider picture only, an evaluator may ignore some clear local input-output mechanisms. These two need to be balanced and this creates a dilemma for evaluators with limited resources. We are in a situation where "there has to be a theory behind all this, otherwise this all falls apart", as one manager of social and health services in our case municipalities put it.

References:

Act on Restructuring Local Government and Services 9.2.2007/169.

Barnes, Marian & Matka, Elizabeth & Sullivan, Helen (2003). Evidence, Understanding and Complexity. Evaluation in Non-linear Systems. *Evaluation* 9(3), 265-284.

Dahler-Larsen, Peter (2001). From Programme Theory to Constructivism: On Tragic, Magic and Competing Programmes. *Evaluation* vol. 7 (3), 331-349.

Davis, Paul (2005). The Limits of Realist Evaluation: Surfacing and Exploring Assumptions in Assessing the Best Value Performance Regime. *Evaluation*11(3), 275-295.

Gershenson, Carlos (2011). *Complexity*. Draft entry for the Encyclopedia of Philosophy and Social Sciences (Sage). Retrieved September 7, 2011 from http://arxiv.org/abs/1109.0214.

Glouberman, Sholomon & Zimmerman, Brenda (2002). *Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?* Discussion Paper no. 8. Commission of the Future of Health Care in Canada. Retrieved from http://www.plexusinstitute.org/resource/collection/6528ED29-9907-4BC7-8D00-8DC907679FED/ComplicatedAndComplexSystems-ZimmermanReport_Medicare_reform.pdf)

Hill, Michael & Hupe, Peter (2011). *Implementing Public Policy* (2nd ed.). Los Angeles: Sage.

Holm Pedersen, Lene & Rieper, Olaf (2008). Is Realist Evaluation a Realistic Approach for Complex Reforms? *Evaluation* 14(3), 271-293.

Holland, John H. & Miller, John H. (1991). Artificial Adaptive Agents in Economic Theory. *American Economic Review.* Papers and Proceedings 81 (May 1991), 365-370.

Ling, Tom (2012). Evaluating complex and unfolding interventions in real time. *Evaluation* 18(1), 79-91.

Mitleton-Kelly, Eve (2003). The Principles of Complexity and Enabling Infrastructures. In Mitleton-Kelly, Eve (ed.): *Complex Systems and Evolutionary Perspectives of Organizations: The Application of Complexity Theory to Organizations.* Pergamon, 23-50.

Mitleton-Kelly, Eve (2004). The Information Systems Professionals as a Hermit of plural rationalities, information rejection and complexity. *Innovation: The European Journal of Social Sciences*, 17(4), 289-323.

Mitleton-Kelly, Eve (2006). A Complexity Approach to Co-creating an Innovative Environment. *World Futures* 62(2006), 223-239.

Niiranen, Vuokko (2006). Reformia, rakenteita ja retoriikkaa. Julkaisussa Aarrevaara, T. & Stenvall, J. (ed.) *Kriittinen ajankuva*. Tampere: Tampere University Press, 66-81.

Pawson, Ray (2003). Nothing as Practical as a Good Theory. *Evaluation* 9(4), 471-490.

Pawson, Ray & Tilley, Nick (1997). *Realistic Evaluation*. London: SAGE Publications.
Raisio, Harri (2010). *Embracing the Wickedness of Health Care: Essays on Reforms, Wicked Problems and Public Deliberation.* Acta Wasaensia 228, Universitas Wasaensis 2010.
Rogers, Patricia J. (2008). Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions. *Evaluation* 14(1), 29-48.
Simon, Herbert A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society* 106(6), 467-482.
Vedung, Evert (2000). *Public Policy and Program Evaluation.* New Brunswick, USA: Transaction Publishers.

---

i

[1] This evaluation research project is a sub-project in a large research program "Research Program for the Evaluation of the Project to Restructure Local Government and Services in Finland" (ARTTU), which is evaluating the ongoing reform. The evaluation research program is carried out together with the Association of Finnish Local and Regional Authorities (AFLRA), several ministries and six universities, as well as with a sample of 40 municipalities and is financed by AFLRA, the ministries and the case municipalities. Our research group at the University of Eastern Finland is responsible for the study of social and health services. For more information see http://www.localfinland.fi/en/association/research/arttu/Pages/default.aspx.

VII

Evaluation in the use of Regional Development

Calculating income and employment for regional development practices in tourism – reliable, realizable, and continual procedure

Ari Karppinen;University of Turku,Turku School of Economics
Saku Vähäsantanen;University of Turku,Turku School of Economics
Teemu Haukioja;University of Turku, Turku School of Economics
Arja Lemmetyinen; University of Turku,Turku School of Economics

Abstract

Tourism has been one of the fastest-growing industries globally, and, nearly all regions have tourism at the core of their development strategies. On the other hand, empirical, but also academic, tourism research is in its infancy, not least because of lack of appropriate definitions for tourists, tourism, industries related to tourism and tourist products. Consequently, there is no single and generally accepted method to calculate tourism income and employment, which, in turn, makes inter-region comparisons unreliable. The purpose of this study is to construct an easily realizable, yet reliable, quantitative evaluation procedure of the economic effects of tourism in a region. Realizable means easy annual updating relative to its developmental purposes; reliability with quantification means that models are based on standard economics, the data is based on public statistics, and the procedure can consistently be transferred to any specific region. Because of excessive statistical lags for regional developing purposes, our procedure includes a short run foresight device for the next period. The calculation procedure is applied to the Satakunta region for the years 2009 and 2010. Economic knowledge produced by the procedure constitutes a solid basis for practitioners to develop and monitor regional tourism.

Keywords: regional tourism income, employment, multiplier, forecast, calculation model

Introduction

Tourism has been one of the fastest-growing industries globally. In the years 2000 to 2010, the real growth rate of the world economy was about 3.7% annually (IMF 2011). During the same period, the growth rate of global traveling measured as international arrivals increased 3.4% on average (UNWTO 2011). Especially in developed economies, the growth rate of tourism has been extensive, 5.6%. It is estimated that this trend is continuing. At present, the total income generated by tourism is about 11% of the GDP in OECD countries, and employment effects are also notable, since business in tourism is labor intensive (OECD 2010).

Tourism is not a well-defined industry. There are only a few commonly accepted features that describe tourism but no exhaustive definition which concurrently would be accurate enough for empirical research needs (Burkart & Medlik 1974; Vanhove 2005). Empirical and theoretical tourism research methods are heterogeneous, which also affects the contents of study-based practical regional development project reports and accounts in tourism, and complicates inter-regional comparisons. Because nearly all regions have tourism at the core of their development strategies, the need to develop consistent knowledge creation practices for regional development purposes is emergent.

The purpose of this study is to construct a practical procedure that generates essential quantitative knowledge about the regional economic effects of tourism for regional developing

purposes. The aim is to construct a reliable and easily realizable model that is simple enough to be updated continually (i.e. annually) to calculate direct and total regional tourism income and employment effects. In other words, we introduce a model by which direct and multiplier (i.e. indirect and induced) effects of tourism can be calculated. Concerning the basic regional tourism data, there is, however, an excessive statistical delay (more than a year) from the perspective of regional tourism developing, and, hence, the procedure for short-run forecasts (one year forward from the latest statistic release) is also introduced here. Annual updating of the model means that, in a year *t*, regional tourism income and employment can be calculated on the basis of released statistics for the year *t-2* and the corresponding foreseeable values for the year *t-1*, further, calculations in a year *t+1* make the earlier foreseeable values definitive. The pilot version of the procedure has been applied to the Satakunta region (Karppinen & Vähäsantanen 2011).

The paper is organized as follows: In chapter two, we present the research framework and some definitions. Chapter three describes the data. In chapter four, the procedure that generates regional direct and indirect effects with forecasts is represented. Chapter five gives the results for the Satakunta case, and chapter six, the conclusion.

Research framework and operational definitions

Measurement of the regional economic effects in tourism was originally developed in the late 1950s and the early 1960s in Sweden (Frimodig 1959; Eriksson & Wikström 1961), and the pioneering studies in Finland were published in the 1960's and the early 1970's (cf. Kauppila 2001, 10). Since then, at least fifty regional, municipal and tourism center-level studies – implemented in varying methods if even reported – have been published in Finland. Thus, comparability and scientific reliability remain obscure (see Karppinen and Vähäsantanen 2011, 15). This paper attempts to accurately obey the applied research methods, while following the typical Nordic tourism research framework (see Kauppila 2001).

Quantitative evaluations of tourism's economic effects on a national level are often based on input-output analyses (e.g. Salma 2002, on criticism see Dwye, Forsyth & Spurr 2004). Regional evaluations are heterogeneous, because typically, there is no data available for input-output analyses (e.g. Walpole & Goodwin 2000). However, as opposed to earlier studies, we are able to evaluate regional economic impacts on tourism, because Statistics Finland (SF) provides input-output matrices for Finnish regions (NUTS 3 nomenclature in the EU). It is speculative to estimate municipal effects with the available information about the regions, because there is no guarantee that regional multipliers present municipal structures. Municipality is such a small unit that regional leakages may be so huge that induced multiplier effects are reduced to a minimum.

There are some provisos that should be mentioned. SF does not publish the Regional Tourism Satellite Account on a regular basis. Consequently, it is not available for the years 2009 and 2010. Also, annual variations in economic effects of single or regular recreation or cultural events are not recognized completely from the aggregate data. These variations are captured over annually changing income and employment data, but not over tourism income and employment shares of different tourism-related industries. Since we consider only commercial tourism income, the input value of owning holiday homes and cottages are omitted; but naturally, the use of local services by the owners is included in the revenue figures.

The operational definitions of the study are based on practices in SF which follow international recommendations for tourism statistics (IRTS 2008). The tool for organizing data resembles the product-industry matrix that is provided by IRTS (2008, p. 51). The matrix

describes relationships between tourism industries, other industries, tourism characteristic products, connected products and other products. The matrix summarizes the total output by products and industries. Originally, this structure is constructed for national accounts of tourism, but with some further information, it is also applicable to regional dimensions. Konttinen (2006) provides estimates for the income shares generated by tourism in Finnish regions.

Data

The data consists of tourism products and tourism industries, which are based on regional tourism account definitions made by SF (see Konttinen 2006, p. 53–58, Statistics Finland 2005). Tourism products are classified into three categories. (1) Tourism characteristic products include accommodation services, catering services, passenger transport services, travel agency, tour operator and tourist guide services, cultural services, recreation and entertainment services, and other tourist services. (2) Connected products include fuel trading (brokerage) and personal local transport. (3) Other products include wholesale and retail trade (commission), and other products. The shares of tourism products for different tourism industries are calculated by the Tourism Satellite Account (RTSA) concerning the Satakunta region (Konttinen 2006). Since then, the Standard Industry Classification (SIC2002) has changed (SIC2008). We have applied the proper classification modification developed by SF. Due to data availability, there are a few provisos: The economic rent of second-home ownership is not taken into account; in cultural services, only market-valued services are included. In the paragraph 'Other tourist services,' only goods rentals are taken into account. Measured tourism income and employment variables are the business revenues (€) and person-years, respectively. This data is based on the Register of Enterprises and Establishments database (SF) in 2009. Tourism income and employment figures for the year 2010 are preliminary. These figures are forecasted from the 2009 figures with business cycle data concerning the year 2010 in Satakunta.

Deriving direct and total effects of tourism with short-term forecasts

Direct income and employment

Direct tourism income (TI_h) and employment (TE_h) for region (h = Satakunta) is calculated as follows:

$$\begin{cases} TI_h = \sum_i^n \sum_j^m \alpha_i X_{jh} \\ TE_h = \sum_i^n \sum_j^m \beta_i Y_{jh} \end{cases} \quad (1)$$

where

α_i = share of revenue generated by tourism concerning tourism related product i,
β_i = share of employment generated by tourism concerning tourism-related product i,
X_j = total tourism revenue in industry j, and
Y_j = total tourism employment in industry j.

Tourism products *(i)* and tourism industries *(j)* are defined by SF. Shares of net sales are estimated by Konttinen (2006)[1]. Since there is no data on employment shares, it is assumed $\alpha_i = \beta_i$. It means our TE_h does not actually overestimate tourism employment effects since typically tourism is a labor-intensive industry.

Regional multiplier effects of tourism

Regional income and employment effects of tourism do not include merely direct effects, but they also cause multiplicative effects (i.e. indirect and induced effects) on a regional economy. Indirect effects are based, firstly, on interdependencies of different industries on the regional economy. Demand caused by regional tourism does not only affect the tourism industries, but also businesses that offer products for tourism enterprises and their business service providers, etc. Hence, tourism income circulates on the regional economy, but in such a way that only that portion of the additional income is taken into account, which is locally produced. In other words, there are leakages resulting from the fact that some of the intermediate products of tourism businesses are purchased from outside the examined area. Another form of leakages are indirect taxes which typically are directed away from a region to the state, albeit some part of this tax income will, further, be returned to the local economy by transfers between central and local governments. Secondly, in the case of tourism, it can be assumed that large amount of the additional demand relates to a local supply, and, hence, leakages are smaller than, for example, in export-led manufacturing. Particularly, this is the case in such tourism products in which value added is based on locality.

Induced effects of tourism arise in consequence of the direct and indirect effects. These effects increase local formation of income (i.e. wages and profits), which in part is directed to additional demand for locally produced goods and services. The amount of leakages in that case depends especially on the characters of a region. A large region, or a region with tourism attraction, has typically bigger multiplier effects. Also, the induced income effects circulate with annual leakages in the regional economy.

Since there is no updated data on industry-level interdependencies concerning the Satakunta region, we have not disaggregated direct tourism income to different industries in Satakunta, but we have calculated the aggregate multiplier. It captures both the indirect and induced effects explained above. Since there are leakages from regional economy, the result of the multiplier effect is the sum of geometrical series, and the multiplier, in its general form, can be presented as follows[2]:

$$k = A\left[\frac{1}{1-[B*C]}\right], \qquad (2)$$

[1] Konttinen (2006) has calculated product-industry matrix (i.e. domestic supply and internal tourism consumption by products) for tourism at the regional level in Finland. We apply the matrix for Satakunta concerning share of revenues *(α_i)*. Since Konttinen, the standard industry classification (SIC2002) has changed in 2008 (SIC2008). In order to reconcile obvious mismatches between the industry classifications, we apply the industry classification key produced by Statistics Finland.

[2] The multiplier is explicitly introduced in Karppinen & Vähäsantanen (2011, 56–59). The general introductions of the Keynesian multiplier effect can be found in many macroeconomics text books. Corresponding multiplier models concerning particularly tourism are introduced in tourism text books, which emphasize quantitative methods, like Ryan (2003), Vanhove (2005), Tribe (2011).

where

A = (1-L) = share of direct tourism income which remains after the first-round leakages *(L)*,
B = share of income of which local entities (residents, firms and local public sector) consume locally produced goods and services.
C = share of local consumption of entities which increase local income.

Since regional multiplier effects are always region specific, Equation (2) needs to be calibrated to concern the case of Satakunta. Also, the modeling-specific calibration is conducted: Since we apply the net sales variable[3] in calculating direct tourism income, and typically the Keynesian multiplier is derived using the value added variable, we apply additional leakages in order to avoid double counting[4] (table 1).

Table 1. Calibration of regional income multiplier for the Satakunta region

Parameter	Calibration value*	Max value	Min value
*L***	0.30	0.40	0.25
B	0.75	0.80	0.70
C	0.65	0.70	0.60
*k****	1.37	1.70*****	1.03*****

Notes: We apply calibrated values instead of scenarios (min-max) so regional developing practitioners can better follow yearly the economic effects of regional tourism (income and employment), ** A = (1-L), *** k = (1-L)/(1-BC), and **** max/min value for A in calculation.*

Source: Karppinen & Vähäsantanen (2011, p. 32)

The proportion on income of which local entities consume locally produced goods and services is based on the assumption that local demand follows average demand structure and willingness to consume as in Finland on average[5]. Finally, parameter *C* is estimated by using assessments from regional industrial structure, regional productivity (reflecting income generation potential), and indirect taxes (as leakages from the regional economy)[6]. Table 1 shows that the tourism income multiplier in Satakunta is sensitive to parameter values: max multiplier is over 1.7 and min value is nearly 1.

Corresponding employment effects are estimated by using constant labor productivity assumption. In other words, the relation between direct tourism income and direct tourism employment is assumed to be constant for the multiplier effects. This is not necessarily an

[3] For practical reasons, we apply net sales instead of value added: Statistical time lag for data on net sales at region-industry level is smaller, and hence tourism income and employment estimations are more updated to the present. Secondly, by using net sales instead of value added we can consistently apply regional-industry business cycle data in order to make short-term forecasts (1 year) to regional tourism income in Satakunta.

[4] The range of leakages *(L)* is based on fluctuations on gross income value added ratios at industry level concerning Satakunta (i.e., *L* is not constant between industries).

[5] Conventionally, regional income shares are lower than corresponding country-level shares because of the larger leakages. However, Satakunta has quite a similar industrial structure as Finland on average.

[6] According to the Herfindahl-Hirschman index, which measures the diversity of regional industrial structure, the Pori sub-region (LAU 1) stands at second place among all sub-regions in Finland (77 in 2006) (Karppinen, Oikarinen & Kaivo-oja, 2010, p. 174). Industries in Satakunta are closely interlinked, and thus leakages (after the first income round) are smaller than in more specified regions.

accurate proceeding, since there are differences in productivity between and within industries. Considering the former, tourism typically possesses lower productivity (labor intensive) than manufacturing. The economic structure in the Satakunta region is industry-intensive, and, thus, our employment estimates for tourism may be exaggerated. Relating productivity differences within industries, our assumption, $\alpha_i = \beta_i$, means that the direct employment effect of tourism is underestimated if net revenue shares *(α)* are smaller (on average) than corresponding employment shares *(β)*, i.e., $(\alpha < \beta)_{|\alpha,\ \beta\ <\ 1}$. That is, tourism products within some industries are more labor intensive than other products in the same industry. Hence, the possible overestimation in the employment multiplier impact is mitigated because of underestimation in direct employment shares.

Calculating short-term forecasts

Due to statistical time lag, direct and total income and employment effects at the regional level are lagged over a year. In many cases, regional agents want more timely data for their purposes. Hence, we forecast the values for one year ahead for the tourism income and employment. If the above calculations are repeated yearly, the forecasted values may be cross-checked every year after. Let's denote in the Equation (3) industries *(j)* which are typical for tourism by superscript *tour*. The anticipated values for direct tourism income and employment effects at time *t+1 (t* = the latest statistical year) are calculated as follows:

$$\begin{cases} TI_{h,t+1} = \bar{\chi}^{tour}_{t+1} TI_{h,t} \\ \\ TE_{h,t+1} = \bar{\upsilon}_{t+1} TE_{h,t} \end{cases} \tag{3}$$

where

$$\bar{\chi}^{tour}_{t+1} = \frac{1}{m}\sum_{j=1}^{m}(\chi^{tour}_{j,t+1}) \ \ and \ \ \bar{\upsilon}_{t+1} = \frac{1}{m}\sum_{j=1}^{m}(\upsilon_{j,t+1}),$$

$$\chi^{tour}_{j,t+1} = \frac{X_{j,t+1} - X_{j,t}}{X_{j,t}} \ \ and \ \ \upsilon_{j,t+1} = \frac{Y_{j,t+1} - Y_{j,t}}{Y_{j,t}}.$$

In other words, the preliminary value *(t+1)* for direct tourism income for a year after the latest statistical announcement *(t)* is calculated as the average growth of net sales of industries. Due to a lack of data, we apply the average growth of employment of all industries in Satakunta at the time *t+1* in calculating the corresponding value for direct tourism employment. Further, the anticipated total tourism income and employment at the time *t+1* in Satakunta is obtained by exploiting the multiplier effects.

Results

The results for the case of Satakunta are presented in Table 2. The direct income of tourism in

Satakunta is 180 million € in 2009, and the corresponding employment is 1610 person-years. The income is slightly less than 1.5 % from total net sales of Satakunta, and the employment is slightly over 2.5 % from the total private sector employment. As we take into account the multiplier effects, the total tourism income in Satakunta increases to 243 million € in 2009, and the corresponding employment grows to 1890 person-years. In the case of Satakunta, the multiplier effects are substantial. This finding is as expected because of the diverse industrial structure and close interconnections between industries in Satakunta.

Table 2 shows that forecasted values for tourism income are increasing. The nominal growth of direct income is 4.4 % in 2010. As we take into account the multiplier effect, the corresponding growth is 4.5 %. Instead, the growth of tourism employment is slightly negative both in direct and total effects, -2.5 % and -1.0 %, respectively.

Table 2. Tourism income and employment in the Satakunta region, 2009 and 2010

DIRECT TOURISM INCOME AND EMPLOYMENT

Variable	2009	2010*	Change 2009-2010
Income** (million €)	180	188	8 (+4.4%)
Income (% of total income in Satakunta)	1.5		
Employment (person-years)	1610	1570	-40 (-2.5%)
Employment (% of total firm sector employment in Satakunta)	2.5		

Total tourism income and employment

Income** (million €)	243	254	11 (+4.5%)
Employment (person-years)	1910	1890	-20 (-1.0%)

*Notes:*Preliminary values, **Net sales* *Source: Karppinen & Vähäsantanen 2011, p. 34–40.*

Discussion

The primary purpose of the study is to construct a simple model framework for the assessment of an annual tourism income and employment at the regional level. The pursued properties are the following: It is relatively easy to update, and it gives sufficiently reliable estimates for practical information needs for the regional tourism development interest groups. Direct and indirect effects of tourism at the regional level are estimated. The procedure is based on regional economics and tourism research, and it is applied to the Satakunta region in Finland. Regional applications require some calibration, but this task is relatively fluent, since major modifications consider the region's industrial structure and business cycles. The model can be used to monitor the progress with just over a one-year delay. Estimates of near-future development may be presented with about four months' time lag.

Without a doubt, this preliminary version may need more or less revision and extensions. We recognize a few. In the present study, employment estimates were derived by assuming shares of employment in tourism industries (unknown) are equal to shares of business revenues (known) in every tourism-related sector. The assumption is naturally accurate for all industries which produce, almost exclusively, tourist goods and services (i.e. typical tourist industries). The

inaccuracy may arise for those industries in which tourism income shares differ significantly, and where labor productivity differs between the firms related to and firms not related to tourism. Despite the apparently clear outcomes from the present procedure, they are restricted to the short-term economic variables. However, modern regional development often requires recognizing sustainable development. Recently, Karppinen, Vähäsantanen, Lemmetyinen & Haukioja (2012) have extended the basic model presented here by taking into account the ecological pressure of regional tourism.

Acknowledgments

The authors would like to thank two anonymous referees for helpful comments and suggestions.

References

Burkart, A.J. & Medlik, S. (1974). *Tourism: Past, Present and Future*. London: Heinemann.

Dwyer, L., Forsyth, P. & Spurr, R. (2004). Evaluating tourism's economic effects: new and old approaches. *Tourism Management* 25, 307–317.

Eriksson, A. & Wikström, U. (1961) (in Swedish). *Turismen i Kiruna*. Kiruna.

Frimodig, L. (1959) (in Swedish). Turism i Bohuslän – en ny industri. *Meddelanden från Handelshögskolan I Göteborg. Geografiska Institution Nr. 58*.

IMF (2011). International Monetary Fund: Data and Statistics 2011. Retrieved July 6, 2011 from http://www.imf.org/external/pubs/ft/weo/2011/01/weodata/weorept.aspx?sy=2000&ey=2010&scsm=1&ssd=1&sort=country&ds=.&br=1&c=001&s=NGDP_RPCH&grp=1&a=1&pr.x=27&pr.y=13.

IRTS (2008). International Recommendations for Tourism Statistics. Retrieved May 30, 2011 from http://unstats.un.org/unsd/trade/IRTS/IRTS%202008%20unedited.pdf.

Karppinen, A., Oikarinen, E. & Kaivo-oja, J. (2010) (in Finnish). Olkiluoto 3-ydinvoimalaitosyksikön rakennusprojektin alueelliset tuotanto- ja -työllisyysvaikutukset, *Kansantaloudellinen aikakauskirja* 2010(2), 171–186.

Karppinen, A. & Vähäsantanen, S. (2011) (in Finnish). Matkailutulo ja -työllisyys Satakunnassa, Porin seutukunnassa ja sen kunnissa 2009 ja 2010. *Turun yliopiston kauppakorkeakoulu, Porin yksikkö, julkaisusarja A*, A38/2011.

Karppinen, A., Vähäsantanen, S., Lemmetyinen, A. & Haukioja, T. (2012). *Calculating global impact of sustainable tourism. Case Satakunta Region in Finland*. AIEST Conference Submitted to the AIEST2012-conference in Thailand.

Kauppila, P. (2001) (in Finnish): Matkailun aluetaloudelliset vaikutukset: pohjoismaisen mallin matkailijatutkimukset. *Naturpolis, tutkimuksia* 3/2001.

Konttinen, J.-P. (2006) (in Finnish). Matkailun aluetaloudelliset vaikutukset – matkailun alueellinen tilinpito. *Kauppa- ja teollisuusministeriö, rahoitetut tutkimukset* 9/2006.

OECD (2010). OECD Tourism Trends and Policies 2010. Retrieved May 3, 2011 from http://www.oecd.org/document.

Ryan, C. (2003). *Aspects of Tourism, 11: Recreational Tourism: Demand and Impacts*. UK (Clevedon). Channel View Publications.

Salma, U. (2002). Indirect economic contributions of tourism, 1997-1998. *Journal of the Bureau of Tourism Research* 4(1), 59–61.

Tribe, J. (2011). *The Economics of Recreation, Leisure, and Tourism (4th edition)*. Oxford:

Butterworth-Heinemann.
UNWTO (2011). UNWTO Tourism Highlights 2011 Edition. UNWTO World Tourism Organisation, Retrieved July 6, 2011 from http://mkt.unwto.org/sites/all/files/docpdf/unwtohighlights11enhr.pdf.
Vanhove, N. (2005). *The economics of tourism destinations*. King's Lynn: Elsevier Butterworth-Heinemann.
Walpole, M. J. & Goodwin H. J. (2000). Local Economic Impacts of Dragon Tourism in Indonesia. *Annals of Tourism Research* 27(3), 559–576.

UNIVERSITY OF TAMPERE

UNIVERSITY CONSORTIUM OF PORI

Sponsors:

TAMPEREEN YLIOPISTON TUKISÄÄTIÖ
TAMPERE UNIVERSITY FOUNDATION

Työsuojelurahasto
Arbetarskyddsfonden
The Finnish Work Environment Fund

FEDERATION OF FINNISH LEARNED SOCIETIES

European Union
European Regional Development Fund

Leverage from the EU
2007–2013