

**Quality monitoring in transition: The emerging challenge of evaluating ‘translational’ research programs in academic biomedicine**

Alexander D. Rushforth

Sarah de Rijcke

Center for Science and Technology Studies (CWTS), Leiden University, the Netherlands

Corresponding Author:

Alexander D. Rushforth

Centre for Science and Technology Studies

Leiden University

P.O. Box 905

2300 AX, Leiden, the Netherlands

[a.d.rushforth@cwts.leidenuniv.nl](mailto:a.d.rushforth@cwts.leidenuniv.nl)

+31 71 527 5392

Word Count: 11, 170

## **Abstract**

Whilst the efficacy of peer review for allocating institutional funding and benchmarking is often studied, not much is known about issues faced in peer review for organizational learning and advisory purposes. We build on this concern by analyzing the largely formative evaluation by external committees of new large, ‘translational’ research programs in a university medical centre in the Netherlands. By drawing on insights from studies which report problems associated with evaluating and monitoring large, complex, research programs, we report on the following tensions that emerged in our analysis: 1) the provision of self-evaluation information to committees and 2) the selection of appropriate committee members. Our paper provides a timely insight into challenges facing organizational evaluations in public research systems where pushes towards ‘social’ accountability criteria and cross-disciplinary research are intensifying. We end with suggestions about how the procedure might be improved.

## Introduction

In OECD countries, publicly funded research is increasingly made accountable by way of periodic external evaluations (Dahler-Larsen 2012; Hicks 2012; Whitley and Gläser 2014). Various efforts to render large-scale university and publicly funded research programs (e.g. multi-university programs and centers of excellence) accountable have brought into the spotlight issues of how best to design and conduct evaluations of such multi-dimensional objects (Cozzens 1997; Hellström 2012; Klein 2008). Whilst periodic program evaluations can rely on a variety of methods, in the main they still tend to be administered through peer review (Feller 2013; Georghiou and Laredo 2006). The continuing reliance on this broad family of techniques to evaluate research programs has been met with growing interest in the efficacy of peer review in arriving at satisfactory judgments and information on which decisions are to be based. This is particularly the case in the context of ‘strong’ evaluation systems (Whitley 2007), where summative judgments of peer committees have direct consequences for institutional funding and benchmarking (Hicks 2012). Relatively overlooked in comparison has been the design and organisation of ‘weaker’ modes of institutional evaluation (Whitley 2007), which tend to emphasize opportunities for organizational learning afforded by more interactive peer review formats over interim periods (Hansson and Monsted 2012; Youtie and Corley 2011). As these peer review formats also require spending of public funds and take-up the time and effort of researchers and administrators, its processes also merit attention.

In this qualitative study we describe in detail the preparation and production of advice in a ‘formative’ peer review procedure, by focusing on the evaluation of an academic medical centre by external expert committees within the national research evaluation system of the Netherlands. The Dutch exercise - which all academic institutions must undergo every six years - can be characterized for the most part as a form of ancillary peer review: the peer review is used to impart expert advice on a number of dimensions relating to research performance, institutional management, and organizational change (Bozeman 1993). Whilst a small number of studies have looked at the application of ancillary formats in evaluating fundamental research programs (Hansson 2010; Rons et al. 2008; Westerheijden 1997), in this paper we explore what happened when the organization in question – the University Medical Centre – decided for the first time to focus this hitherto largely routinized procedure around six new ‘translational research’ programs. Translational research is an increasingly popular and consequential slogan. Among other things, it is being used to express a desire to confront obstacles in the organization and management of biomedical research said to inhibit translation of research into development of new drugs, therapies, diagnostics, or public health practices (Molas-Gallart et al. 2015). By drawing on insights from studies that report problems associated with evaluating and monitoring large, complex, research programs (Cozzens 1997; Hellström 2012; Lawrenz et al. 2012; Quinlan et al. 2008), we hope to make sense of challenges that may begin to confront (self-) evaluations of in-house translational initiatives in medical research institutes. This is a timely matter, as in the Netherlands - and indeed many

other OECD countries – a transitional move towards translational programs is well underway, with targeted priority setting, funding instruments, and explicit evaluation and accountability criteria introduced to steer academic medical institutions in such a direction (Macleod et al. 2014). In the Netherlands, University Medical Centers were initially conceived as organizational initiatives to stimulate interactions between academic research of university medical faculties on the one hand, and patient care and research in academic hospitals on the other. Yet by 2012 a sense of continued distance between academic research and patient care led the Ministry of Health, Welfare and Sport to instruct all University Medical Centers to prioritize around focused cross-disciplinary translational research programs, which would close-off ‘translation’ gaps between the research and healthcare being conducted within these organizations (NFU 2012). In practice this has led to pressure on research departments traditionally administered within large clinical divisions to collaborate and build bridges through participation in new disease-focused program structures (for instance by making access to centralized research funding and resources available only via the new translational programs). In light of such efforts at top-down steering of medical research organizations, at present it is not clear to which extent modes of evaluative expertise, criteria, and indicators established to evaluate more traditional discipline-based modes of research remain relevant to this new task. Given that the goals of translational initiatives typically cut across individual and institutional actors and between bodies of work (Molas-Gallart et al. 2015; Rey-Rocha and Martín-Sempere 2012), the issue of whether new methods and practices can or should be formulated to support or perhaps even overhaul existing forms of evaluation such as bibliometrics and expert peer review has gained importance (Hansson 2010; Hemlin and Rasmussen 2006).

In this study the decision taken by the University Medical Centre management board to focus on evaluating translational research programs meant a novel experiment for all concerned. As such our analysis provides an excellent opportunity to take stock of challenges encountered when initiating evaluation of translational initiatives at an institutional level. The management board aspired to depart significantly from long-established bibliometrics-informed peer evaluations of departments within the traditional disciplinary structures of clinical divisions. As a result, most of the elements involved in the procedure were caught-up in various states of transition between the existing routines for evaluating fundamental research and aspirations to evaluate how departments and programs performed with respect to translational research in the six disease areas. This state of flux could be observed across several layers of the Dutch evaluation system’s ‘machinery’ (Lamont 2009). For example, the organizational structure being evaluated and the very knowledge being produced within the organization were in a more-or-less advanced state of transition. At the same time the UMC staff responsible for designing, organizing, and participating in the self-evaluation and site visits, and external independent committee members, were all participating in the new-look procedure for the first-time. Moreover, the formal evaluation guidelines of the Netherlands – *the Standard Evaluation Protocol* – which traditionally has

been geared more towards evaluating fundamental research programs, provided little instruction on such an experiment. Whilst our study is about an experimental evaluation procedure in a particular organization and country, difficulties described with respect to evaluating ‘social’ accountability criteria and cross-disciplinary research programs are clearly of relevance beyond this particular site and indeed this particular national setting (Bozeman and Boardman 2009; de Jong et al. 2016; Samuel and Derrick 2015).

The structure of this paper is as follows. In the following section we provide an overview of literature on organizational evaluations from which our analysis has benefitted, especially studies laying-out common challenges that can thwart formative, feedback-oriented forms of peer review. We then outline the scope of our fieldwork, and the data and methods on which our findings are based. Before presenting the findings, we sketch out further the design of the Dutch university research evaluation system and how the main guiding document – *the Standard Evaluation Protocol* – was translated by the management board of the UMC in setting up this evaluation procedure. Our findings are grouped along two themes which emerged as especially significant in the course of committee deliberations and in subsequent interactions with informants: 1) the production and presentation of self-evaluation and self-reported information produced prior to site visits by UMC staff for expert committees and b) the selection of peers and decisions to depart from routine methods for evaluating fundamental research programs (namely professional bibliometric reports) and the move towards narrative impact studies loosely influenced by the UK Research Excellence Framework (REF). Whilst these issues are often broadly found across all kinds of panel deliberation processes (Lamont 2009; Langfeldt 2004; Lawrenz et al. 2012), we aim to explore the distinct characteristics these issues took on in this experimental **translational** evaluation context.

Although there are no doubt more issues that could pose pertinent challenges towards ancillary peer review in a medical research context, the information committees receive, methods of evaluation, and selection of peers have the making of recurring struggles actors will likely face when confronted with the task of evaluating translational initiatives at the organizational level. Expanding ancillary peer review formats to evaluate translational research programs requires greater endeavor and risk compared with more routine bibliometric-informed procedures typical of evaluations in Dutch University Medical Centers to date. The problems we address in our findings sections make clear some of the emerging risks and challenges. For example, insufficient information for expert committees that the organizational staff generated during the self-evaluation phase shows the sheer difficulty in self-reporting about what are rather novel, experimental organizational structures in medical research. For these reasons a priori guidelines may help guide self-reporting procedures (particularly those sensitive to process evaluations, see Hellström 2012; Molas-Gallart et al. 2015). Second, in considering what constitutes favorable characteristics of committee members, we suggest that there are at least three important qualities of

experts in evaluating these kinds of program structures: 1) flexibility towards local circumstances, 2) responding quickly towards what the organization requests, and 3) knowing **when** certain performance indicators are (not) appropriate. For medical research organizations interested in improving translational ‘proximities’ (Molas-Gallart et al. 2015) and more broadly research leaders interested in improving internal managerial processes, appointing committee members with experience at cultivating or managing strategic research programs may be an advantage. Developing further guidance about committee selection would be a valuable addition to the formal evaluation protocol document of the Netherlands.

### **Ancillary Peer Review of Research Programs**

Although our main focus and contribution is towards problems of evaluating translational research initiatives at an institutional level, insights from research literature on organizational peer review proved instructive when interpreting and coding our materials. In this section we provide a brief synopsis of literature that helped us to frame the challenges observed, and we also return to the studies throughout our later analysis.

To reiterate in more detail a point already made, the Dutch peer review procedure is quite different from equivalent procedures for auditing university research in countries like the United Kingdom (van Leeuwen 2004). According to one prominent classification of research evaluation systems in OECD countries, the Dutch procedure can be considered ‘weak’: it is neither summative, strongly standardized, nor transparent (Whitley 2007). Periodic external peer review of Dutch university programs and institutions does not automatically determine funding distributions, but instead aims to guide ‘quality improvement’ decisions taken locally by research management (Van Der Meulen 2007). The evaluation format delivers reports in the form of ratings, narratives, and supporting evidence produced by independent external experts which equip research managers with ‘solidly legitimate arguments on which to base strategic decisions’ (Westerheijden 1997, 397). In contrast with peer review formats mobilized in ‘strong’ evaluation systems like the UK Research Excellence Framework (REF), relatively little attention has been given to design and organization of weak evaluation systems where research programs undergo external ancillary reviews. Although perhaps not be the most visible issue debated in research evaluation, ancillary peer review-based evaluations of programs have become steadily more routinized and established in certain research systems since the 1980s (Daniel et al. 2007).

Generally a number of issues in ancillary peer review can be associated with generic aspects of committee behavior, such as bandwagon effects, group think, and grandstanding (Bozeman 1993). As a result, several observations of committee behavior in well-known studies such as Lamont’s (2009) can be expected to be recognizable within our own materials, although her study was focused on another area of the sciences (social sciences and humanities) and was operating with a different peer review format (project selection by committees). For example, questions concerning the extent to which

committees feel the information they have been provided offers sufficient context (Bozeman 1993; Cozzens 1997), or whether short site visits by external parties are able to give anything more than ‘impressionistic’ accounts of large-scale complex program activities will likely persist (c.f. Lawrenz et al. 2012; Patton 2015), particularly if committees are composed of international peers with little knowledge of the local research system they are evaluating (Langfeldt 2004). Research on evaluations of fundamental research programs has shown that effectiveness of committees in imparting useful advice is likely to be highly dependent on appointing appropriate experts (Langfeldt 2006). In the context of fundamental research, previous studies have therefore recommended appointing individuals with appropriate **disciplinary** knowledge (Cozzens et al. 2002; Hansson 2010).

At the same time, some issues at stake in studying ancillary peer review are not generic to all kinds of peer review format. Rons and colleagues (2008) suggest the need for focusing upon aspects of design and organization that facilitate or restrict production of meaningful and useful advice for managers and researchers. This concern is very central to our analysis. Furthermore, whilst literature on ancillary peer review has tended to focus on evaluations within a disciplinary program context (Hansson 2010; Quinlan et al. 2008), the extent to which the same kinds of problems persist when ‘translational’ programs become objects of evaluation is an important but hitherto under-explored question. The tensions between existing components and routines around evaluation machineries on the one hand and new imperatives to promote translational research at an institutional level on the other form the backdrop to our analysis. In the first section of findings we report on how, in light of this translational experiment, one aspect of the Dutch ‘machinery’ – the self-evaluation and self-reporting of information by staff to expert committees – generated information of variable detail and coherence. This had an important bearing on reviewers’ abilities to impart useful advice for research management and program leaders. The second part of our analysis addresses how another part of the ‘machinery’ – the expert committees themselves – responded to the absence of bibliometric reports (the usual de facto method for evaluating quality of fundamental research in biomedicine). Whilst committee selection also remains highly consequential in this procedure, our findings lead us to question the extent to which disciplinary expertise is still the main criterion on which selection of peers should be based when the goal is to evaluate translational research programs.

## Methods

Our collection of empirical materials emerged from a combination of methods. Firstly, we traced debates on organizational evaluations within research policy and evaluation literature, across key titles including *Science and Public Policy*, *Research Policy*, *Research Evaluation*, *Social Studies of Science*, and *Evaluation and Program Planning*. We also read through national level research policy documents

on evaluations in the Netherlands and elsewhere, as well as expert reports by organizations like *OECD* and *Technopolis Inc.*

Secondly, between April 2014-February 2015 we conducted ethnographic fieldwork in a University Medical Centre, as part of a larger comparative project into the epistemic effects of evaluation and indicators on Dutch medical research. During this period we gained close access to this particular centre's institutional peer review procedure, carried out nationally every six years. For this we attended two site visits in April and May 2014: one of a committee assessing a strategic program and the other the main committee assessing the overall performance of the medical center. During the site visits materials were drawn from observing and audio recording meetings, accessing various documents including a committee secretary's written notes, and official strategic documents and previous self-evaluation reports made available to expert committees. We also took detailed fieldnotes of informal conversations between committee members, and between ourselves and committees during the site visit process. After the site visits we attended meetings with members of the UMC board and committee secretaries and also held a meeting in which we fed back our findings to the UMC board. Given the committee members were established scientists confident in their abilities to spot research quality and given the time constraints of the site visits, it is highly unlikely our presence was significantly disruptive. Accessing these usually hidden settings provided a rich amount of qualitative data revealing what committees and senior members of the research organization declared challenging in the course of formulating their advisory reports.

Thirdly, we draw on semi-structured interviews with the Dean of the Medical Centre and the organization's head administrator. These informants were selected due to their prominent roles within the organization and in shaping this procedure. We also conducted contextualizing interviews with two independent committee secretaries hired to coordinate the site visit and prepare the evaluation report, to take into account their experiences of staging this particular procedure and how this one differed (or not) from procedures they usually coordinate. Overall the guide was designed to gain insights into informants' perceptions of how the procedure was conducted and where things went awry or could have been improved. Our schedule for the interviews with members of the organization covered a number of topics, with overarching questions about the goals of the evaluation, how these were executed, the impact of the report on the organization, and how successfully informants viewed the experimental format to have been. In both sets of interviews, scheduled questions were followed by prompts and follow-up questions on developments we learnt about only during the interviews but deemed relevant. As is customary in qualitative research, names of individuals and the organization have been made anonymous in order to shield the identity of informants.

The audio recordings of meetings and interviews were transcribed in full and along with observation notes were coded for emerging themes in the Atlas Ti software. Our coding strategy took the form of



generating *pattern codes* in the data (Miles and Huberman 1994). In part inductive and in part informed through reading scholarly literature on research evaluation, the objective here was to identify emerging 'leitmotifs or patterns... discerned in local events and relationships' (Miles and Huberman 1994) which would provide relevant insights into the challenge of evaluating translational research programs. Earlier codes were refined through subsequent reading and analysis and through iterative reading of the fieldnotes and transcribed materials, and other data sources. As a result, **self-reporting and choice of methods** and **selection criteria of committee members** emerged and were selected as particularly prominent sensitizing concepts, which we present in the subsequent sections. Rather than aim for generalizability, our analysis is intended to open up a space for further conceptual and empirical probing into challenges cross-disciplinary translational research programs pose for actors made responsible for their evaluation. Before presenting these findings, for background information we now provide a quick sketch of the design and organization of this particular procedure.

### **Evaluation of the UMC using the Dutch *Standard Evaluation Protocol***

#### *Design*

As with other European research systems, over the past two decades the Netherlands has introduced its own distinct form of evaluation procedures and accountability measures. Every six years all academic research units must invite external expert committees to evaluate the research being performed in their organizations following procedures set-out by the *Standard Evaluation Protocol* (SEP) (KNAW 2009). This Protocol sets out the broad vision and purpose of its evaluations, as well as the formal procedures and criteria that must be followed. In the 2009-2015 protocol the formal criteria on which all units must be scored by committees were: **research quality, productivity, vitality and feasibility**, and **societal relevance** (KNAW 2009). The societal relevance criteria was introduced as explicit criterion for the first time in this version, marking the crystallization of a wider research policy transition towards social accountability in the Netherlands (de Jong et al. 2016; Van Drooge et al. 2013).

Although obligatory, the protocol's requirements are deliberately underspecified, granting significant discretion to those administering the procedure to design and organize the protocol's elements as they see fit. The fact that this is a peer review-driven procedure also means significant discretion is allotted to committees to translate the terms of reference provided by the organization in the course of their deliberations (Lamont 2009). This is a largely 'hands-off' procedure, in that beyond the fact of having a protocol which must be followed, few if any direct communications subsequently occur between the bodies organizing evaluations and the authorities setting rules and requirements. We shall briefly outline how the Protocol's template was translated in the course of the UMC evaluation.

#### *Letter to the University Rector and Committee Appointments*

All management board of listed university research units are contractually obliged to undergo *the SEP* procedure every six years. In the case of the UMC, the management board wrote a formal letter to the rector of the university, setting out their vision for the evaluation of six translational research programs by six separate external committees. The rector approved.

Leaders of each strategic program, in consultation with program-affiliated staff, then decided on the composition of committee members. Each committee was to consist of four-to-six members subject to availability. In an official *Terms of Reference* document sent to committee members before the site visits, the following criteria of selecting expert evaluators was explicated: “[they should] be distinguished scientists, have scientific experience and expertise outside their own discipline as well, [and aside from chairpersons and vice chairs] not work in the Netherlands” (Anonymised TOR Document). The rector or management board could overrule the proposed choices made by program leaders, for instance if they detected a conflict of interest e.g. the proposed choice being an extensive collaborator or former employee of the UMC.

At the start of each procedure an independent committee secretary must be hired from a professional evaluation agency in the Netherlands. This figure mediates between committee deliberations, *Terms of Reference* and *the SEP*; helps to organize self-reporting and site visits; records information from the deliberations; and drafts the final report.

### *Self-reporting*

Involving those who will be evaluated in the design and preparation is often seen as a tried-and-tested means of imbuing an evaluation procedure with enthusiasm and legitimacy (Bozeman 1993). The self-evaluation of strategic programs was overseen by the board of the UMC at arms lengths, with each program leader able to suggest composition for their own committees and coordinate self-evaluations and evidence submission in conjunction with selected researchers they nominated.

It was decided program leaders would take a selection of six or seven research departments to showcase in the self-reporting and site visit (modelled in part on the UK REF’s ‘impact case study’ approach). Program leaders thus nominated principle investigators from departments they considered their ‘best shots’ to appear before committees. This phase included a SWOT exercise (the common strategy tool designed to ‘brainstorm’ Strengths, Weaknesses, Opportunities, and Threats) and meetings to draw-up points for committees to focus upon. The protocol does not make explicit which particular kinds of indicators should be provided by the institution, stating only ‘The committee may use qualitative and quantitative indicators and indications’ (KNAW 2008, 8). Leaving self-reporting to the discretion of the program leaders meant inclusion of bibliometric reports was optional, although members of the board stated in interviews they had discouraged comparative bibliometric analyses of programs and groups in order for the evaluation to do more justice to ‘societal relevance’ and to the trans-disciplinary

organization of the programs and groups. Despite the consultations, program leaders and the UMC board held editorial control over what was finally presented to the committees.

### *Guidance to committees*

Instructions on how the management board wished the committees to execute the SEP were set-out within the *Terms of Reference* (TOR) document, which committee members received four weeks before the site visits. When comparing the 2009-2015 SEP and the UMC's TOR, despite many overlaps, we noticed some important differences in how the two documents presented formal evaluation criteria used for all research assessments nationally – **quality, productivity, relevance, vitality and feasibility**. Some differences are worth mentioning briefly, as they indicate how the UMC's management sought to recast some of the basic tenets of the national protocol. One very striking difference was how the documents recommended 'quality' be assessed: the SEP definitions of quality chimed with more familiar notions of 'scientific impact', with reviewers encouraged to consider: "Originality of the ideas and the research approach, including technological aspects; Significance of the contribution to the field... (Inter)national position and recognition; Prominence of the programme director and other research staff" (KNAW 2009, 11). In contrast the UMC's TOR included an expanded definition for the research quality category: "The scope of the term research is not limited to the research results. Research management, research policy, research facilities, PhD training and societal relevance of research are considered integral parts of the quality of work in an institute and its programs" (Anonymized TOR document). The inclusion of such additional criteria in the TOR is emblematic of efforts by the UMC board to move beyond 'traditional' approaches to **measuring** research quality in this type of assessment. Another marked contrast between the two documents was how each defined the category 'productivity'. Whereas the TOR emphasised productivity as being "judged in relation to the mission and resources of the institute" (Anonymized TOR document), the 2009-2015 SEP set-out a more straightforward counting approach for committees to follow, in the form of: "Scientific publications and PhD-theses; Professional publications; Output for wider audiences; Use of research facilities by third parties" (KNAW 2009, 11). The mediating role of TOR document is key for making sense of the experimental nature of this procedure and the problems we report in the analysis. This document is arguably as important locally as the national SEP standards.

### *The Site Visit*

A narrative-based self-evaluation report was sent to committee members and secretaries approximately two weeks before the start of site visits. The committees were also provided with mid-

term reviews and the previous *SEP* evaluation report of the UMC. The site visit for each program typically occurred over two back-to-back working days, with a two-week interval between each program committee visit. The format was partly interactive and partly based on closed discussions. In the morning of the first day the committees met behind closed doors with the management board for questions and answers over the vision of the six programs and sought clarifications on issues not clear from the self-reporting information they were provided. The committees also met with strategic program leaders and management separately, and with the principle investigators nominated by the program leaders. The first days concluded with internal discussions among the committees and an evening dinner between committees, UMC management board, and program leaders. The second days of site visits included a closed meeting with societal stakeholders, a ‘consulting hour’ with the UMC management board, preparation of advice, and finally PowerPoint presentation of the preliminary report in front of those who had participated over the two-day process.

For the main committee site visit, similar combinations of interactive and closed discussions took place with some of the above-mentioned participants. In the afternoon of the final day the preliminary report findings were presented in a large, public lecture theatre in the UMC, followed by a closing drinks reception.

### *The Report*

The final report, consisting of written feedback and scores on the program(s), is drafted by independent committee secretaries based on their notes and recollections from the site visits. During the writing process the secretaries send drafted versions to committee chairs for adjustments and corrections, taking between one and two months. The report must then be published on the university website. Typically the management board is expected to write and make publicly available a response to the content of the report. On this occasion the management board also met with the committee secretary and their team in order to discuss how the evaluation went.

The document was divided into sections, including a separate review for each of the six programs containing each committee’s scoring and a short narrative explaining the result. These six sections were accompanied by a separate section providing a similarly formatted review of the UMC overall, and a preface authored by the board of the UMC setting-out the vision of the programs and thanking the committees for their work. Each program review section was nominally written by the committee chair, and contained a short two-page summary giving ‘overall impressions’, including a preamble praising the work of each program. Each section then allocated a score on the *SEP*’s four criteria of **research quality**, **productivity**, **vitality** and **feasibility**, and **societal relevance** (KNAW 2009). The score is made on a simple scale ranging from one to five (one being poor, five being excellent). The score is accompanied by narrative feedback explaining the figure and where appropriate, where the program could be improved upon. At the end of each section was an annex of information. This made public the

composition of peers, with a brief biography of each committee members. In some program sections their biographies were coupled with information listing ‘top publications’, H-index scores, and funding and awards. The transparency mechanism of naming committees is designed to engender trust and legitimacy in the report, which is built in no small part upon the reputation and credentials of the independent peers (Westerheijden 1997). Also included in the annex were schedules for each site visit and a precis of main results of the SWOT analysis and self-reporting for each program.

Although the compositions of committees is made public, the text and scoring is written as a collective, thereby anonymizing input made by individual committee members to the report’s contents.

### *Changes to the Procedure from Previous UMC Evaluations*

In preparing the assessment, the UMC board attempted to make at least three alterations from previous assessment formats. First, assessing the strategic programs meant other parts of the organization - clinical Divisions, Departments, and their principle investigators - could no longer be the primary objects of committees’ evaluations. Whereas this had previously involved only a single committee to evaluate the whole UMC, on this occasion each strategic program was evaluated by its own specially appointed committee, with a main evaluation committee composed of program committee chairs invited back for a site visit to produce an overall evaluation of the UMC. Second, mimicking procedures being tried out around the same time in the United Kingdom’s research assessment system, the Board attempted to make narrative-based ‘impact stories’ the primary mode through which committees would receive information about the programs. Previous *SEP* evaluations in the UMC relied heavily on proxies for measuring short-term outputs, namely (comparative) bibliometric analyses of productivity and citation impact of divisions and departments, with some narrative text to support the scoring. The practice of hiring professional bibliometricians is typical of UMC evaluations across the Netherlands up until the time of writing and underscores an historic tendency in biomedical research for administering evaluation through numbers, particularly in the Netherlands (Rushforth and de Rijcke 2015). Third, for the first time ‘stakeholders’ such as representatives from patient groups, industry, and health insurance were invited to meet with expert committees and offer their accounts of interacting with the strategic programs. According to the Terms of Reference “Stakeholder representatives will play an advising role to the expert panel in developing an opinion on the impact element of the assessment” (Anonymous TOR Document). Only senior research managers from the clinical divisions met with committees on previous occasions.

Although the Standard Evaluation Protocol lays out a series of obligatory steps, in practice research organizations are granted significant discretion to appoint peers, perform self-evaluations, and influence the peers to focus on certain facets of organizational performance and activities over others. Given this jurisdiction for research organizations to act, the procedure we describe in this paper - although

experimental when compared with the routines Dutch medical centers have grown accustomed – did appear to comply with the protocol.

## **1. Preparing Self-Reporting Information**

An important part of the machinery of ancillary peer review in the Dutch research evaluation system is the self-evaluation and self-reporting phase. In this phase, organizational members under evaluation generate information on which expert committees are to base their judgments and comments, in conjunction with the contextual information they gather from site visits. In this section we report on some of the difficulties committees and independent committee secretaries articulated about the information received prior to site visits. In some ways hearing such complaints is not so unusual. The sheer scale of such research programs and the need to produce detailed advice within a short time-frame can risk ancillary peer review evaluations remaining rather impressionistic (Howell and Yemane 2006). Yet several of the problems we report in this section appeared to relate to the novelty of the translational programs and by extension the novelty of self-evaluation and self-reporting for UMC staff on these organizational structures. Criticisms of the information are broken down into two sub-themes: a) lack of information about coherent vision and goals of programs and departments, and b) difficulties disentangling overlaps across translational programs in terms of research groups, outputs, and research focus.

### *Lack of Coherent Vision*

During both site visits we attended it was made clear by committee members that some of the self-reporting on groups and programs had not produced coherent accounts of their future ambitions. As committees are instructed by the protocol to focus on qualitative information about policy and strategy in making advice about the institute and programs, the lack of such information being provided by program leaders was problematic: it inhibited the provision of formative feedback about certain programs in the final report. In the following discussions the main committee chairs are deciding how to score programs according the four criteria listed in the *Standard Evaluation Protocol* (see above), with objections being raised at various moments of the discussion that insufficient information was provided to make judgments on the ‘vitality’ and ‘feasibility’ of groups and/or programs (the more formative criteria).

Peer A: But that was the problem for [Program Z] – we didn’t have insight precisely about what are your future plans, what are you going after, is that creative? And we didn’t have that information.

One of the explanations committee chairs gave as to why some programs and departments provided stronger information was the fact that institutionally some had pre-existed the translational programs in the UMC, whereas others were built more-or-less from scratch. One of the programs reportedly provided more coherent information because it had evolved out of a small number of established departments in the old university hospital, taking the strongest research groups on certain diseases from this old structure and successfully applying for the status of a strategic program. For the committee this meant they already had a strong sense of what their focus and priorities were. Such a criticism was made by a committee chair about their program:

Peer A: Going over the other [program committees'] reports, there was a clear description about what their ideas were, what they have achieved in the past. But that was not done in our information ... There was only one exception it was [names group] – that was very well laid out – what they were planning to do, what they had performed in the past, and that was the only one, so we were struggling.

The group which this chair claims had provided well laid out information was in fact a long established research division within the academic medical school, first within the university and later within the UMC. It had an internationally renowned focus on a core set of themes and reportedly already resembled a strategic program (it was claimed by the group leader that the programs were themselves partly modelled on the group's structure and activities). The fact such programs were in a sense already established meant they also had a track record of international publications and grants to present before committees, as well as local infrastructure for performing translational and clinical studies with patients in the clinical divisions of the UMC. This suggests the intelligibility of self-reporting information may well in part be related to the phase of the research program or group and time lags it takes for tangible health and care outcomes to accrue from different research lines (Martin et al. 2004). This is a problem which may well surface in novel translational programs, particularly those focusing on frontier topics whereby clinical aims and promises are perhaps more speculative, or where there is less by way of existing capacity in relation to a given focal area.

The inability of program leaders to articulate clearly and persuasively future plans and missions can be consequential. In one instance the shortcomings of information were used by committee members a criticism of a programs' long-term viability (i.e. the fact it did not have a clear plan in place). The following point was made during scoring deliberations on the research programs during the main site visit:

Peer B: so we can go along the research programs and give groups 'satisfactory' and so on. But there are programs that do not fulfil the criteria nowadays: for instance [for Program X] you cannot be happy with the program at this moment because we do not see the future.

The poor outline of missions and strategies here led to discussion about ‘marking down’ particular programs. We observed that programs bearing closer resemblance to long-established departments fared better in committee discussions and indeed formal written feedback. As such program leaders and researchers may end-up being adversely affected by shortcomings in the information they provide to committees.

#### *Overlaps of groups across programs*

Another problem the self-reporting machinery of the Dutch peer review procedure faced in transitioning towards evaluating these new organizational structures came in presenting work done in each program in a way amenable to treating each as a bounded unit of assessment. Since the new translational initiatives had been initiated, research groups in the UMC had come to be affiliated across multiple programs, driven perhaps by curiosity to link up research with other areas, or by strategic reasons like fear of missing out on access to institutional funding and infrastructural resources. This is not ostensibly a bad thing as crisscrossing is something these strategic translational programs are supposed to facilitate and encourage. Yet this posed difficulties for evaluators, as in previous versions of the UMC evaluations departments’ research outputs were determined as belonging to a specific division through routines of bibliometric reporting.

A difficulty in articulating uniqueness of the new translational programs was that since the new translational initiatives had been initiated, a number of research departments (e.g. those working on stem cell research) had started to work across different areas of clinical need, meaning their research may fall within two or more of these new programs. This was not so much a problem for the actors on the ground than for committees trying to interpret programs as coherent units of accounting, whose activities, outputs, and personnel could be represented in coherent, standardized forms, with relations between groups, programs, and outputs clearly delineated. The decision to evaluate programs instead of divisions meant efforts to evaluate outputs like publications ran into difficulties, as authors from one group had affiliations with more than one program. The bibliometric reports usually conducted in advance of site visits are able to follow standards for identifying all outputs of groups (e.g. departments) and programs (e.g. divisions) and take steps to avoid duplication and overlaps, were no longer available for committees. The issue of multiple overlaps between groups and programs brought about by the novelty of having research programs as objects of evaluation appeared not to have been anticipated in preparing the information in advance of the site visits. Encountering overlaps between programs and departments may become more likely for committees when only a fixed number of clinical areas are prioritized as important to the UMC research. One specific problem this raised for committees was attributing which programs were responsible for looking after a given research line, given overlaps between groups and programs. Addressing other committee chairs, one of the chairs stated:



And looking at [program X], I was quite surprised because there is a lot of overlap between stroke [in this program] and [stroke in] the [program Y]. In relation when reading about [Program X] I thought ‘oh it’s in relation to the elderly and with genetic components’, but then they came up with bits of surgery, which is also in our program [Z]. (Peer A)

Some of the groups that were presented in self-reporting information as being part of the programs portfolio, provided little information or publications that were affiliated with that particular program’s research area.

As well as not being able to present a clear connection between groups and program themes, also problematic was the issue of determining the relative contribution by groups in programs to research involving collaborations with actors outside of the programs. For example, a prestigious laboratory outside the UMC program structure was suspected as being a creative driving force behind a set of publications and potential clinical advances, making the crediting of collaborators within the programs a sticky issue. This point was raised by a program chair in a consultancy hour meeting with the management board.

Where are the papers? We couldn’t sort them out – what was [produced by] UMC and what was [names a laboratory in the University’s Life Sciences faculty]? (Peer D)

This self-reporting problem appears to exemplify wider challenges of evaluating individual groups and staff in translational initiatives which spill across singular organizational boundaries, for example, universities, hospitals, and companies (Molas-Gallart et al. 2015). It is also perhaps illustrative of challenges that translational research poses to traditional bibliometric information sources as proxies for impact.

In this section we have focused on how the self-evaluation phase of the peer review machinery came under strain in the course of the effort to make a transition towards evaluating translational programs. On the basis of the observations we infer that problems in producing useful information for committees may be particularly acute where the staff are self-reporting for the first time on organizational structures that are themselves still work-in-progress, have multiple, ambiguous goals, and are the result of political negotiations (e.g. between existing departments and clinical divisions), and therefore may be lacking a single coherent research vision to present before committees. In previous exercises when research was grouped into more traditional disciplinary-oriented units of assessment (e.g. in divisions and departments) the evaluation process tended to be more routine: most research groups were formally affiliated within a single division of the UMC and their individual outputs like publications could be easily grouped into bibliometric forms of accounting. In experimenting with such a novel evaluation format, inexperience in self-reporting on these organizational initiatives combined with the ambiguous

character of how research is coordinated through translational research programs may well lead to shortcomings in the information that committees receive. Furthermore, formal criteria in the Dutch evaluation protocol appeared too broad to help overcome such challenges and provide committees with the information they wanted. Criteria specifically developed to evaluate translational research programs promise to go some way to correcting these issues (e.g. Hellström 2012; Molas-Gallart et al. 2015) although such models remain empirically largely untested.

## **2. Changing Methods, Retaining the Peers**

In asking expert committees to focus on the new translational research programs as objects of evaluation, the UMC board assumed that the type of knowledge under evaluation, and the kinds of knowledge produced by committees (or desired of committees), and methods through which committees could identify and judge quality, would all need to change from previous versions of this evaluation exercise. However, in this section we describe how bibliometric methods associated with evaluating fundamental research, both in Dutch UMCs in particular and biomedicine more generally, resurfaced during the exercise to the detriment of efforts to develop a more process-oriented translational approach. Whereas efforts to bring indicators back in do not completely negate other qualities that the respective committees may have contributed, such incidences raise a general query about the appropriate selection criteria for appointing individuals as experts for evaluating the kinds of activities and goals associated with translational research programs. Based on the findings we present below, we go on to consider some of the possible criteria and skills needed to evaluate translational research programs in such an advisory capacity.

Whilst some biomedical sub-cultures strongly resemble so-called ‘Mode 2’ research, others retain more traditional disciplinary approaches toward knowledge production (Boggio et al. 2016). As such, simple output indicators based on citations have become part-and-parcel of evaluating medical research in some epistemic sub-cultures of biomedicine and remain less important in others (Rushforth and de Rijcke 2015). Aside from technical limitations of popular indicators like the Journal Impact Factor or H-index, basic output indicators are not particularly useful as information when the purpose of an evaluation is to learn about process improvements to ongoing research programs (Cozzens et al. 2002). Nonetheless three of the six program committees were reportedly keen to use bibliometric publication information to judge the research quality of groups and their principle investigators. The board explained this to us in terms of different traditions of evaluation across committees linked to the profile of the programs: some covered research areas around an established disciplinary base, whilst others encompassed more recent transdisciplinary areas and were thus less aligned with citation indicators. Remarking on this dynamic in what he considered a more transdisciplinary programs, the Dean stated:

It has never been a discipline on its own; it’s truly a transdisciplinary group, they do completely different, new things. They do 3D printing, engineering on cartilage in your

knee, for instance, which never existed... This is what you see [in the transdisciplinary program committee], and so they looked very differently [at] the research that got presented to them, than for instance the classical guys [in more 'traditional' programs] ...who have many journals with high impact factors. (Dean Interview)

The program committee we observed was reported to be one of those most informed by disciplinary standards of excellence. This was confirmed during our observations of their deliberations, as indicated by the following conversation, which took place towards the end of the site visit:

Peer 1: I think in some ways it would be more helpful for the groups to have done more quantitative analysis – citations...

Peer 2: I echo that. We had that thought yesterday – you called it lay people writing ... but it's really difficult to follow quantitatively...

Peer 1: I mean in the areas I know well, I know how to rank them, but the areas I don't know well I do not, because there is not the citation analyses.

Peer 3: It's a terrible way of doing it, but it's the only way we have.

Peer 2: And even if [the Dean] doesn't like it...

Peer 3: If [the Dean] doesn't like it... well

There may well be an element of 'grandstanding' in such discussions, insofar as the committee members themselves had high H-index scores and often published in high impact journals, suggesting they would be confident that this is the appropriate criterion for defining successful individuals and groups (and that program evaluations should take this into account). There was also some grandstanding in terms of evaluating individuals outside of their specialties, on the basis of *where* they published, evaluated in terms of brand of the journal and/or its impact factor:

Peer 3: I was Pubmed-ing some of them, you know some of it was impressive; some of it was less impressive. You know is the highest level you are publishing in PlosOne? It's not a terrible journal, but it gives some indication unfortunately...

The availability of wireless internet access on their personal laptops also meant peers could go online to search informally for the H-index scores of group leaders through platforms like *Google Scholar*, *Scopus*, or *Web of Science* (which produce different versions of the score). Having all agreed that more quantitative publication information was needed on the groups, the committee requested the management board to provide a list of H-index scores for all principle investigators within their program. Given the lack of time afforded by the site visit, the H-index scores never arrived. The committee made

do without, but made critical remarks in their program report that ‘more systematic’ publication information should be provided next time, otherwise they could only provide ‘impressionistic’ feedback. In a separate committee, further publication information was requested, with a professional bibliometric report sent to committee members after the site visit. This meant this information was evaluated remotely rather than through deliberation. Much of the section of the final report for these programs where committees had requested publication information concentrated on the quantities of publications each group had in high impact titles.

To some extent a call for quantitative ratings is understandable, insofar as norms of fairness are often central to the legitimacy and confidence in the worth of peer review (Lamont 2009). To balance the powers of cognitive particularism in which peers are likely to look more favorably upon work in their own vicinity of research (Lamont 2009), ranking groups through standardized indicators may appear a reasonably fair and convenient solution (given also time constraints). Furthermore, shortcomings we reported in the previous section about qualitative information provided by program leaders (in collaboration with the UMC board and program staff) may make a clamor by committees for returning to the routines of bibliometric evaluations more likely. However, there are a number of reasons to dispute the assumption that H-index scores and JIF or brand name of where groups publish are an appropriate set of indicators to inform translational program evaluations. The preference for H-index and JIF-informed evaluation of small numbers of papers is itself arguably a form of cognitive particularism, insofar as it provides a rigid one-size-fits all measure which is insensitive to epistemic differences across projects and towards the aims of evaluating these particular groups and programs in terms of their processes and health-related impacts. Similarly excessive attention towards these indicators may risk committees becoming insensitive towards new forms of expert input which are increasingly valued in order to evaluate research quality in terms of broader social impacts (such as stakeholder testimonies) (see, Hemlin and Rasmussen 2006). Deferring to these kinds of indicators can marginalize alternative methods like narratives (unflatteringly referred to in an above quote as ‘lay people writing’), which were most probably more suited for evaluating what the board had asked to learn about, namely how to support research defined as ‘excellent’ in terms of impacting health and patient care issues within the UMC as an academic hospital (as was set out in the Terms of Reference document received by committees).

The UMC board was not in agreement with committees’ criticisms about the change in methods (this scope for disagreement is arguably an advantage of the Dutch system over more rigid mechanical procedures like the UK REF). The UMC board members explained this outcome to us as relating to selection of committees. Such complaints are often voiced within the research literature on peer review committees, whereby if individuals have been judged themselves according to particular criteria throughout their careers (and evidently succeeded according to them), then they may well wish to uphold such standards for measuring excellence across a range of evaluation contexts (c.f. Lamont 2009, 31-

32). The program committees that erred towards use of evaluative bibliometrics appeared to have been selected by program leaders at least partly on the basis of their publication records. For example, in two of the three committees in which desire to use quantitative information was reported as most prominent, the biographies contained in annexes of the final report listed the size of their publication output and respective H-index scores. Biographical sections on what were described to us as the more ‘transdisciplinary’ oriented programs did not include this information. Selecting peers with high citation counts within a particular field may be a logical means of selecting experts to evaluate disciplinary research initiatives, where such measures are often taken as votes from peers that they have successfully contributed to frontiers of knowledge over time and are therefore themselves likely to be able to exercise good scientific judgment (Lamont 2009). Of course this turn of events may also betray the fact that not all program leaders will necessarily always be in-sync with the management board’s aims of evaluating translational research programs in new experimental ways, and may for various reasons wish still to have their research evaluated according to more traditional bibliometric criteria by well-known scientists from their field. Given the interactive character of the Dutch evaluation this is likely to be an unavoidable trade-off, as involving those who are being evaluated is crucial for generating legitimacy of report findings within research organizations (Westerheijden 1997). Despite this possibility, it remains useful for those embracing evaluations of translational research programs to consider what constitutes sufficient criteria for picking out independent experts, and if not publication output or citations, then what other indicators of expertise can be used to inform selection? As these considerations were not obvious in advance - even for members of the board championing the new procedure - we return to these questions in our concluding section, where we suggest a number of idealized qualities that may be a useful to consider when appointing peers.

## **Conclusion**

Having reported some of the challenges that emerged in this shaken-up version of the Dutch evaluation format, we shall now outline our contributions to emerging interests in the challenge of evaluating translational research initiatives. We will also suggest some tentative steps that different actors associated with this kind of procedure might take to ensure that evaluations of translational research programs will run more smoothly.

In the first section of our findings on self-reporting, committees found a lack of contextual information about the relationship between the new strategic programs and the existing clinical divisions, which were highly consequential for how the former operated on a day-to-day basis. Addressing those who prepared the information (program leaders overseen by the management board), we found that although the divisions are no longer the primary unit of evaluation for this procedure, the committee members found them to remain an important presence in relation to the programs and therefore wished they had been included in information provided. We also saw that multiple affiliations of research groups and the

multi-level and multi-organizational divisions of research labor brought about difficulties in allocating credit to groups and programs and for pinpointing the sources of research quality – which is central to the objectives of the evaluations. The disconnect between information provided on this occasion and what the committees wished to know was in fact one of the criticisms the final evaluation report highlighted. Although shortcomings of information might be explained in terms of individual faults of those conducting self-reporting processes or ‘inherent’ weaknesses in departments or even the programs, we suspect it also underlines the sheer difficulty of self-reporting about what are rather novel, experimental organizational structures in medical research. Whereas in principle any interactive ancillary peer review format may be thwarted by inconsistencies or lack of thoroughness in organizational self-reporting, the fact that a number of departments were having to define their research mission in explicitly translational terms for the first time (or were in such early stages of development that such promissory outcomes or processes were hard to identify) arguably made this problem particularly acute in this research evaluation setting. For these reasons the self-reporting procedures may benefit to some extent from flexible use of standardized, under-determined a priori guidelines more specifically oriented towards detecting and evaluating process criteria (Hellström 2012; Molas-Gallart et al. 2015). One test of such emerging evaluation criteria would be to see whether they are sufficiently flexible and informative in helping to link self-reporting and committee deliberation phases in such a peer review format. Given the Standard Evaluation Protocol is designed to govern all types of research activity in Dutch universities, and hence only lays-out very general parameters, this would not be a suitable venue for instructing how academic medical centres should prepare information for committees on translational programs. For those coordinating and engaging in self-evaluations this may be a matter of informal learning, for whom we hope these recommendations are useful.

In the second part of our analysis, we reported issues concerning the selection of committees and questioned what constitutes expertise in evaluation translational research programs. The granting of significant discretion is arguably one of the benefits of the Dutch peer review format with regards producing information from which management and program leaders can learn and act upon. A potential risk is that the likely usefulness is highly sensitive to the selection of peers (Langfeldt 2006). Committees are accountable to a number of audiences, both proximal and distant, mediating between standards and norms of excellence they are committed towards upholding as professional scientists; the specific and local audience who is conducting the research being evaluated; and the (shifting) formal criteria the Dutch evaluation protocol requires them to follow in holding university research to account. For potential future committee members presiding as external experts on similar assessments, we suggest that default reliance on publication and citation-based measures of research excellence can be quite limited in the context of evaluating translational research programs, especially as the management board explicitly did not wish to learn about more traditional dimensions of research performance. Indeed an interesting and important puzzle raised by these moments concerns what constitute the more desirable

characteristics of expertise. Here a number of the features Lamont (2009) suggests constitute a good committee member (preparedness, politeness, open-mindedness etc.) surely still hold in this expanded context. For those in research organizations tasked with selecting expert committee members, the assumption that individuals should be selected on the basis of their disciplinary credentials might also have to be approached with some caution in the context of translational research programs. In our view, the notion set out in the Terms of Reference that committee members should be knowledgeable of research outside their own specialty should be a minimum. One can go further and state that an important quality is *flexibility towards local circumstances*. The variability of goals in translational research programs and resulting variability of quality criteria and indicators needed to evaluate them (Klein 2008) requires flexibility and sensitivity among peers. Committees need to respond quickly towards what the organization requests, rather than applying de-contextualized measures of quality. Put differently, expertise may be equated with knowing *when* certain performance indicators are appropriate and when not.

As yet individuals receive no formal training for how to evaluate these kinds of initiatives and the transition towards use of narrative methods may seem particularly abrupt for research communities like medicine, with little experience or tolerance for ‘soft’ methods. One strategy for policymakers would be to promote and organize further training opportunities and ensure opportunities to participate are made available in universities. Another would be to assume that over time individuals who sit on these committees will become more familiar with the kinds of process criteria needed to make evaluative judgments, and to assume that reviewers will become less perturbed by the ambiguity of the narrative-based method currently used by ancillary peer review and in assessing social relevance. Certainly our findings suggest some variation across committees in terms of continued commitment towards ‘traditional’ standards for evaluating translational research programs. Thus, comparing how different epistemic communities and their cultures of evaluation meet with the challenges of working with narrative-based methods and emerging ranges of alternative indicators to evaluate translation initiatives would be a fruitful avenue for further investigation. Although medicine may not be alone in facing challenges of moving towards evaluating cross-disciplinary research programs and their long-term societal outcomes, it is arguably the field in which pressures to make transitions away from disciplinary definitions of excellence towards more process-oriented evaluation criteria have been (and will continue to be) most pronounced, given the social values and accountability expected of the biomedical field.

In some instances, program leaders appeared to select committee members from their international scientific networks on the basis of them being renowned scientists from prestigious institutions in North America and Europe, with high number of citations to their name (this selection logic was also present in the Terms of Reference). Of course some may continue to desire being judged and receiving attention from those who embody the highest standards of excellence in scientific terms (Lamont 2009), and may even have a political stake in continuing to be evaluated through established standards of ‘excellence’.

But for members of research organizations interested in learning about organizational processes and translational ‘proximities’ (Molas-Gallart et al. 2015), appointing committee members with experience of building or managing strategic research programs, or with wider experience of engaging in third mission activities in the course of their research may also be beneficial. Guidance on committee selection might be set out in future iterations of the *Standard Evaluation Protocol*, modified in light of what an organization wants to achieve from an assessment (more summative judgment versus formative advice). These are only modest suggestions, as committee selection always remains something of a lottery, given organizers can never fully anticipate group dynamics or the attitudes or taste of individuals plucked from remote parts.

Finally the problems reported in this paper suggest that expanding an existing evaluation machinery to focus on translational research programs requires from managers and researchers involved in administering and appearing before committees, much more preparation, imagination, and risk than the more routine fundamental research-oriented evaluations that had typically taken place until this point in Dutch University Medical Centers. Fixing the issues we identified may be easier said than done, as most of our suggestions to improve the procedure would involve even more work from the board, administrators, and committee members, after what was already an expanded process. This dilemma relates to the burden of evaluations and whether these kinds of exercises are deemed worthwhile (Youtie and Corley 2011). Whilst we cannot answer these questions, we hope our contribution will lead to further scrutiny of the tensions that come with evaluating translational initiatives at an organizational level and to further reflections on what can be done to improve upon existing evaluation procedures.

## References

- Boggio, Andrea, Ballabeni, Andrea, and Hemenway, David (2016), 'Basic Research and Knowledge Production Modes: A View from the Harvard Medical School', *Science, Technology & Human Values*, 41 (2), 163-93.
- Bozeman, Barry (1993), 'Peer review and evaluation of R&D impacts', *Evaluating R&D impacts: Methods and practice* (Springer), 79-98.
- Bozeman, Barry and Boardman, Craig (2009), 'Broad impacts and narrow perspectives: Passing the buck on science and social impacts', *Social Epistemology*, 23 (3-4), 183-98.
- Cozzens, Susan E (1997), 'The knowledge pool: measurement challenges in evaluating fundamental research programs', *Evaluation and Program Planning*, 20 (1), 77-89.
- Cozzens, Susan E, Bobb, Kamau, and Bortagaray, Isabel (2002), 'Evaluating the distributional consequences of science and technology policies and programs', *Research Evaluation*, 11 (2), 101-07.
- Dahler-Larsen, Peter (2012), *The evaluation society* (Stanford, California: Stanford Business Books) x, 265 pages.
- Daniel, Hans-Dieter, Mittag, Sandra, and Bornmann, Lutz (2007), 'The potential and problems of peer evaluation in higher education and research', *Quality assessment for higher education in Europe*, 71-82.
- de Jong, Stefan P. L., Smit, Jorrit, and van Drooge, Leonie (2016), 'Scientists' response to societal impact policies: A policy paradox', *Science and Public Policy*, 43 (1), 102-14.



- Feller, Irwin (2013), 'Peer review and expert panels as techniques for evaluating the quality of academic research', in Albert N. Link and Nicholas S. Vonortas (eds.), *Handbook on the Theory and Practice of Program Evaluation* (Cheltenham: Edward Elgar), 115-42.
- Georghiou, Luke and Laredo, Philippe (2006), 'Evaluation of publicly funded research: recent trends and perspectives', *Report to the OECD DSTI/STP 7*.
- Hansson, F. (2010), 'Dialogue in or with the peer review? Evaluating research organizations in order to promote organizational learning', *Science and Public Policy*, 37 (4), 239-51.
- Hansson, F. and Monsted, M. (2012), 'Changing the Peer Review or Changing the Peers - Recent Development in Assessment of Large Research Collaborations', *Higher Education Policy*, 25 (3), 361-79.
- Hellström, Tomas (2012), 'Epistemic capacity in research environments: a framework for process evaluation', *Prometheus*, 30 (4), 395-409.
- Hemlin, Sven and Rasmussen, Søren Barlebo (2006), 'The Shift in Academic Quality Control', *Science, Technology & Human Values*, 31 (2), 173-98.
- Hicks, Diana (2012), 'Performance-based university research funding systems', *Research Policy*, 41 (2), 251-61.
- Howell, Embry M and Yemane, Alshadye (2006), 'An Assessment of Evaluation Designs: Case Studies of 12 Large Federal Evaluations', *American Journal of Evaluation*, 27 (2), 219-36.
- Klein, Julie T (2008), 'Evaluation of interdisciplinary and transdisciplinary research: a literature review', *American journal of preventive medicine*, 35 (2), S116-S23.
- KNAW (2009), *Standard Evaluation Protocol 2009-2015* (The Hague: KNAW, NWO, VSNU).
- Lamont, Michele (2009), *How professors think : inside the curious world of academic judgment* (Cambridge, Mass. ; London: Harvard University Press).
- Langfeldt, Liv (2004), 'Expert panels evaluating research: decision-making and sources of bias', *Research Evaluation*, 13 (1), 51-62.
- (2006), 'The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments', *Research Evaluation*, 15 (1), 31-41.
- Lawrenz, Frances, Thao, Mao, and Johnson, Kelli (2012), 'Expert panel reviews of research centers: The site visit process', *Evaluation and Program Planning*, 35 (3), 390-97.
- Macleod, Malcolm R., et al. (2014), 'Biomedical research: increasing value, reducing waste', *The Lancet*, 383 (9912), 101-04.
- Martin, B. R., Allwood, C., and Hemlin, Sven (2004), 'Conclusions: how to stimulate creative knowledge environments', in Sven Hemlin, C. Allwood, and B. R. Martin (eds.), *Creative Knowledge Environments: The Influences on Creativity in Research and Innovation* (Cheltenham: Edward Elgar), 193-220.
- Miles, Matthew B. and Huberman, A. M. (1994), *Qualitative data analysis : an expanded sourcebook* (2nd edn.; Thousand Oaks: Sage Publications) xiv, 338 p.
- Molas-Gallart, Jordi, et al. (2015), 'Towards an alternative framework for the evaluation of translational research initiatives', *Research Evaluation*, rvv027.
- NFU (2012), *NFU priorities in health research*, ed. ZonMW (Utrecht).
- Patton, Michael Quinn (2015), 'Evaluation in the Field: The Need for Site Visit Standards', *American Journal of Evaluation*, 36 (4), 444-60.
- Quinlan, Kathleen M, Kane, Mary, and Trochim, William MK (2008), 'Evaluation of large research initiatives: outcomes, challenges, and methodological considerations', *New Directions for Evaluation*, 2008 (118), 61-72.
- Rey-Rocha, Jesús and Martín-Sempere, María José (2012), 'Generating favourable contexts for translational research through the incorporation of basic researchers into hospitals: The FIS/Miguel Servet Research Contract Programme', *Science and Public Policy*, 39 (6), 787-801.
- Rons, Nadine, De Bruyn, Arlette, and Cornelis, Jan (2008), 'Research evaluation per discipline: a peer-review method and its outcomes', *Research Evaluation*, 17 (1), 45-57.
- Rushforth, Alexander and de Rijcke, Sarah (2015), 'Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands', *Minerva*, 53 (2), 117-39.

- Samuel, Gabrielle N. and Derrick, Gemma E. (2015), 'Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014', *Research Evaluation*, 24 (3), 229-41.
- Van Der Meulen, B. (2007), 'Interfering Governance and Emerging Centres of Control: University research evaluation in the Netherlands', in R. Whitley and J. Glaser (eds.), *The Changing Governance of the Sciences* (Netherlands: Springer).
- Van Drooge, Leonie, et al. (2013), *Twenty years of research evaluation* (The Hague: Rathenau Institute).
- Westerheijden, DonF (1997), 'A solid base for decisions', *Higher Education*, 33 (4), 397-413.
- Whitley, Richard (2007), 'Changing governance of the public sciences', in Richard Whitley and J. Glaser (eds.), *The changing governance of the sciences: The advent of research evaluation systems: Sociology of sciences yearbook* (Dordrecht: Springer), 3-30.
- Whitley, Richard and Gläser, Jochen (2014), *Organisational transformation and scientific change: The impact of institutional restructuring on universities and intellectual innovation* (42: Emerald Group Publishing).
- Youtie, Jan and Corley, Elizabeth A (2011), 'Federally sponsored multidisciplinary research centers: Learning, evaluation, and vicious circles', *Evaluation and program planning*, 34 (1), 13-20.