fteval JOURNAL

for Research and Technology Policy Evaluation

JULY 2025

#57

Proceedings of the REvaluation Conference 2024

RTI POLICY REvaluation '24 CONFERENCE

NAVIGATING TIMES OF CHANGE

FEDERAL MINISTRY FOR WOMEN, SCIENCE AND RESEARCH
Minoritenplatz 5, 1014 Vienna

Dr. Patrick Svensson-Jajko, E: Patrick.Svensson-Jajko@bmfwf.gv.at
Mag.ª Simone Mesner, E: simone.mesner@bmbwf.gv.at

FEDERAL MINISTRY FOR ECONOMY, ENERGY AND TOURISM
Stubenring 1, 1010 Vienna

Mag. Bernd Zimmer, bernd.zimmer@bmwet.gv.at

FEDERAL MINISTRY FOR INNOVATION, MOBILITY AND INFRASTRUCTURE
Radetzkystraße 2, 1030 Vienna

Dr. Rupert Pichler, E: rupert.pichler@bmimi.gv.at
Philipp Witibschlager, MSc, BA, E: philipp.witibschlager@bmimi.gv.at
Lisa Koch, MMA MSc, Kmsr., E: lisa.koch@bmimi.gv.at

ACR – AUSTRIAN COOPERATIVE RESEARCH
Sensengasse 3/3/3, 1010 Vienna

Dr.in Sonja Sheikh, E: sheikh@acr.ac.at

AIT – AUSTRIAN INSTITUTE OF TECHNOLOGY
Giefinggasse 4, 1210 Vienna

Mag. Michael Dinges, E: michael.dinges@ait.ac.at
Anahi Montalvo-Rojo, E: anahi.montalvo-rojo@ait.ac.at

AQ AUSTRIA – AGENCY FOR QUALITY ASSURANCE AND ACCREDITATION AUSTRIA
Franz-Klein-Gasse 5, 1190 Vienna

Dr.in Ulrike Najar, E: ulrike.najar@aq.ac.at

AWS – AUSTRIA WIRTSCHAFTSSERVICE GMBH
Rechte Wienzeile 225, 1120 Vienna

Mag. Norbert Knoll, E: N.Knoll@aws.at

CDG – CHRISTIAN DOPPLER RESEARCH ASSOCIATION
Boltzmanngasse 20, 1090 Vienna

DIin Magª. Brigitte Müller, E: brigitte.mueller@cdg.ac.at

CLIMATE AND ENERGY FUND
Leopold-Unger-Platz 2/1/142, 1190 Vienna

Mag. Bernd Vogl, E: bernd.vogl@klimafonds.gv.at
Dr. in Chiara Cardelli, MSc, E: chiara.cardelli@klimafonds.gv.at

CONVELOP – COOPERATIVE KNOWLEDGE DESIGN GMBH
Kaiserfeldgasse 7/II, 8010 Graz

DIin Dr.in Karin Grasenick, E: karin.grasenick@convelop.at
Konstantin Melidis, E: konstantin.melidis@convelop.at
Erdbergstraße 82/4, 1030 Wien

FFG – AUSTRIAN RESEARCH PROMOTION AGENCY
Haus der Forschung, Sensengasse 1, 1090 Vienna

Marlene Lang, MA, E: marlene.lang@ffg.at
Mag. Leonhard Jörg, E: leonhard.joerg@ffg.at
Dr. Jakob Eder, MA, E: jakob.eder@ffg.at

FH CAMPUS WIEN
Favoritenstraße 226, 1100 Wien

Dr. Mario Steyer, E: mario.steyer@fh-campuswien.ac.at

FORWIT - AUSTRIAN COUNCIL FOR SCIENCES, TECHNOLOGY, AND INNOVATION
Pestalozzigasse 4/DG 1, 1010 Vienna

Priv.-Doz.in DIin Mag.ª Dr.in techn. Alexandra Mazak-Huemer, E: amh@rfte.at
Dr. Anton Graschopf, E: ag@rfte.at

FWF – AUSTRIAN SCIENCE FUND
Sensengasse 1, 1090 Vienna

Dr.in Tina Olteanu, E: tina.olteanu@fwf.ac.at
Dr. Falk Reckling, E: falk.reckling@fwf.ac.at

GESELLSCHAFT FÜR FORSCHUNGSFÖRDERUNG NIEDERÖSTERREICH M.B.H.
Hypogasse 1/1, 3100 St. Pölten

Mag. Georg Pejrimovsky, E: g.pejrimovsky@gff-noe.at
Mag. Mario Enzenberger, E: m.enzenberger@gff-noe.at

IHS – INSTITUTE FOR ADVANCED STUDIES
Josefstädter Straße 39, 1080 Vienna

Dr.in Angela Wroblewski, E: wroblews@ihs.ac.at

INDUSTRIEWISSENSCHAFTLICHES INSTITUT – IWI
Mittersteig 10, 1050 Wien

FH-Hon.Prof. Dr. Dr. Herwig W. Schneider, E: schneider@iwi.ac.at
Mag. Philipp Brunner, E: brunner@iwi.ac.at

JOANNEUM RESEARCH FORSCHUNGSGESELLSCHAFT MBH
Leonhardstraße 59, 8010 Graz

Michael Ploder, E: Michael.Ploder@joanneum.at
Mag. Jürgen Streicher, E: juergen.streicher@joanneum.at

KMU FORSCHUNG AUSTRIA –AUSTRIAN INSTITUTE FOR SME RESEARCH
Gusshausstraße 8, 1040 Vienna

Mag. Peter Kaufmann, E: p.kaufmann@kmuforschung.ac.at
Harald Wieser, PhD, E: h.wieser@kmuforschung.ac.at
Jakob Kofler, MSc MPA, E: j.kofler@kmuforschung.ac.at

LUDWIG BOLTZMANN GESELLSCHAFT
Nußdorfer Straße 64/6, 1090 Vienna

Elvira Welzig, E: elvira.welzig@lbg.ac.at
Sherin Alias, E: Sherin.Alias@lbg.ac.at
Georg Russegger, E: georg.russegger@lbg.ac.at

OEAD - AGENCY FOR EDUCATION AND INTERNATIONALISATION
Ebendorferstraße 7, 1010 Vienna

Mag. Ulrich Hörmann, E: ulrich.hoermann@oead.at

OEAW – AUSTRIAN ACADEMY OF SCIENCE
Dr. Ignaz Seipel-Platz 2, 1010 Vienna

Ing. Mag. Michael Ofner, E: michael.ofner@oeaw.ac.at

TECHNOPOLIS GROUP AUSTRIA
Rudolfsplatz 12/11, 1010 Vienna

Mag.a Katharina Warta, E: warta@technopolis-group.com
Tobias Dudenbostel, MA, E: tobias.dudenbostel@technopolis-group.com

VIENNA BUSINESS AGENCY.
A SERVICE OFFERED BY THE CITY OF VIENNA.
Mariahilfer Straße 20, 1070 Vienna

DI Mag. Alfried Braumann, E: braumann@wirtschaftsagentur.at

WIFO – AUSTRIAN INSTITUTE OF ECONOMIC RESEARCH
Arsenal, Objekt 20, PF 91, 1030 Vienna

Dr. Jürgen Janger, E: juergen.janger@wifo.ac.at
Mag.a Dr.in Agnes Kügler, E: agnes.kuegler@wifo.ac.at

WPZ RESEARCH GMBH
Mariahilfer Straße 115/16, 1060 Vienna

Dr.in Brigitte Ecker, E: brigitte.ecker@wpz-research.com
Prof.(FH) Mag.a Dr.in Verena Régent, E: verena.regent@wpz-research.com

WWTF – VIENNA SCIENCE AND TECHNOLOGY FUND
Schlickgasse 3/12, 1090 Vienna

Dr. Michael Stampfer, E: michael.stampfer@wwtf.at
Dr. Michael Strassnig, E: michael.strassnig@wwtf.at

ZSI – CENTRE FOR SOCIAL INNOVATION
Linke Wienzeile 246, 1150 Vienna

Dr. Klaus Schuch, E: schuch@zsi.at
Stefanie Schuerz, MA, E: schuerz@zsi.at
Mag. Dr.in Elisabeth Unterfrauner, E: unterfrauner@zsi.at

EDITORIAL

DEAR READERS,

Evaluation is a field that rarely takes things at face value, which makes it even more important to meet face-to-face from time to time to exchange perspectives, clarify assumptions, and critically reflect together. The REvaluation Conference 2024, held in Vienna from December 4-6, offered such a moment. Against the backdrop of a shifting research and innovation policy landscape, the conference brought together researchers, policymakers, and practitioners to collectively explore what it means to evaluate in times of transformation. With its mix of conceptual reflection, methodological debate, and practical insight, the conference made one thing clear: evaluation is not merely reacting to change, it is becoming part of the change.

This issue of the *fteval Journal for Research and Technology Policy Evaluation* contains the proceedings of the REvaluation Conference 2024 and documents the outcomes of the discussions and exchanges that shaped the event. Structured around four thematic clusters, this issue reflects the strands of discussion that defined the REvaluation 2024 programme. The conference's agenda – from transformation-oriented policies to novel digital tools – is mirrored here and further developed through the selected contributions. Each block connects directly to sessions and debates held in Vienna, offering both, summaries and fresh perspectives.

The first cluster, ***Challenges and New Approaches in Research Assessment***, features contributions by Aura Kivilaakso and Johanna Kolhinen; Lottie Provost and Zenia Xenou; Hendrik Berghäuser; Ioanna Grypari, Sergio Di Virgilio, Haris Papageorgiou, Aris Fergadis and Dimitris Pappas; as well as Florian Bayer. The cluster reflects the ongoing shift from traditional, metric-driven evaluation towards more inclusive and context-sensitive approaches. The contributions examine how research assessment can better reflect societal relevance, interdisciplinarity, and diverse researcher profiles. Across disciplines and institutions, the papers explore both the conceptual foundations and the practical tensions of reform. What unites them is a shared effort to expand what "excellence" can mean and how assessment can support rather than constrain responsible, impactful research.

The second cluster, ***Learning from Programme Evaluations***, includes contributions by Verena Régent and Brigitte Ecker; Dagmar Simon; as well as Lisa Neusel and Simon Hirzel. The contributions in this cluster illustrate how programme evaluations are evolving into instruments of strategic and transformative learning. While each article focuses on a different national context and policy field - gender equality in Austria, climate and energy policy in Germany, and alternative funding logics in philanthropic research support - they share a commitment to using evaluation not merely for accountability, but as a tool for reflection and development.

The third cluster, ***New Methods and Approaches***, brings together contributions by Florence Benoit; Valentina Di Girolamo, Dario Diodato, Erik Canton and Julien Ravet; Kathleen Toepel; Christina Schuh, Daniel Schwertfeger and Sonja Fringes; Sarah Seus, Florian Wittmann and Nele Weiblen; Erika Hajdu, Giovanna Lima and Stefan De Jong, as well as Michael Dinges, Christiane Kerlen, Surya Knöbel and Kathleen Toepel. This cluster brings together novel conceptual, methodological, and practical perspectives that push the boundaries of how research and innovation (R&I) are evaluated. The contributions span diverse topics – from complexity-based metrics to gamified theory-of-change tools

– but converge in their ambition to better capture dynamic, non-linear, and system-level change. Several papers experiment with new forms of data and modelling, such as the use of complexity and relatedness indicators for smarter policy design or innovation biographies that reconstruct micro-level impact pathways. Others reflect the changing role of evaluators and the limitations of traditional methods in capturing processes like digital transformation, societal impact, or unintended effects. Whether through mixed-method systemic evaluations, qualitative case reconstructions, or participatory tools, the contributions advocate for more adaptive, reflexive, and creative approaches to evaluation. Collectively, they make a strong case for expanding the methodological repertoire to keep pace with the evolving goals of STI policy.

The fourth cluster, ***Evaluation Systems and Cultures***, features contributions by Peter Kolarz and Diogo Machado; Vitalii Gryga, Olha Krasovska and Yuliia Ryzhkova; Leonie van Drooge and Niki Vermeulen; as well as Klaus Schuch, Vitalii Gryga, Attila Havas, Renata Anna Jaksa, Olha Krasovska, Arzu Kurşun, Sofia Liarti, Eriona Shtëmbari, Miroslav Štefánik, Lena Tsipouri and Albena Vutsova. This cluster reflects the growing recognition that evaluation does not operate in a vacuum; institutional traditions, policy regimes, and the maturity of evaluation cultures shape it. The contributions explore how evaluation systems are being constructed, reformed, or contested in different settings, often revealing a tension between formalised procedures and the informal dynamics that sustain them. Across the board, a key theme emerges: building robust and responsive evaluation systems requires more than methodological sophistication - it demands attention to context, capacity, and culture.

Taken together, the four thematic clusters reflect the diversity and depth of current debates in research and innovation policy evaluation. They offer insights into how assessment practices are being reimagined, how programs are becoming sites of learning, how methodological innovation is broadening the evaluator's toolbox, and how systems and cultures shape what evaluation can achieve. What binds these contributions is a shared understanding that evaluation is not an external observer of change, but an active participant in shaping it. The REvaluation Conference 2024 provided the space to engage with these developments critically and collaboratively. This issue captures that spirit of reflection, experimentation, and shared purpose.

Enjoy reading, and we hope to keep the discussions going at the next volume of REvaluation Conferences in 2027!

Michael Dinges & Isabella Wagner

fteval JOURNAL
for Research and Technology Policy Evaluation

July 2025, Vol. 57, pp. e1, 1-16
DOI: 10.22163/fteval.2025.693

REFRAMING RESEARCH ASSESSMENT: TOWARDS A COMPREHENSIVE FRAMEWORK FOR RESEARCHER PROFILES

LOTTIE PROVOST AND ZENIA XENOU
DOI: 10.22163/FTEVAL.2025.693

ABSTRACT

The reform of research assessment is a top priority in the European Research Area. Recognising its crucial role in a strong Research and Innovation system, recent policies call for new approaches. Traditional methods rely heavily on publication metrics, failing to reflect the collaborative and interdisciplinary nature of modern research. The CoARA Agreement on Reforming Research Assessment, which was officially opened for signature on 28 September 2022 and counted 832 signatories as of 14th, March 2025, calls for better recognition of the diversity of research contributions, outputs, and career paths, and to base research assessment primarily on qualitative evaluation supported by a responsible use of quantitative indicators. The movement for reform also calls for better acknowledgement of contributions to Open Science.

This contribution presents a framework for "Researcher Profiles" under development within the Horizon Europe project GraspOS (Grant Agreement n.101095129). This service aims at supporting organisations in implementing the CoARA commitments and to offer a flexible framework for assessing researchers which values diverse practices, and prioritises comprehensive quality and societal impact of research.

Keywords: Responsible Research Assessment, Researcher Profile Framework, Open Science Infrastructure, Research Curricula

1. INTRODUCTION

The European Commission has placed the reform of the research assessment system at the top of the European Research Area Policy Agenda 2022-2024, emphasising that the way research projects, researchers, research units, and research institutions are assessed is fundamental for a well-functioning Research and Innovation system.

Policy efforts have sought to accelerate the shift away from the established, publication-based assessment methods, underlining their limitations in reflecting the increasingly collaborative and interdisciplinary nature of research (European Research Area policy agenda, 2022). Consolidated evidence shows that publication-based metrics such as the Journal Impact Factor[1] and the h-index[2] fail to reflect the broad range of activities that make up research, and are widely (mis)used as proxies for assessing the quality, performance and impact of research and researchers (Institut de France, 2011; Hicks et al., 2015; Pontika et al., 2022; DORA, 2024).

Critics have also drawn attention to how the current assessment system has fostered perverse incentives for researchers, encouraging them to prioritise aspects such as publication venue and number of citations (Edwards et al., 2017), often at the expense of essential aspects of scientific knowledge production such as research quality, collaborative open research methods, and the impact of research on society (Di Donato, 2024). These incentives can shape not only how research is conducted, but also which questions are pursued, steering scholars toward topics more likely to yield high-impact publications (Van Wesel, 2015).

In response to the identified challenges, the European Commission has driven the efforts seeking to establish a clear and common direction for the reform of research assessment practices.

In 2021, the European Commission Scoping Report "Towards a reform of the research assessment system" (European Commission, 2021) called for research proposals, researchers, research units and research institutions to be "evaluated on their intrinsic merits and performance rather than on the number of publications and where they are published, promoting qualitative

1 https://www.nihlibrary.nih.gov/about-us/faqs/what-are-journal-impact-factors

2 https://guides.lib.umich.edu/c.php?g=282982&p=1887449

judgement with peer-review, supported by responsible use of quantitative indicators." Echoing this call, signatories of the Agreement on Reforming Research Assessment (ARRA) (CoARA, 2022), who were 832 in total as of 14th, March 2025[3], have undertaken to uphold a series of commitments, including to recognise and value diverse contributions to and careers in research, and to base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators[4].

However, for most ARRA signatory organisations, implementing such changes remains a challenge. In particular, tailoring research assessment practices to different disciplines, career stages and research outputs further increases this challenge and the lack of a high-quality and open infrastructure[5] appears to be a major obstacle.

A number of EU-funded projects are tasked with supporting the ongoing policy reforms and designing new ways to incentivise higher quality research, collaboration and Open Science practices (European Commision, 2024). Among these, the Horizon Europe project GraspOS[6] addresses the need for new services and tools to support a research assessment system that incentivises Open Science practices. The project aims to develop a data infrastructure facilitating qualitative and quantitative assessments, ultimately supporting the practical implementation of the reform at various levels and the transition towards an Open Science-aware responsible research assessment.

3 https://coara.eu/agreement/signatories/

4 Commitments 1 and 2 of the ARRA: "Recognise the diversity of contributions to, and careers in, research in accordance with the needs and nature of the research" and "Base research assessment primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators". https://coara.eu/agreement/the-commitments/

5 Research infrastructures are facilities that provide resources and services for research communities to conduct research and foster innovation. They may be single-sited, distributed, or virtual and can include major scientific equipment or sets of tools and instruments; collections, archives or data; computing systems and communication networks; and any other research and innovation infrastructure open to external users. https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/european-research-infrastructures_en

6 Next Generation Research Assessment to Promote Open Science (Grant Agreement n.101095129) https://graspos.eu/

2. THE FRAMEWORK FOR RESEARCHER PROFILES: AN INNOVATIVE TOOL TO SUPPORT ORGANISATIONS IN ADOPTING RESPONSIBLE RESEARCH ASSESSMENT PRACTICES

2.1 NEW TOOLS SUPPORTING THE TRANSITION TO RESPONSIBLE RESEARCH ASSESSMENT PRACTICES

To support the emerging policy reforms and pave the way towards an Open Science-aware Responsible Research Assessment system, GraspOS is developing an innovative tool designed to support research funding and performing organisations in implementing the ARRA commitments. At the same time, it also enables researchers to provide a more comprehensive view of their contributions to science and society.

The tool is envisaged as a framework for Researcher Profiles, aligned with the latest policy guidance promoting a responsible approach to research assessment. In particular, its development is guided by the SCOPE Framework (International Network of Research Management Societies - Research Evaluation Group, 2023) and the DORA Guidance on the responsible use of quantitative indicators in research assessment (DORA, 2024).

The SCOPE Framework was developed by the International Network of Research Management Societies (INORMS) as a structured process to guide responsible research evaluation and to help research managers and evaluators in designing and implementing assessments which align with best practices and institutional values. The acronym SCOPE stands for the five stages of the process:

1. **Start with what you value:** Make sure that the evaluation process effectively measures and assesses what you or your institution value.

2. **Context considerations:** Ask yourself who are you evaluating? And why is the evaluation taking place? This should allow for more contextual evaluations.

3. **Options for evaluating:** Be careful about considering and balancing quantitative and qualitative measures and avoid using quantitative measures to evaluate qualities.

4. **Probe deeply:** Be aware of the unintended consequences that a certain evaluation approach may bear, such as unfair discrimination or eventual gaming strategies.

5. **Evaluate your evaluation:** In this last stage, reflect on the aims of the evaluation and assess whether these have been achieved.

The SCOPE Framework emphasises the need to prioritise core values and contextual factors in research assessments and calls for the recognition of diverse research contributions whose quality and impact cannot be assessed through quantitative metrics.

While metrics and indicators can serve as useful benchmarks for measuring research performance, they are inherently limited. Indeed, they often fail to capture the complexity and societal relevance of research. As outlined in the DORA Guidance on the responsible use of quantitative indicators in research assessment (DORA, 2024), a contextualised approach is essential—one that combines quantitative indicators with qualitative insights to reflect the broader impact and quality of research.

These principles underpin the development of the Researcher Profiles framework, which seeks to foster a research culture that values overall quality and societal impact over mere numerical output.

At the same time, it is important to recognise that elements of competition continue to shape the research environment, and efforts to promote more responsible, open, and fair assessment practices must take this reality into account. The Researcher Profiles framework does not attempt to eliminate competition, but to rebalance assessment criteria so that under-recognised qualities such as collaboration, societal impact, and openness are adequately valued and rewarded. In fact, qualitative insights provide critical context and help highlight such dimensions often overlooked in traditional evaluation systems.

Moreover, research funders operate under practical constraints. Limited time, administrative burdens, and the need to compare diverse applicants often compel them to rely on metrics that are easy to collect and compare. Recognising these pressures is essential to understand how responsible assessment reforms can be implemented within existing institutional and operational constraints.

By integrating both qualitative and quantitative elements, the Researcher Profiles framework offers a practical path forward which aligns with responsible assessment principles while remaining compatible with the operational needs of institutions and funders. In doing so, it supports more balanced and context-aware assessment processes that recognise a broader spectrum of scientific contributions.

2.2 DESIGNING THE FRAMEWORK FOR RESEARCHER PROFILES: METHODOLOGY

The design of the framework started with a landscape analysis of existing services and indicator frameworks used to describe research activity. This analysis confirmed an overreliance on publication-based metrics[7] and highlighted the need for a more comprehensive approach that includes a broader range of activities. In response, GraspOS aims to integrate both established and emerging indicators that better reflect diverse academic outputs and practices.

An essential foundation for this work is the integration of insights and resources from two Horizon Europe projects—OPUS and PathOS—whose characteristics are summarised in Table 1.

	GraspOS	OPUS	PathOS
Full title	next Generation Research Assessment to Promote Open Science	Open and Universal Science	Open Science Impact Pathways
Project website	https://graspos.eu/consortium-partners	https://opusproject.eu/	https://pathos-project.eu/
DOI	10.3030/101095129	10.3030/101058471	10.3030/101058728
Start date	1 January 2023	1 September 2022	1 September 2022
End date	31 December 2025	31 August 2025	31 August 2025
Funded under	Research infrastructures	Reforming and enhancing the European R&I System	Reforming and enhancing the European R&I System

7 Services providing indicators focusing mostly on scientific publications include Google Scholar, Academia.edu, Web of Science, and ResearcherID.

Topic	Services and tools to underpin a research assessment system that incentivises open science practices	Support to changes in the assessment of research and researchers to reward the practice of open science	Modelling and quantifying the impacts of open science practice
Consortium	18 partners	18 partners	10 partners
EU funding	€ 2 985 441,00	€ 1 726 898,00	€ 1 999 990,00

Table 1. Overview of Key Characteristics of Horizon Europe Projects: GraspOS, OPUS, and PathOS

Specifically, the OPUS Researcher Assessment Framework (RAF) (O'Neill, 2023) and the PathOS Open Science Indicator Handbook (Apartis et al., 2024) proved particularly useful, providing a solid basis for the design of the framework, as illustrated in Figure 1 below.

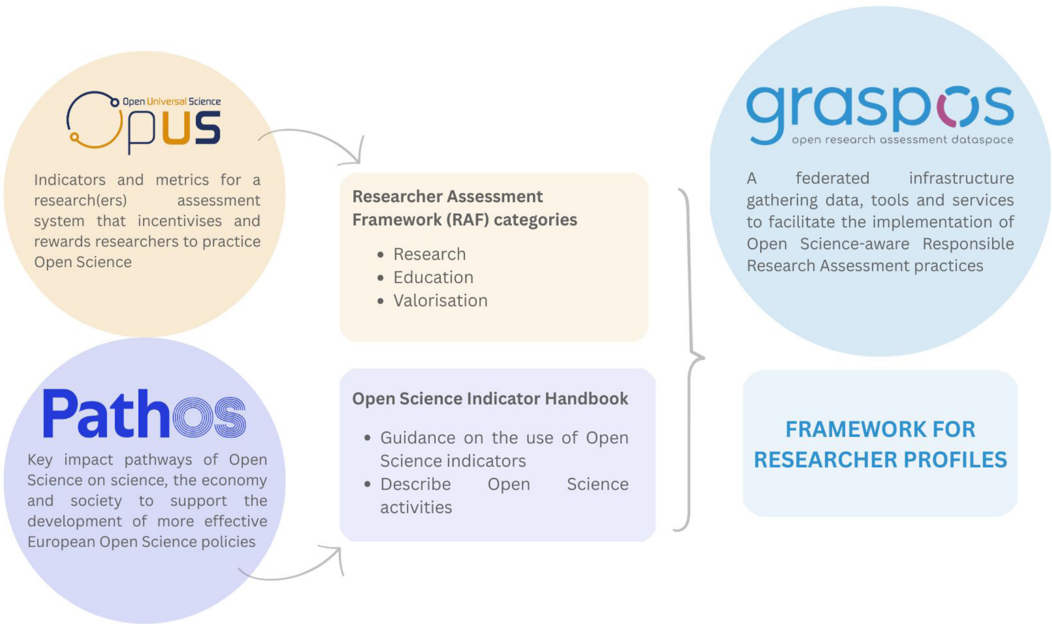


Figure 1. OPUS and PathOS project contributions to the GraspOS framework for Researcher Profiles

The OPUS project is working on a framework to assess researchers (RAF) including Open Science dimensions, to ensure that such practices are explicitly recognised and rewarded (O'Neill, 2023). From this framework, three main categories were identified – “Research”, “Education” and “Valorisation”– in which data collected through the landscape analysis was classified. These categories served as a structured framework to organise and interpret the diverse information gathered from various sources.

The PathOS project published a first version of the Open Science Indicator Handbook (Apartis et al., 2024), providing guidance on the use of a wide range of Open Science indicators. It served as a basis to describe Open Science activities in the framework for Researcher Profiles. Open Science is not limited to indicators relating to Open Access publishing, but is rather considered in a more holistic manner, encompassing and including key pillars as defined in the UNESCO Recommendation on Open Science (UNESCO, 2021).

Data on researchers’ contributions will be sourced from ORCID[8] and the OpenAIRE Graph[9]. These platforms provide reliable and comprehensive data on researchers and outputs. Moving forward, the aim is to integrate additional data sources to enhance the breadth and depth of the information collected. Additionally, functionalities allowing users to manually edit and update their data will be implemented, ensuring flexibility and accuracy in maintaining researcher profiles and related information.

Finally, the framework will undergo iterative refinement in collaboration with the nine GraspOS Pilots[10], each representing a specific context in research assessment (National research funding and performing organisations, Current Research Information Systems, universities and university departments, disciplines). The pilots will provide practical feedback on the suggested components of the framework.

8 https://orcid.org/

9 https://graph.openaire.eu/

10 https://graspos.eu/case-studies

3. KEY CHARACTERISTICS OF THE RESEARCHER PROFILE

3.1 NARRATIVE CV

A Narrative CV section will gather qualitative input on a researcher's skills and experiences. Following the four module model of the Royal Society's Résumé for Researchers[11], this approach supports a more contextual and qualitative assessment of their diverse contributions to science and society, including:

1. Contribution to the generation of knowledge
2. Contribution to the development of individuals
3. Contribution to the wider research community
4. Contribution to broader society

This key feature of the framework for Researcher Profiles will enable researchers to provide context and explain the impact of their research, and to highlight specific stories of particular interest in the context of an assessment event. Additional modules to present other types of experiences, such as extra-curricular or voluntary work will be included, thereby providing a more complete view of a researcher's profile. This Narrative CV section will serve as the core feature of this profile, providing a comprehensive overview of achievements and contributions, supported by evidence-based indicators.

3.2 INTERACTIVE TIMELINE

The interactive timeline is intended to provide a dynamic, visual representation of events or milestones which users can explore by clicking or hovering on different elements to reveal more information, allowing them to explore data and narratives in a more engaging way. The timeline complements the Narrative CV section by providing information about the evolution of a researcher's interests and research topics over time, highlighting the research and policy areas to which they have contributed in a chronological order.

11 https://royalsociety.org/news-resources/projects/research-culture/tools-for-support/resume-for-researchers/

3.3 DIVERSE RESEARCH OUTPUTS

Taking into account the need to recognise the variety of outputs produced in science, the Research Outputs section will gather a broad range of results including publications, preprints, datasets, software, patents, books, and other research-related products. For each output, a narrative box enables researchers to provide additional information on the context, the rationale, the activities carried, the outcomes, or any other relevant information.

3.4 OPEN SCIENCE

A section will be dedicated to recognising engagement with Open Science, rewarding researchers who contribute to making scientific knowledge openly available, accessible and reusable for all; to increasing scientific collaborations and sharing of information; and to opening the processes of scientific knowledge creation and communication to societal actors. The Researcher Profile aims to take into account three pillars of Open Science (UNESCO, 2021): open scientific knowledge, open science infrastructures, and open engagement of societal actors.

3.5 VALORISATION

Valorisation refers to the process of increasing the value and societal relevance of research by translating knowledge into practical applications, products, services, or broader social benefits. It includes both the refinement and dissemination of research findings and their contribution to solving real-world challenges.

In the Researcher Profiles framework, valorisation includes not only the immediate, applied impact of research but also the foundational contributions of basic, curiosity-driven research, which may not yield direct applications in the short term but are essential for long-term knowledge advancement. Researchers are encouraged to document and reflect on these contributions, ensuring their value is acknowledged in assessment processes.

A dedicated Valorisation section complements the Narrative CV, focusing on the wider impact of research beyond academia. This includes activities such as industry collaboration, policy engagement, public communication, and contributions to sustainable development. The framework also allows for the valorisation of conceptual and theoretical advances, whose importance may lie in shaping future directions of scientific inquiry.

By combining qualitative narratives with supporting quantitative indicators, the framework enables a context-sensitive and evidence-informed approach to assessment. In doing so, it supports a more responsible and balanced evaluation of research, where quality, originality, and societal relevance are valued alongside traditional outputs.

4. CONSIDERATIONS FOR FUTURE DEVELOPMENTS AND IMPLEMENTATION

The main aim of the GraspOS project is to develop tools and services to support and facilitate the transition towards Open Science-aware responsible research assessment practices. In light of the movement for reform and the growing emphasis on recognising and valuing a wider range of contributions to science and society, including Open Science practices, the framework under development promotes a balanced approach in the use of quantitative indicators and qualitative perspectives. However, as with any new tool, the design and development of the framework for Researcher Profiles should carefully take into account a variety of potential pitfalls.

IMPLEMENTATION CHALLENGES

A key challenge lies in the capacity of RPOs to adopt new standards and tools. While offering more depth and nuance in the research evaluation process, the effective shift to more qualitative and narrative-based assessment models requires more than just adopting new tools: it demands time, resources, and training.

BALANCING COMPARABILITY AND FLEXIBILITY

While uniformity across Researcher Profiles is not a desirable outcome, we are aware that research evaluators will need to compare profiles in a meaningful way. This is why the framework is structured according to standardised but adaptable categories that apply across disciplines, career stages and types of research—including basic research. These categories offer a common structure without enforcing a one-size-fits-all model. This flexibility ensures that assessments can be contextualized and sensitive to the diverse nature of scientific research.

OVER-RELIANCE ON KEY PERFORMANCE INDICATORS (KPIS)

Another challenge associated with the development of new research assessment models is the risk of replacing old metrics with new, but equally narrow, sets of KPIs. This could unintentionally reproduce the same limitations found in traditional assessment processes. Relying on a narrow set of indicators would undermine the very objectives of the framework, which are to recognise and value the wide range of contributions that researchers make to science and society. To counter this, the framework for Researcher profiles puts strong emphasis on qualitative evaluation, and researchers are encouraged to include narrative descriptions and context to explain their work and its impact.

NARROWING DOWN TO OPEN SCIENCE METRICS

Similarly, there may be a risk that specific quantifiable Open Science practices or outputs substitute previous misused metrics, missing the overall need to monitor the comprehensive transformation towards a new research culture12. In addition, there is a need to assess the values and impacts of science by focusing on the people who conduct, engage with, and benefit from it, while also addressing the current lack of relevant policies and training. Existing methods to assess the adoption of Open Science practices should therefore be strengthened (UNESCO, 2023), particularly to track changes in research culture and to value open and reproducible research processes.

SUPPORTING DIVERSE TYPES OF RESEARCHERS

Ensuring that researchers from diverse backgrounds, with different skills and competencies, are evaluated equitably is a critical consideration in the transition towards a more inclusive research assessment system. International researchers may struggle with integration into local networks and collaborations, and introverted researchers may face challenges in systems that prioritise public visibility and engagement. The framework addresses these concerns by allowing researchers to provide contextual information in their profiles. International researchers can highlight cross-border collaborations and explain challenges unique to their research environments. Introverted researchers can showcase less visible, yet essential, contributions, such as technical innovations, infrastructure development, or mentorship, through narrative descriptions.

12 The Royal Society defines research culture as follows: "Research culture encompasses the behaviours, values, expectations, attitudes and norms of our research communities. It influences researchers' career paths and determines the way that research is conducted and communicated." https://royalsociety.org/news-resources/projects/research-culture/

The development of the framework addresses several important considerations related to the flexibility of researcher profiles across diverse fields of study. One of the key challenges is ensuring that the tool can adapt to different contexts and needs across disciplines. Research contributions in fields such as humanities, social sciences, natural sciences, and applied sciences are fundamentally different in the way they are produced, disseminated, and evaluated. In fact, the framework for researcher profiles needs to be sufficiently flexible enough to be adapted to various local contexts and to cater to research institutions' diverse values, needs and goals. Ultimately, the aim is to design Researcher Profiles that are customisable and context-aware, allowing researchers to highlight achievements that are most relevant to their work.

5. CONCLUSIONS

The development of the framework for Researcher Profiles was presented for the first time at the REvaluation 2024 Conference[13] on 4th, December 2024 as a means to engage with and gather input from research assessment experts and community members interested in advancing research evaluation systems[14].

Responsible Research Assessment represents a critical evolution in how researchers' contributions are evaluated, emphasising fairness, transparency, and inclusivity across disciplines. The GraspOS project aims to address these needs by designing a dynamic Researcher Profile framework that combines qualitative narratives with a responsible use of quantitative indicators, promoting a comprehensive and contextualised view of contributions to research.

The inclusion of the interactive timeline feature and the Narrative CV section is aimed at further enhancing and enabling qualitative evaluations. Paired with evidence-based information, this qualitative perspective should allow for a more comprehensive assessment of research quality and productivity and should leave enough flexibility for research institutions or researchers to adapt the tool to their needs.

As research assessment evolves, it is crucial to avoid simply replacing traditional metrics with new indicators that risk being misapplied. The focus

13 https://www.revaluation2024.eu/

14 The presentation is available in Open Access on Zenodo. https://doi.org/10.5281/zenodo.14531056

should remain on fostering a responsible evaluation culture that values both the process and the broader impact of scientific work. By collaborating with the GraspOS Pilots and continuously refining the framework based on practical feedback, the project aims to ensure its relevance and adaptability across various contexts.

REFERENCES

Apartis, S., G. Catalano, G. Consiglio, R. Costas, E. Delugas, M. Dulong de Rosnay, I. Grypari, et al. 2024. "Open Science Impact Indicator Handbook." Zenodo. 2024. https://doi.org/10.5281/zenodo.14538442.

Coalition for Advancing Research Assessment, Agreement on Reforming Research Assessment, 2022. https://coara.eu/agreement/the-agreement-full-text

Declaration on Research Assessment (DORA), Guidance on the responsible use of quantitative indicators in research assessment, 2024. https://doi.org/10.5281/zenodo.13844873

Di Donato F., "What we talk about when we talk about research quality. A discussion on responsible research assessment and Open Science", Bollettino telematico di filosofia politica, March 2024. https://commentbfp.sp.unipi.it/quality-fdd/; https://doi.org/10.5281/zenodo.10890788

Edwards, M. A., & Roy, S., Academic Research in the 21st Century : Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. Environmental Engineering Science, 34(1), 51-61, 2017. https://doi.org/10.1089/ees.2016.0223

European Commission, Directorate-General for Research and Innovation, Towards a reform of the research assessment system – Scoping report, Publications Office, 2021. https://data.europa.eu/doi/10.2777/707440

European Commission, Directorate-General for Research and Innovation, European Research Area policy agenda – Overview of actions for the period 2022-2024, Publications Office of the European Union, 2022. https://data.europa.eu/doi/10.2777/52110

European Commission, Action Plan by the Commission to implement the ten commitments of the Agreement on Reforming Research Assessment (ARRA), 2024. https://research-and-innovation.ec.europa.eu/document/download/e69aff11-4494-4e5f-866c-694539a3ea26_en?filename=ec_rtd_commitments-reform-research-assessment.pdf

Hicks, D., Wouters, P., Waltman, L. et al., Bibliometrics: The Leiden Manifesto for research metrics. Nature 520, 429–431, 2015. https://doi.org/10.1038/520429a

Institut de France, Académie des Sciences, "On the proper Use of Bibliometrics to Evaluate Individual Researchers", Report presented to the French Minister for Higher Education and Research, 17 January 2011. https://www.academie-sciences.fr/archivage_site/activite/rapport/avis170111gb.pdf

International Network of Research Management Societies - Research Evaluation Group, The SCOPE Framework, The University of Melbourne, Report, 2023. https://doi.org/10.26188/21919527.v1

O'Neill, G., OPUS Deliverable 3.1: Indicators and Metrics to Test in the Pilots. Zenodo, 2024. https://doi.org/10.5281/zenodo.10497434

Pontika, N., Klebel,T., Correia, A., Metzler, H., Knoth, P., Ross-Hellauer, T., Indicators of research quality, quantity, openness, and responsibility in institutional review, promotion, and tenure policies across seven countries. Quantitative Science Studies; 3 (4): 888–911, 2022. https://doi.org/10.1162/qss_a_00224

UNESCO, Recommendation on Open Science, 2021. https://doi.org/10.54677/MNMH8546

UNESCO, Open science outlook 1: Status and trends around the world—UNESCO Digital Library, 2023. https://doi.org/10.54677/GIIC6829

van Wesel, M., Evaluation by Citation: Trends in Publication Behavior, Evaluation Criteria, and the Strive for High Impact Publications. Sci Eng Ethics 22, 199–225 2016. https://doi.org/10.1007/s11948-015-9638-0

AUTHORS

LOTTIE PROVOST
Consiglio Nazionale delle Ricerche
Via Giuseppe Moruzzi, 56124 Pisa, Italy.
Email: lottiemiaprovost@cnr.it
ORCID: 0000-0001-5279-797X

ZENIA XENOU
OpenAIRE AMKE
Artemidos 6 & Epidavrou 15125 Maroussi, Athens, Greece
Email: zenia.xenou@openaire.eu
ORCID: 0009-0004-0561-5343

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e2, 1-17
DOI: 10.22163/fteval.2025.694

RESEARCH ASSESSMENT AS A TOOL FOR STRATEGIC MANAGEMENT

PROMOTING INTERDISCIPLINARITY THROUGH ENHANCEMENT-LED EVALUATION AT THE UNIVERSITY OF HELSINKI

AURA KIVILAAKSO AND JOHANNA KOLHINEN
DOI: 10.22163/FTEVAL.2025.694

ABSTRACT

The paper examines the evolution of research assessment as a tool for strategic development and management. At the University of Helsinki, research assessment framework's enhancement-led approach emphasizes continuous development and quality enhancement, supporting the implementation and planning of the organisation's strategy. In 2025, the statutory research assessment is harnessed as a tool for promoting multi- and interdisciplinary research by rethinking what and how to evaluate. Integrating interdisciplinarity into comprehensive research assessment exercise is an ambitious attempt in the field of research assessment. To succeed, it requires careful and regenerative planning.

Keywords: Research assessment, enhancement-led evaluation, interdisciplinarity, strategic management

INTRODUCTION

TOWARDS ENHANCEMENT-LED APPROACH OF RESEARCH ASSESSMENT

The University of Helsinki has a long tradition of evaluating its research activities. The latest assessment is underway in 2025, with previous assessments conducted in 2018–2019, 2012, 2005, and 1999. In 2018–2019, the assessment framework underwent significant revisions as the University renewed it with an enhancement-led approach to better reflect the evolving landscape of academic research and its societal impact. Enhancement-led evaluation emphasizes continuous improvement and future-oriented goals. It clarifies current conditions and trends, supporting ongoing development and a forward-looking mindset. This inclusive, interactive approach builds trust, encourages participation, and promotes collaboration among all stakeholders. A key feature is the flexible use of diverse, suitable and adaptable methods. By treating evaluation as a learning process, it supports change, offers feedback on strengths and areas for improvement, provides actionable recommendations, and enables progress tracking over time.[1]

At the University of Helsinki, key driver for the research assessment framework development in 2018–19 was to support the University's 2021–2030 strategy implementation.[2] At the same time, the comprehensive assessment was to fulfil the national legal requirements and cover all research, carried out at the University. In Finland, the Universities Act requires universities to assess their own research and its effectiveness frequently and publish the results openly.[3] However, there is no common framework specifically focused on research quality, whereas the Quality Audit system for Higher Education Institutions in Finland, conducted by FINEEC, covers education, research, and outreach as general elements of institutional quality systems.[4] Therefore, universities can freely define how they assess their research and what they aim to achieve with the assessment.

1 Moitus & Kamppi 2020, 6.

2 University of Helsinki 2024a.

3 Universities Act 558/2009.

4 Moitus & Kamppi 2020.

In 2025, the University of Helsinki conducts the overall assessment of its research again. The positive experience, gained from the previous Research Assessment at the University of Helsinki (RAUH 2018–19), along with subsequent strategic developments, enables the RAUH 2025 assessment to build on some of the same elements. However, the University decided to renew the assessment process from the perspective of the University's management and strategic thinking. In 2025, the statutory research assessment is harnessed as a tool for strategic management to promote interdisciplinary research. The goals of RAUH 2025 include highlighting strong research areas, identifying emerging fields, and evaluating the University's success in addressing societal challenges through research, as outlined in the University's strategy 2021–2030. It focuses on the management and state of multi- and interdisciplinary research at the University.[5] In addition, the results of the 2025 assessment will be used in preparation of the strategic period starting in 2031.

This article examines the development and evolution of the research assessment framework at the University of Helsinki from the perspective of strategic management. How can a research assessment support strategic renewal and foster interdisciplinarity at both university and unit levels? How can the academic community be engaged in the assessment project and commit the planned changes to the assessment framework? The article first approaches the University of Helsinki as an organization with a strategic goal to promote interdisciplinary research and then focuses on the assessment framework methodology at the University of Helsinki. It becomes clear that the units being assessed, the assessment criteria and the panel structure cannot be treated as separate parts of the assessment project. Instead, the assessment framework is an entity, where each component influences the others.

UNIVERSITY OF HELSINKI AND ITS STRATEGIC GOAL TO INVEST IN INTERDISCIPLINARY RESEARCH

The University of Helsinki is the oldest and largest institution of academic education in Finland, an international scientific community of 40,000 students and researchers. It operates on four campuses in Helsinki and in 10 other localities in Finland. The University accommodates 11 faculties, four independent research-focused institutes, as well as multidisciplinary research networks and

5 University of Helsinki 2024b.

campus units.[6] The University of Helsinki is also a proud founding member of the League of European Research Universities (LERU)[7]. Through the power of science, the University has contributed to society, education, and welfare since 1640.[8]

The University of Helsinki is a distinguished research university, renowned for its extensive range of scientific disciplines. For example, the City Centre Campus hosts five faculties dedicated to the social sciences and humanities (SSH), while the Viikki Campus, with its focus on life sciences, is home to Finland's only Faculty of Veterinary Medicine. The Kumpula campus serves a hub for the exact sciences, encompassing disciplines such as physics, chemistry, mathematics, and computer science, while the Meilahti Campus ranks among Europe's leading medical campuses, combining top-tier research, education, and patient care.[9] This diversity underscores the University's commitment to fostering a multidisciplinary academic environment, making it an attractive destination for researchers and students across various fields.

The broad range of disciplines at the University of Helsinki enables it to thrive in both discipline-specific and interdisciplinary research. Solutions to complex questions require a creative and integrative approach that brings together diverse actors, perspectives, and structures. Recognizing this, the four international assessment panels of the RAUH 2018–19 emphasized the importance of nurturing curiosity-driven and interdisciplinary research, core characteristics of a research-intensive university, while also promoting inclusive research environments and coherent institutional practices.[10] In response to the panels' recommendations, the University of Helsinki has taken strategic steps to strengthen interdisciplinarity, across its research and academic initiatives. As outlined in the strategic plan of the University of Helsinki 2021–2030, "*In 2030, the University will enjoy an increasingly established international standing as a scientific partner, especially thanks to its ground-breaking discipline-specific expertise as well as its multidisciplinary and cross-disciplinary research.*"[11]

6 University of Helsinki 2024c.

7 LERU 2025.

8 University of Helsinki 2024d.

9 University of Helsinki 2024e.

10 Mälkki & al. 2019, 6.

11 University of Helsinki 2024b.

In this article, the term “interdisciplinarity” is used as an umbrella term to describe research that is multidisciplinary, interdisciplinary, or transdisciplinary. The difference between these three research approaches relates to the degree of disciplinary boundaries in collaboration. In multidisciplinary research, actors from different disciplines work together, each bringing their domain expertise into the collaboration, whereas in clearly interdisciplinary research, interaction is multi-level and open to achieve collaboration that utilizes a deep, integrative approach. In interdisciplinary research, collaboration is seamless and participants’ understanding of their partners’ disciplines increases. Transdisciplinary research, on the other hand, almost completely merges the boundaries between disciplines to create a new, problem-oriented approach.[12] Interdisciplinarity not only enables studying complex challenges but also promotes the usability and value creation of research results for the world; maximizing the social and economic value of research and innovation, also known as knowledge valorisation.[13]

Scientific research is always built upon previous scientific studies and is influenced by strong discipline-based research paradigms. Interdisciplinary research is thus also built upon the foundation of previous scientific studies, with strong paradigms, rooted in disciplines guiding the research process. What makes interdisciplinary research special is its integrative nature that builds on strong disciplinary expertise – the purpose is to facilitate the encounter of different perspectives to generate so-called ‘landscapes of knowledge’.[14] The University of Helsinki, aiming to foster novel insights and breakthroughs, began investing in promoting interdisciplinarity at the start of its 2021 strategic period. This commitment is reflected through new internal funding instruments and various strategic development initiatives, including the RAUH 2025 research assessment exercise discussed in this article. While it acknowledges the value of strong disciplinary foundations, it promotes the integration of diverse perspectives to drive scientific innovation. Through the interdisciplinary cooperation, the University aims to reach novel research directions that seek solutions to major global problems.

12 Willamo & al. 2022, 12–14, 36; Wernli & Ohlmeyer 2023, 5.

13 European Commission 2021, 10.

14 Huutoniemi & al. 2010, 80.

THE EVOLUTION OF THE ASSESSMENT FRAMEWORK

The RAUH has evolved between the 2018–2019 and 2025 assessments, reflecting the changing priorities of the University of Helsinki's quality management policy, strategic choices, and methodologies for evaluating research quality and impact (Table 1). The RAUH 2025 project started with the objective to streamline the assessment processes and to construct a tool that promotes interdisciplinary research.

	RAUH 2018-19	RAUH 2025
Units of assessment	**39 units:** Faculties, Institutes, Departments, disciplines or combinations of disciplines	**15 units:** Faculties and research-focused independent institutes
Assessment criteria	▪ Scientific Quality ▪ Societal Impact ▪ Research Environment **and Unit Viability**	▪ Scientific Quality ▪ Societal Impact ▪ Research Environment
Evidence base	1. Metric data 2. Self-assessment 3. Panel Meeting	1. Metric data 2. Unit-level self-assessment reports **3. University-level self-assessment on interdisciplinarity** 4. Panel Meeting
Panels	**4 panels** **46 international experts**	**1 panel with 3 sub-groups** **27 international experts**

Table 1. Assessment frameworks in RAUH 2018–19 and RAUH 2025 projects.

Unlike, in 2018–2019, the timeline for the assessment was set shorter than 1.5 years, from the planning of the assessment framework update to the completion of the panel report. Also, it was predetermined that the number of the units of assessment should be lower to condense the assessment effort. The well-regarded assessment framework from 2018–2019 was to be retained as much as possible, but it was clear that changes were necessary due to the shift to larger units of assessment and the implementation of the strategic goal of promoting interdisciplinarity. Previously, the classification of the 39 units of assessment varied considerably, from large faculties and departments to individual disciplines or even combinations of disciplines. Now, the aim was to make the units of assessment more comparable, even though the sizes of faculties vary significantly within the University. RAUH 2025 streamlined the units of assessment to 15, focusing on faculties and research-focused independent institutes. At the same time, it was important to maintain a dialogical and strongly participatory approach with the University's academic community.

To allow academic units to monitor the success and progress of their development after the RAUH 2018–19 and to keep the continuous development cycle as productive as possible, the RAUH 2025 Steering Group decided that the three main assessment criteria of the RAUH 2018–19 assessment – Scientific Quality, Societal Impact, and Research Environment – will also be applied in RAUH 2025. However, changes were made in almost all other areas of the assessment framework, starting with the definition the units of assessment.

In the RAUH 2018–19 assessment, the evidence base was comprehensive, incorporating metric data, self-assessments, and panel meetings including site-visit and interviews. Self-assessment was seen primarily as a tool for improving operations. Therefore, the units were instructed to reflect upon the research and the research environment in a nuanced way to have a truly useful basis for further development. The assessment was conducted by four external panels comprising 46 international experts, utilizing an enhancement-led evaluation approach. The panels were asked to focus on the units' preparedness to address potential shortcomings by describing actions already taken or planned, rather than concentrating on the deficiencies themselves. Additionally, the panels valued the units' capacity for critical self-reflection, as RAUH 2018–19 used a rating scale of Weak – Good – Very Good – Excellent to categorize the performance of the assessed units.[15]

In the RAUH 2025 assessment, the evidence base was expanded to include a new university-level self-assessment on interdisciplinarity, while unit-level self-assessment reports and metric data on funding, personnel, publications, other outputs, doctoral research, and panel site-visit had already been used in the RAUH 2018–2019 assessment. The assessment panel was consolidated into one panel with three sub-groups, involving 27 international experts. The methodology continued to follow the enhancement-led evaluation. Notably, the steering group decided that RAUH 2025 will not use a rating scale; instead, the assessment was defined as focusing solely on identifying strengths and areas for improvement at both unit and university levels.

15 Mälkki & al. 2019.

The reduction in the number of units assessed and the inclusion of a university-level self-assessment on interdisciplinarity suggest a strategic move to foster collaboration and integration across different research areas at the University of Helsinki. Additionally, the consistent use of international experts and the enhancement-led evaluation method underscore the University's commitment to maintaining high standards and continuous improvement on its research activities.

In the 2018–2019 assessment, the self-assessment report was structured according to the predesigned headings, but the units were encouraged to freely decide on the use of any sub-headings. The units were expected to carry out as reflective self-assessments as possible, to identify areas in need of development, and to provide concrete descriptions of the operations and results. The first part of the self-assessment report focused on background information. The core of the self-assessment was the second part: the description of the organisation, profile, mission, and goals of the unit. The unit's performance and operations were primarily assessed against those measures. The self-assessment included reflection on the strengths and weaknesses of the described actions. In section three, the units described the self-assessment process.

In the RAUH 2025 assessment, the structure of the unit-level self-assessment report largely followed the format used in 2018. However, due to the larger size of the units being assessed, especially the guiding questions for self-evaluating scientific quality needed an update. From the perspective of these larger units, the previous criteria focused too much on project-level details, with monodisciplinary emphasis. Therefore, for example, the assessment of scientific quality now included more specific criteria such as research profile and scientific impact, which are evaluated based on both outputs and the unit's own qualitative self-assessment report. The assessment units are also expected to outline their future vision and provide more detailed goals for their research.

The RAUH 2025 project highlighted that the units of assessment, criteria, and panel structure are closely interconnected and must be addressed as an integrated whole. Instead, the assessment framework is an entity where each component influences the others. For example, when the decision was made to move away from traditional gradings, it became necessary to establish new, more detailed assessment criteria that aligned with the self-assessment report and served to guide the panel's qualitative, narrative-based evaluations. Defining qualitative assessment criteria turned out to be more challenging than initially expected, particularly in terms of ensuring consistency and transparency. Drafting a comprehensive and responsible assessment framework, required deep and profound discussions between the University's Research Assessment Office and with both the RAUH 2025 steering group and representatives of the units of assessment. In these discussions, clarity of the assessment objectives and an understanding of the differences between the units of assessment were key elements. Updating the assessment framework and considering every detail of the guidelines, templates and criteria has proven to be a complex exercise. However, this process has been a valuable learning experience for everyone involved.

ENHANCEMENT-LED EVALUATION AS AN ASSESSMENT METHOD – PARTICIPATION AS A CORE VALUE

Since the 2018–19 assessment, the RAUH framework has evolved from a one-time evaluation exercise into a dynamic and integrated tool for strategic development, fostering a culture of continuous improvement and alignment with the University's long-term goals. RAUH 2018–19 offered a solid basis for the mid and long-term development of research activities within the University's academic units, and the assessment results have been actively used since the assessment report's publication. Each academic unit has utilized the development areas identified by the assessment panel in their annual implementation plans, following the strategy of the University. The RAUH follow-up 2022–23 worked as a mid-term evaluation checkpoint, focusing on the development steps taken by the faculties and independent institutes. The follow-up phase included 15 faculty-level qualitative self-assessment reports, which addressed the main insights from RAUH 2018–19, the principal development measures undertaken based on the results, and the extent to which these actions align with the University's current strategic plan. In this context, it was observed that awareness had shifted positively, with representatives of academic units recognizing the value of RAUH as a tool for strategic development and expressing a desire for continuity between the frameworks of the RAUH 2018–19 and RAUH 2025 assessments.

In addition to the RAUH research assessment exercise, a more ambitious measure was taken in 2021–2022, when the University and its academic units elaborated 'Roadmaps for Implementing Research Themes'. These Roadmaps contributed to the goal of fostering interdisciplinary research. The roadmap process was part of the implementation of the University of Helsinki's strategy, but it also provided input for the RAUH follow-up in 2022–23. It became clear that the units benefited from the work they had done to develop their Roadmaps in connection with the mid-term RAUH self-assessment. To make the most of the constructive dialogue with the external panel, the self-assessments were submitted to the RAUH 2018–19 panel chairs, who provided feedback on the progress made. A joint discussion for the unit leadership, research managers and administrators, and other interested services and members of the university community was organised on 29 March 2023. Linking the two major university-wide strategic research management development tools – the RAUH and the Roadmap processes – was a natural outcome of building on synergies and the momentum of organisational learning.

In accordance with enhancement-led evaluation, self-assessment has been primarily a tool for improving operations in both RAUH 2018 –19 and RAUH 2025 assessments. The purpose of the RAUH assessment is to strive for meaningful impact and operate in a manner that is both positive and respectful towards all participants. The design and implementation of the assessment processes are intentionally crafted to foster continuous learning, facilitate the sharing of valuable information, and build a stronger, more cohesive understanding among all stakeholders. The true impact of all efforts is realized through the learning process itself. This process leverages evaluation data and expertise to enhance understanding, inform decision-making, and develop effective solutions. These outcomes drive development initiatives at both the University and unit levels, ensuring that the University's efforts are both comprehensive and impactful.

In practical terms, the enhancement-led research assessment method used in RAUH encourages the units of assessment to critically assess their own activities. This self-assessment helps identify strengths and areas for improvement in relation to the specific goals of the units of assessment. When necessary, this process is supplemented by external assessment feedback, providing an additional layer of insight and validation. Throughout the process, the assessment method emphasizes the importance of building and maintaining trust. The foundation for creating and sustaining meaningful, productive relationships both within the University and with external partners, ensuring that all participants feel valued and respected.[16]

The University of Helsinki has concluded that assigning grades in research assessment does not best support the intended trust-based approach. Instead, it is better to focus on qualitative, descriptive feedback, which genuinely aids the development of activities. This mindset aligns well with the spirit of the

16 Moitus & Kamppi 2020.

Coalition for Advancing Research Assessment (CoARA)[17]. Updating the RAUH assessment framework, in connection with the 2025 assessment, was therefore one of the actions in the University of Helsinki's CoARA Action Plan[18].

In the realm of research assessment, the emphasis on participation cannot be overstated. The RAUH process exemplifies this commitment, embedding participation as a core value in its enhancement-led methodology. This approach recognizes that the subjects of an assessment are not mere data points but are integral to the interpretation and success of the assessment. By involving the academic community of the University of Helsinki in the co-design and co-interpretation of the assessments, RAUH ensures that the assessment is conducted with researchers and not just about their work, fostering a deeper understanding and more meaningful results. The participatory method aligns with the SCOPE principle of evaluating with the evaluated, ensuring that evaluations are not only thorough but also respectful and inclusive, enhancing the relevance and impact of the research outcomes.[19]

In RAUH 2025, the University's Research Council – comprising vice-deans and directors of academic units – has taken a central collaborative role in shaping a shared, university-level self-assessment with a focus on interdisciplinary research. This represents a new form of participatory collaboration. The same academic leaders, who are responsible for coordinating the self-assessment processes within their units, have come together as a multidisciplinary body to influence the strategic development of the University. As there was no pre-existing model for such a university-wide self-assessment at the University of Helsinki, the process was approached with a strong orientation toward learning. The outcome was successful: the self-assessment phase not only deepened institutional understanding of the state of interdisciplinary research at the University but also strengthened the Research Council as a cohesive group.

The RAUH enhancement-led approach to research assessment is well aligned with current developments with responsible research and researcher assessment (e.g., CoARA). Thanks to the established qualitative approach, it allows the University leadership to gain a rich understanding of the development work and its meaning for the units and faculties.

17 Coalition for Advancing Research Assessment 2022.

18 University of Helsinki 2024f.

19 INORMS Research Evaluation Group 2021, 5.

The results of the follow-up show a variety of concrete steps stemming from the panel feedback, leading to impactful development activities. The quality of operations ensures the quality of results, too.

MAKING USE OF THE RAUH RESULTS IN CONTINUOUS STRATEGIC UNIVERSITY DEVELOPMENT

Research assessment at the University of Helsinki has been closely integrated into all strategic development, related to research activities and reflecting the continuous development driven quality culture at the University. The RAUH process interconnected especially with the Roadmaps for Implementing Research Themes (2021–2022). Initially, the Roadmap project was only loosely connected to research assessment, as its primary purpose was to provide tools to support the University's strategic goal of promoting multidisciplinarity and interdisciplinary research. The roadmaps were designed to serve as an intermediary between the university-level strategic goal and the academic unit implementation plans, providing direction for defining the concrete actions in them. Since the completion of the Roadmaps, they became foundational to the planning of RAUH 2025. This integration was driven by the recognition that the Roadmaps offered a comprehensive framework that aligned well with the University's strategic objectives, ensuring that research assessment and strategic development were seamlessly connected.

While the Roadmaps provided a foundation for planning RAUH 2025 and updating the assessment framework, there are several other concurrent actions that support RAUH 2025. In spring 2025, the University of Helsinki launched a training program on strategic management of research activities for the members of the University's Research Council. The university-level self-assessment report, with a focus on interdisciplinary research, which was part of the RAUH 2025 assessment, was prepared as part of the training by the Research Council's teamwork, enhancing strategic thinking at both the University and unit levels. This prompted the academic leaders at the University's Research Council to define the success from their own perspective and use the understanding gained towards reflecting on the RAUH 2025 assessment criteria.[20] This form of participatory collaboration has impacts on the University's future strategic development, too.

---

20 Kivilaakso 2025.

Through the 2018–19 and 2025 assessments, RAUH has evolved and became part of the University of Helsinki's quality system. The quality system at the University provides the necessary structures and defines the procedures and responsibilities to ensure that the University can maintain and develop the quality of its operations in a spirit of continuous development. Through this, the University can monitor its progress towards goals, identify areas needing improvement, and make necessary adjustments. The approach not only supports the University's strategic objectives but also ensures that society can trust in the University's operations and results.[21]

CONCLUSION – THE FUTURE OF RESEARCH ASSESSMENT

As demonstrated in the article, the feedback received during the RAUH 2018–19 assessment – particularly the panels' encouragement to promote interdisciplinarity – led to impactful strategic development actions. This is now reflected in the University's strategic goal, the implementation plans of both the University and its academic units, and the Roadmaps for Implementing Research Themes. The recommendation to invest in interdisciplinarity even influenced the focus of the RAUH 2025 assessment. This trajectory has played a central role in shaping the University's strategic direction and illustrates how the assessment process can serve as a catalyst for long-term development.

In 2025, RAUH is not just an overall assessment of research, but rather an important next step towards the implementation of the University's strategic goal of promoting interdisciplinarity. What will be the focus of the next research assessment in the early 2030s? The open-endedness continuous development philosophy allows this unfold along the strategic developments occurring at the end of the 2020s, for which RAUH 2025 is already laying the groundwork. In any case, in the spirit of enhancement-led assessment at the University, there will likely be some form of follow-up for RAUH 2025 before the next assessment cycle.

Looking ahead to the future of research assessment at the University of Helsinki, the focus increasingly will be on enhancing inclusivity and participation. The objective is to foster a cultural shift within the University community by creating long-term impacts that strengthen the academic community's inclusion and ownership of the assessment processes.

21 University of Helsinki 2024a.

This ensures that the academic community views these processes as collective efforts rather than mere administrative tasks. Emphasizing faculty leadership, the goal is to fully integrate the RAUH process into the units' strategic development. This approach aims to strengthen the community's inclusion and experience of the benefits of research assessment in continuous strategic and operational development. Starting from spring 2026, the RAUH 2025 results will be discussed thematically. These discussions will promote interdisciplinary dialogue, further embedding the principles of inclusivity and collective ownership in the research assessment process.

Integrating strategic alignment into the overall assessment of research is an ambitious attempt in the field of research assessment. To succeed, it requires careful and regenerative planning, building on existing processes and learned organisational strengths. One of the key lessons has been that research assessment can be conducted more ambitiously than merely evaluating current operations based on output analyses. It is essential to consider the collective learning process involved in examining both unit-level and university-wide activities. The idea of continuous development is practiced in research assessment, drawing from previous exercises such as Roadmaps for Implementing Research Themes and trust. Emphasizing the importance of participation is crucial, especially in building trust. Only through jointly planned and discussed efforts can the desired impact be truly achieved and owned. When assessment is used as a tool for strategic management, it simultaneously promotes organisational learning by enhancing organisational system thinking, team learning, and supporting the establishment of a shared vision within the University community.[22]

22 Senge 2006.

REFERENCES

Coalition for Advancing Research Assessment 2022. Agreement on Reforming Research Assessment. Retrieved on 12 May 2025. Available at https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

European Commission 2021. European Research Area Policy Agenda – Overview of actions for the period 2022–2024. Retrieved on 20 December 2024. Available at https://research-and-innovation.ec.europa.eu/system/files/2021-11/ec_rtd_era-policy-agenda-2021.pdf

Huutoniemi, K., Klein, J.T., Bruun, H. & Hukkinen, J. 2010. Analyzing interdisciplinarity: typology and indicators. Research Policy, 39 (1), 79–88. Retrieved on 20 December 2024. Available at https://doi.org/10.1016/j.respol.2009.09.011

INORMS Research Evaluation Group 2021. The SCOPE Framework. Emerald Publishing. Retrieved on 20 December 2024. Available at https://doi.org/10.26188/21919527.v1

Kivilaakso, A. 2025. Oppivan ja uudistuvan johtajuuskulttuurin jäljillä: tutkimustoiminnan johtamisen valmennuskonseptin kehittäminen Helsingin yliopistoon. Master's Thesis. Lahti: LAB University of Applied Sciences. Retrieved on 12 May 2025. Available at https://urn.fi/URN:NBN:fi:amk-202504166677

LERU 2025. League of European Research Universities. Retrieved on 12 May 2025. Available at https://www.leru.org/

Moitus, P. & Kamppi, P. 2020. Kehittävä arviointi Kansallisessa koulutuksen arviointikeskuksessa. Tiivistelmä 8:2020. Kansallinen koulutuksen arviointikeskus. Retrieved on 20 December 2024. Available at https://www.karvi.fi/sites/default/files/sites/default/files/documents/KARVI_T0820.pdf

Mälkki, A., Kolhinen, J., Raassina, M. & Väänänen, R. 2019. Research Assessment 2018–19. University of Helsinki. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/assets/drupal/2021-06/rauh_final_report_060919_full.pdf

Senge, P. 2006. The Fifth Discipline: The Art and Practice of the Learning Organization. London: Random House.

Universities Act 558/2009. FINLEX, Finnish Ministry of Justice. Retrieved on 20 December 2024. Available at https://www.finlex.fi/en/laki/kaannokset/2009/en20090558_20160644.pdf

University of Helsinki 2024a. Quality Management. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/en/about-us/strategy-economy-and-quality/quality-management

University of Helsinki 2024b. Strategic Plan 2021–2030. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/en/about-us/strategy-economy-and-quality/strategic-plan-2021-2030

University of Helsinki 2024c. Organisation. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/en/about-us/university-helsinki/organisation

University of Helsinki 2024d. We are the University of Helsinki. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/en/about-us/university-helsinki

University of Helsinki 2024e. Campuses. Retrieved on 20 December 2024. Available at https://www.helsinki.fi/en/about-us/visit-us/campuses

University of Helsinki 2024f. University of Helsinki CoARA Action Plan 2024. Retrieved on 12 May 2025. Available at https://zenodo.org/badge/DOI/10.5281/zenodo.12525695.svg

Wernli, D. & Ohlmeyer, J. 2023. Implementing interdisciplinarity in research-intensive universities: good practices and challenges. Advice paper no. 30 – March 2023. League of European Research Universities LERU. Retrieved on 20 December 2024. Available at https://www.leru.org/files/Publications/Implementing-interdisciplinarity-in-research-intensive-universities-good-practices-and-challenges_Full-paper.pdf

Willamo, R., Erkkola, M., Kettunen, H., Rekola, M., Salmesvuori, P. & Virtala, A. 2022. Tieteidenvälisillä poluilla. Kokemuksia tieteidenvälisten tutkielmien ohjauksesta ja arvioinnista. Helsinki: Helsingin yliopisto.

AUTHORS

AURA KIVILAAKSO
University of Helsinki
P.O. Box 53 (Fabianinkatu 32)
00014 University of Helsinki, Finland
Email: aura.kivilaakso@helsinki.fi
ORCID: 0009-0006-4583-6620

JOHANNA KOLHINEN
University of Helsinki
P.O. Box 53 (Fabianinkatu 32)
00014 University of Helsinki, Finland
Email: johanna.kolhinen@helsinki.fi
ORCID: 0000-0001-8138-2570

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e3, 1-25
DOI: 10.22163/fteval.2025.695

THE ROLE OF CHANGE AGENT CHARACTERISTICS IN RESEARCH ASSESSMENT:

EXPERIENCES FROM AN EVALUATION OF A RESEARCH FELLOWSHIP PROGRAM IN GERMANY

HENDRIK BERGHÄUSER
DOI: 10.22163/FTEVAL.2025.695

ABSTRACT

Research funders are increasingly faced with the challenge to design funding programs generating both scientific and social impact, which in turn places new demands on research assessment processes. Using an evaluation of a german research funding program targeting scientists from the Global South, this paper examines how different funding goals and selection criteria - in this case research excellence on the one hand and social or development relevance on the other - can be combined in research assessment and to what extent change agent qualities of the applying researchers can be integrated into the assessment process in order to increase the chances for social impact. Drawing on interviews, a survey of funded researchers as well as a survey of a control group of non-funded scientists the study shows that funded researchers indeed, have much stronger change agent attitudes than scientists in the non-funded control group. Differences with regard to concrete change agent actions, on the other hand, cannot be determined. All in all, change agent characteristics represent a promising element for research assessment procedures, but the selection criterion suffers from an insufficient definition and conception, especially for the science sector.

Keywords: Research assessment, Social impact, Change agents, Global South

1. INTRODUCTION

As research increasingly seeks to contribute to solutions for social problems, with strong emphasis on transfer activities and social engagement (Belcher et al. 2016; Bornmann 2013), research funders are faced with the challenge to design funding programs which generate both social and scientific impact. This implicates a need for research assessment processes that take into account and deal with tradeoffs in both research excellence and the potential of social impact, especially when diverse actors and researchers outside mainstream circles are addressed (Kraemer-Mbula et al. 2020; Ferretti et al. 2018). This paper provides insights on an evaluation of a research fellowship program that aims to enable excellent scientists from developing and emerging countries to spend a research period in Germany. The fellowship program wants to recognize the relevance of researchers from these countries in achieving the 17 UN Sustainable Development Goals (SDGs) (UN General Assembly 2015). Thus, classic indicators of research excellence such as key publications, scientific impact (h-index) etc. fall short or need to be contextualized, given the heterogeneity of researchers from 137 eligible countries and the partly difficult conditions in the science systems in the Global South (Tijssen and Winnink 2022). At the same time, the notion of “social impact” or “contributions to SDGs” appear abstract and elude standardized, indicator-based objective evaluation. Therefore, in addition to scientific merits, the research assessment jury of this funding program also looks for change agent characteristics among the applicants (Caldwell 2005; Hutton 1994; Ottaway 1983). This is linked to the assumption that certain personal characteristics increase the likelihood that the results of the funded research projects will eventually find their way into application in the home countries of the researchers and thus generate a developmentally relevant impact.

The focus of this paper is therefore twofold: First, it examines and discusses possibilities and limits when linking two different funding objectives - research excellence and social or developmental impact - as well as potential conflicting goals. Second, the paper looks at the potential of change agent characteristics of the applicants as a criterion for research assessment in individual funding programs.

The evaluation object is the Georg Forster Research Fellowship of the Alexander von Humboldt Foundation, which is funded by the Federal Ministry for Economic Cooperation and Development in Germany. The selected researchers receive a monthly sponsorship for a research stay in Germany of up to two years and are hosted by a german scientific institution during their research period.[1] Decisions to award fellowships are based on the assessment of the applicant's academic qualification, their contribution to development-related research as well as their future potential.

2. THE ROLE OF CHANGE AGENT CHARACTERISTICS IN RESEARCH ASSESSMENT

While studies on the contribution of success in academic research can be traced back to over a half a century (Merton 1957; Zuckerman and Merton 1971; Hagstrom 1975) debates about research assessment processes intensified in the 1980s, when the growth in public research funding led to a growing workforce and a stronger need for clear and transparent distribution of scarce financial resources for funding allocation, hiring, tenure and promotion became obvious (Alberts et al. 2014; Bonn and Bouter 2023). As a consequence, publication metrics, in particular publication counts, citations, the H-index as well as the journal impact factors started being used in research assessment in order to provide a greater sense of objectivity than traditional forms of peer-review qualitative assessment (Gingras 2014; Bonn and Bouter 2023). However, due to the strong focus on these mostly quantitative metrics, research assessment has come under increasing pressure. In particular, the underlying concept of research excellence and the application of too narrow criteria and indicators of research quality was criticized. According to this critique, the strong focus on these metrics or the narrowing of quantitative criteria and indicators reduces the diversity of research missions and purposes, leading researchers to adopt similar strategic priorities or to focus on lower-risk, incremental work. Furthermore, the systemic biases against those who do not meet these narrowed indicators of quality or impact, or who do not conform to certain career pathways, reduce the diversity and representative legitimacy of the research community (Curry et al. 2020; Moed 2020; Tijssen and Winnink 2022). In addition, the application of these metrics, combined with an increasing competition for research funding might distort incentives, create unsustainable pressures on researchers and thus lead to unethical behavior (Edwards and Roy 2017; Moher et al. 2018). Therefore, efforts were coalescing around the idea of a more responsible approach of

1 For more information about the funding program, see: https://www.humboldt-foundation.de/en/apply/sponsorship-programmes/georg-forster-research-fellowship.

metrics and its use that lead to new initiatives in research assessment like the Declaration on Research Assessment (DORA) in 2012 (DORA 2021), the Leiden Manifesto for research metrics (Hicks et al. 2015) or the Metrics Tide (Wilsdon et al. 2015) that recognized the need to improve the ways in which researchers and research outputs are evaluated.

At the same time, an increased focus on directionality and mission-orientation can be observed in science policy. This is accompanied by expectations of research to develop concrete solutions for societal challenges (Boon and Edler 2018; Mazzucato 2018), leading scholars to conceptualize impacts beyond academia and the contribution of science to addressing current or future social, environmental, economic, policy and other needs or problems (D'Este et al. 2018; Reale et al. 2017) as well as to develop frameworks to evaluate such impacts (Kok and Schuit 2012; Joly et al. 2015; Smit and Hessels 2021; Spaapen and van Drooge 2011; Matt et al. 2017; Donovan and Hanney 2011; Lauronen 2020).

While research assessment procedures are under pressure, the demands placed on researchers applying for research funding have also changed, especially in funding programmes targeting both social and academic impact. Since funded researchers are expected to promote not only scientific but also social impacts through their research work, personal characteristics, which can generally be described as change agent characteristics, are increasingly coming to the fore in research assessment alongside aspects of research excellence. The first comprehensive definition of 'change agents' was developed by Beckhard (1969) who describes change agents as people either inside or outside of an organization who provide technical, specialist or consulting assistance in the management of a change effort (Beckhard 1969). Roger and Shoemaker define change agents as professionals who influence innovation decision in a desirable direction (Rogers and Shoemaker 1971). They bring about purposeful transformation and help people to change the way they think - changing the norms and changing the organization's systems and processes (Hutton 1994). Moran and Brightman (2000) look at the necessary characteristics of change agents and find out that persons need to have necessary skills to initiate and manage change processes, e.g. in leadership, creativity and problem-solving (Moran and Brightman 2000). More recent studies underline the personal skills needed for change agency like openness, empathy, energy and networking Lunenburg (Lunenburg 2010). As a consequence, researchers can also act as change agents and contribute to social impact through promoting public values and knowledge transfer activities or by focusing on the social relevance of research (Bornmann 2013).

The concept of change agents could therefore also link the concept of social impact and research excellence. Tijssen and Winnink (2022) differentiate in this context between 'global excellence' and 'local excellence'. Whereas global excellence is primarily determined by acknowledged scientific visibility and (partially) measurable reputation within the international research community, local excellence instead relates to the utilisation of knowledge and know-how among non-scientific users and local communities. Nevertheless, the concept of change agents in science remains vague and ill-defined.

3. METHODOLOGICAL APPROACH

The methodological approach used here is based on a triangulation of different quantitative and qualitative as well as reactive and non-reactive empirical methods used for the evaluation of the mentioned funding programs. The analysis relies on participant observation of two meetings of the research assessment jury based on an observation guide, five interviews with program managers and representatives of the research assessment jury, an online survey of all research fellows who have received program funding between 2010 to 2020 and a control group survey with non-funded scientists from developing and emerging countries with comparable researcher characteristics identified by using bibliometric methods.

4. PROGRAM STAKEHOLDER INTERVIEWS AND PARTICIPANT OBSERVATION OF RESEARCH ASSESSMENT JURY MEETINGS

The evaluation process started with a group interview with several program managers from the funding agency in Spring 2022. The goal of this initial interview was to obtain more detailed information about the funding program, to discuss the overall goals of the program and to better understand the underlying program logic. In addition, the procedures of the triannual research assessment jury meeting and the criteria for selecting applicants for the research fellowship were discussed during this interview. Consequently, this group interview provided the basis for the subsequent planned participant observation of two research assessment jury meetings.

After the interview and an analysis of funding guidelines and further program information for applicants, the preparation for the participant observation of the research assessment jury meetings started. The participant observation used in this case referred to an unstructured procedure with an explorative objective of recording background conditions (Kawulich 2005). Specifically, two research assessment jury meetings were observed (in June and October 2022). As a prerequisite, all stakeholders involved (program managers and jury members) were informed about the process of the participant observation. Consent was obtained from all participants. The participant observation was then conducted in an open and non-active participatory manner, i.e. the members of the research assessment jury were aware of the participation of the evaluators, yet the evaluators did not actively participate in the discussion but merely observed the process instead. The observation followed a previously developed observation scheme in which the objects of observation and the tasks are concretized. The strength of this method is that it allows the observation of research assessment and selection processes and thus actual behavior regarding the implementation of the two selection criteria. The observation guide was derived from the previous interview and the written selection criteria:

Observations regarding assessment criteria "research excellence"	Observations regarding assessment criteria "social / developmental relevance"
▪ Observations regarding the topic of scientific career and scientific achievements (mobility, determination, breadth of expertise, scientific productivity)	▪ Observations regarding the relevance of the research projects for the further development of the country or region of origin (including the 17 SDGs of the UN)
▪ Observations regarding the quality of the key publications name in the application (originality, degree of innovation, own contribution in the case of multi-author publications)	▪ Observations regarding change agent characteristics of the applicant or possible multiplier effect in research, teaching, science management and development-relevant processes and activities outside academia
▪ Observations regarding the originality and innovative potential of the proposed research project (significance for the further development of the field, convincing choice of scientific methods, possibilities for further scientific development, feasibility, etc.)	▪ Observations regarding further development-relevant aspects (for example gender-specific aspects, special need for support with regard to other specific region or the concrete research area)
▪ In the case of postdocs, how is the scientific potential of the applicant assessed (further scientific development, career prospects etc.)? In the case of experienced researchers: How is the stand-alone scientific profile of the researcher evaluated?	

In addition to operationalizing the two selection criteria - research excellence and social / developmental relevance - the participating observers were also looking for “soft” aspects regarding the discussion dynamics in particular:

- Dominant actors (jury members dominating the discussion beyond their role),
- Gender-perspective (are all participants fairly engaged in the discussions, regardless of their socio-demographic characteristics? Is there any evidence for homosocial biases, i.e. men favoring men or women favoring women?),
- Balance of power between different actors within the research assessment jury meetings (e.g. how passive are the funders vis-à-vis the experts?),
- How are controversial decisions handled? How are conflicts resolved?

As the research assessment jury meetings took place in a hybrid format, both meetings were observed by two participants, one taking part virtually and one taking part physically. In this way, the authors were able to evaluate how the two main selection criteria - research excellence and social / developmental relevance of the applying researchers and their proposed research projects were being assessed and how these two selection criteria were connected in the discussion within the jury.

The two observers recorded their observations of the two half-day research assessment jury meetings in a result protocol. The results were then compared and discussed. Divergent perceptions and observations were not included in further analysis. Subsequently, the results logs were coded and analyzed using MAXQDA, a qualitative content analysis software.

After the participant observation, four interviews were conducted with members of the research assessment jury. The interviews were semi-standardized and based on an interview guide. The goal of the interviews was to find out about the individual understanding and interpretation of the selection criteria, as well as to mirror observations of the jury assessment meetings (Laudel and Gläser, 2007). Accordingly, the interviews addressed the perception and linkage of research excellence and social or developmental relevance, including change agent characteristics, in the selection process. When selecting the jury members for interviews, we paid attention for a high diversity with regard to disciplinary orientation (e.g. humanities, natural sciences, engineering), gender and region-specific knowledge, which were of great importance in the evaluation of the applicants from the global south. The interview partners were selected after the jury meetings. Therefore, also the behaviour of the reviewers during the jury meetings was also taken into account. For example, among those selected were reviewers who highlighted or critically questioned certain selection criteria (e.g. ‘excellence’ or ‘social / developmental relevance’) during the discussion of the applications. The interviews were recorded and transcribed. Subsequently, the transcripts were also coded and analyzed using MAXQDA software.

4.1 ONLINE-SURVEY AMONG FUNDED RESEARCH FELLOWS

In fall 2022, preparations began for the quantitative analyses, in particular the bibliometric analyses as well as the online surveys of the selected and funded research fellows and of the control group. On December 5, 2022, the survey of the funded fellows was launched, using the EFS survey tool by Questback. A total of 754 persons were invited to take part, of which 20 persons could not be reached for technical reasons (final adjusted total gross sample: 734). On December 14, a reminder mail was sent out, in which all those persons who had not yet participated by that time or who had dropped out of the survey were contacted again. A second and final reminder campaign was launched on January 9, 2023. The survey was closed on January 24. The results dataset was then downloaded and cleaned. In the end, 505 valid responses were counted, which corresponds to a substantial participation rate of 68.8%.

4.2 CONTROL GROUP SURVEY AMONG NON-FUNDED BIBLIOMETRIC TWINS

In the evaluation, bibliometric analyses were mainly used to investigate the publication performance of the funded fellows (before and after the funded research stay) and to measure the development of the fellows’ international scientific collaboration (through co-publications). In addition, bibliometrics was also used as an approach to generate a control group to be surveyed subsequently. For the bibliometric formation of the control group, Scopus database was used. For each funded research fellow, one or more “twins” were bibliometrically generated by searching for a scientist with similar characteristics. There characteristics were:

- Scientific discipline: Here the journals, in which the funded fellow mainly publishes served as a reference,
- Gender (male / female): this is generated from the first name of the scientist,
- Scientific age: This is calculated based on the first scientific publication covered in Scopus,

- The country / region of origin: a minimum number of publications in the years prior to the funding year (of the funded fellow) in the funded fellow's home countries or regions,

Publication performance: A similar publication output of a researcher and citation rate at the time the fellow is funded. In a first step, however, the funded fellows needed to be identified in Scopus. Of the 809 funded research fellows, 651 could be clearly identified by name and e-mail addresses (80.5%).

For these 651 researchers, "twins" were then searched for in Scopus using the above defined criteria. In order to obtain a larger sample for the control group survey, multiples were also counted (i.e. if there was more than one bibliometric twin for a fellow). However, this also means that some research fellows had a stronger weight in generating the comparison group than others. As a result, complete representativeness was not achieved. But this was taken into account when comparing the survey data. Nevertheless, these methodological limitations were accepted in order to achieve a sufficiently large sample that promised a solid response in an online survey. In the end, 1918 researchers were identified as potential twins for the control group survey. The following table compares the two groups surveyed in terms of key characteristics.

Research fellows		Control group	
Gender		**Gender**	
Male	71.2%	Male	67.5%
Female	27.5%	Female	29.8%
Diverse / other	0.4%	Diverse / other	0.4%
n/a	0.8%	n/a	2.4%
Discipline		**Discipline**	
Agriculture, forestry, horticulture and veterinary medicine	9.6%	Agriculture, forestry, horticulture and veterinary medicine	11.4%
Biology	16.7%	Biology	13.3%
Chemistry	16.2%	Chemistry	7.5%
Humanities	15.6%	Humanities	16.1%
Geosciences	3.8%	Geosciences	2.7%
Engineering sciences	11.3%	Engineering sciences	10.6%
Mathematics	2.8%	Mathematics	1.6%
Medicine	3.8%	Medicine	5.5%
Physical sciences	3.8%	Physical sciences	3.5%
Social and behavioural Sciences	10.3%	Social and behavioural Sciences	16.5%
Other	6.0%	Other	11.4%
Scientific age (here: date of doctorate degree)		**Scientific age (here: date of doctorate degree)**	
Before 2000	1.2%	Before 2000	8.7%
2000 – 2004	9.3%	2000 – 2004	6.2%
2005 – 2009	23.0%	2005 – 2009	16.6%
2010 – 2014	36.2%	2010 – 2014	29.1%
2015 – 2019	20.6%	2015 – 2019	15.2%
2020 and later	0.6%	2020 and later	8.7%
n/a	9.1%	n/a	15.6%
Country of origin		**Country of origin**	
651 research fellows from 62 countries in the Global South were identified in Scopus		Comparable twins were identified from 59 of 62 countries.	
Publication performance		**Publication performance**	
individual		Identical number of pub.	58.2%
		+ / - 3 publications	26.6%
		+ / - 5 publications	15.2%

Some of the key characteristics, for identifying bibliometric twins were validated in both surveys and could thus be determined even more precisely. The participants of the surveys were asked for examples about their scientific discipline, their gender and the year of their doctorate's degree. In spite of the mentioned methodological limitations, the two groups are very similar in terms of key characteristics, as table 2 shows. There are slight differences in the distribution of scientific age and publication performance.

The control group survey started on March 1, 2023. Of the 1918 identified researchers, 1753 could be reached. An initial reminder campaign was launched on March 12 and a second and final one on March 26. On April 3, 2023, the control group survey was closed, and the results dataset was cleaned according to the same procedure as the dataset of the fellowship survey. The final number of participants was 189, which represents a participation rate of 16.5%.

In order to avoid a selection bias, the control group survey was intentionally framed rather broadly in terms of content. In the invitation email, the aim of the study was described as basically asking about the motives and attitudes of researchers in the Global South. In this context, corresponding questions on the role of science in society and about change agent characteristics were asked and analysed. Therefore, a possible selection bias can be largely excluded. The large difference in the participation rate between both surveys is, primarily due to the close relationship between the funded researchers and the funding agency and the explicit reference in the survey of funded research fellows, that the survey is related to the evaluation of the funding programme.

5. RESULTS

5.1 PROGRAM STAKEHOLDER INTERVIEWS AND PARTICIPANT OBSERVATION OF RESEARCH ASSESSMENT JURY MEETINGS

The jury meeting was preceded by an external, written assessment of all applications by at least two reviewers. In the jury meetings, only those applications were presented and discussed for which the external reviews did not provide a clear picture. Applications with two poor external reviews for example were already excluded before the jury session. The same applies to excellent applications that received two very good external reviews. These clear funding cases ("F-cases") were also not discussed at the jury meeting. These candidates received funding without an additional discussion and jury vote. Consequently, the assessment jury only dealt with the so-called discussion cases ("D-cases") which, however, make up the majority of applications. The research assessment jury meetings proceeded in a highly structured manner. Each jury member presented the applications that were assigned to them based on their professional profile and regional expertise. Within a few minutes, the respective applicant and the proposed research project were briefly presented and discussed. In doing so, the jury member gave his or her assessment about the research excellence and the developmental relevance and made a final recommendation for or against funding. In the assessments, reference was also made to the external review reports that were available to all jury members. Afterwards, there was a short question and answer session before the next application was presented. Voting was done in secret at the end of the meeting (each jury member could distribute 0,1 or 2 points for each candidate).

During the presentation and discussion of the applications, a rather benevolent review could be observed overall. The jury members predominantly expressed respect and appreciation for the candidates ("very good candidate", "very interesting project", "very determined candidate", "has high potential", "very impressive" etc.).

In addition, the candidates were presented and assessed in comprehensive form. The positive aspects and negative impressions were weighed up. In the discussion, one single criterion was rarely decisive for the assessment. Moreover, all selection criteria (e.g. academic career, scientific productivity, key publications or developmental relevance) were considered with regard to the individual context. Individual characteristics such as age, gender, origin as well as private life circumstances were explicitly taken into account in the evaluation ("The candidate is already somewhat older. But she has raised three children in Africa and earned a doctorate in Japan. That is impressive"). The individual context conditions in the home countries of the applicants were also considered in the assessment ("the publication performance of this candidate is not outstanding, but it must be taken into account that the candidate, as is common in this country, was employed as a lecturer for a long time, which meant a high teaching load and little room for research and publication").

However, it remains unclear, how exactly the different levels of development in the applicants' home countries were weighted. At the opening of the meeting, the funding authority pointed out the applications from scientists from least developed countries (LDCs) and asked for a particularly favorable evaluation. In the subsequent interviews with jury members after the jury meetings, it was said that applicants are not directly compared with each other. Nevertheless, a very good candidate from a very poor country may receive a higher rating than an equivalent candidate from a more developed country, since the performances are to be weighted differently with regard to the different conditions in the home countries. However, in some cases, especially in the case of candidates from the least developed countries, it was discussed whether and to what extent the candidates would be able to continue their work in his or her home country after a research stay in Germany. In one case, this was doubted. It was argued that there were no appropriate infrastructures in the home country to continue this research or even to establish it in the first place. In this case, funding was not recommended, since the probability of generating social impact through the fellowship was considered rather low and the risk of brain drain through the funded research stay was considered to be high. The discussion made it clear that academic expertise in each discipline coupled with regional knowledge are necessary in order to be able to thoroughly and fairly evaluate a candidate and to contextualize individual achievements ("All the applicants' key publications are in Spanish. But we have to consider that in [this country] science is mainly published in Spanish and addresses local and regional scientific audiences first"). In assessing the potential of the applicants, particular consideration was given to their academic performance over the course of their academic career ("his curve goes steeply upwards" or "In the beginning it was a little less, but now she is very active and publishes a lot"). In addition, a higher age (e.g. 60 years or older) tended to be evaluated rather negatively, since the chances regarding scientific potential or social impact were estimated to be rather low.

Developmental and/or social relevance counts, at least formally, more than scientific excellence in this funding program. In fact, developmental and/or social relevance was evaluated first in the selection process before scientific excellence was evaluated. If no developmental relevance was seen, or if it was deemed to be too low, the application was immediately sorted out and not admitted to the final vote at the end of the session. This procedure also describes how the two selection criteria - scientific excellence on the one hand and developmental and/or social relevance on the other hand - are linked to each other in this funding program: Developmental and/or social relevance is basically the prerequisite for receiving any chance of funding at all. The two selection criteria are not weighed against each other.

At the same time, the assessment of developmental and/or social relevance often remained very vague. Often, the justification for the assessment was not further elaborated ("the proposed project is clearly developmentally relevant" or "the developmental relevance is given"). Sometimes it remained unclear what the developmental relevance refers to, e.g. which regional focus is considered (developmentally relevant for a specific region, country, continent or even reference to global challenges), or which of the SDGs are addressed or for which social areas or social groups possible impacts are foreseeable. Discussions about developmental relevance mostly arose among more basic research-oriented projects, as well as among applications from social scientists and humanities scholars. Here, members of the research assessment jury more often asked about the development relevance of the research project than in the case of applied-oriented research projects that are more associated with general development cooperation, like projects dealing with micro plastics or the contamination of drinking water. In this context, there were also occasional lively discussions and appeals from jury members not to focus solely on application-oriented research projects, even if the developmental relevance may be more obvious here at first glance. The relevance of basic research, which might not have an immediate impact on society but can contribute to social impacts eventually, was then frequently emphasized. Discussed examples included projects that deal with specific mathematical or physical models that can provide contributions for subsequent weather forecasts or projects in biology for the classification of plant or animal species, in order to obtain contributions for a better understanding of biodiversity in a particular region. All in all, the impression was that developmental relevance needed to be justified more strongly in the case of basic research-oriented projects or in the case of social sciences and humanities. This tension was confirmed in the subsequent interviews with jury members: "We always get to that point with theorists regarding their developmental relevance" (Interview no. 3).

Overall, a very broad understanding of developmental relevance could be observed. According to the jury, the vast majority of the proposed research projects of the candidates were assessed as developmentally relevant to some degree. In fact, only a few applicants were denied on the grounds of the developmental and/or social relevance of their proposed research projects. Consequently, development relevance seems more like a soft selection criterion.

A special aspect in the evaluation of developmental relevance are the change agent characteristics. These too, have not been defined in any further detail at any point. This usually involved a special commitment to science and/or society. Examples given included special commitment to the training of doctoral students and undergraduates, the establishment of study programmes, research areas or entire research institutes or involvement in various professional societies or committees. In addition, special emphasis was placed on engagement with social actors that are closely related to the applicant's scientific work. This often refers to various transfer activities in which the applicants were engaged. In this context, various examples of knowledge and technology transfer (i.e. candidates planning science-industry cooperation, citizen science projects or candidates consulting policy makers) as well as civic engagement were described (private involvement with an NGO, founding a civil society association, etc.). In addition, professional experience outside academia, e.g. in industry, civil society or politics, was also positively highlighted in the presentation of the candidates. It was noticeable that, on the one hand, special engagement efforts from applicants were appreciated and strongly emphasized as a positive example of change agent characteristics. Yet, on the other hand, a lack of such notable engagement was not usually highlighted negatively, indicating that change agent characteristics is a rather "soft" selection criterion.

5.2 RESULTS OF THE ONLINE SURVEYS

The aim of the control group approach was to identify possible differences in change agent characteristics between funded research fellows and their bibliometric twins as these characteristics are a key evaluation and selection criterion for the funding programme. Due to the vague definition and operationalization of the concept 'change agents' in the scientific context, both in the academic literature and in the funding guidelines of the programme, the topic was approached in the two online surveys by asking about the social or developmental relevance of the research and about individual exchange, transfer and engagement activities. Furthermore, both funded and non-funded scientists were asked about their attitudes regarding priorities and motivations in their scientific work and the role science in general and scientists in particular should play in society.

The results of the comparative study are somewhat surprising. When asked "to which areas of society outside of science is your research relevant" (see Figure 1) the participants in the control group indicated a higher social relevance of their research for every social area compared to the funded research fellows (here: politics, business / industry, end-users and professionals, civil society organisations, citizens, media and culture and arts). This is surprising, since social or developmental relevance is an important assessment criterion in this research funding programme.

Similarly, in the subsequent question "Have you had any interaction with this group / these groups at any stage, before, during or after the research process?", the participants in the control group indicated that they interacted more frequently on average with actors from all areas of society than the group of the funded research fellows did (see Figure 2).

The picture is not quite clear for the question "What kind of transfer channels do you use to make research results accessible outside of academia?" (see Figure 3). Here, the answers of the participants of the funded research fellows and the participants of the control group survey don't differ very much with regard to the frequency of the mentioned transfer, engagement and exchange activities (research collaborations with non-scientific partners, contract research, scientific consultancy services, further education for non-scientists, commercialization activities, public engagement / citizen science, service learning, science communication), even if the frequencies indicated for these activities also tend to be somewhat higher for the control group than for the group of the funded research fellows.

On the other hand, when asked "Scientists may have different views about which goals should be given high priority in the science system. What priority do you think the following goals should have in the scientific system?", all priorities were rated higher in the survey among the funded research fellows compared to the control group survey (see Figure 4). This includes priorities regarding responsible research, transfer and exchange activities as well as social relevance of research.

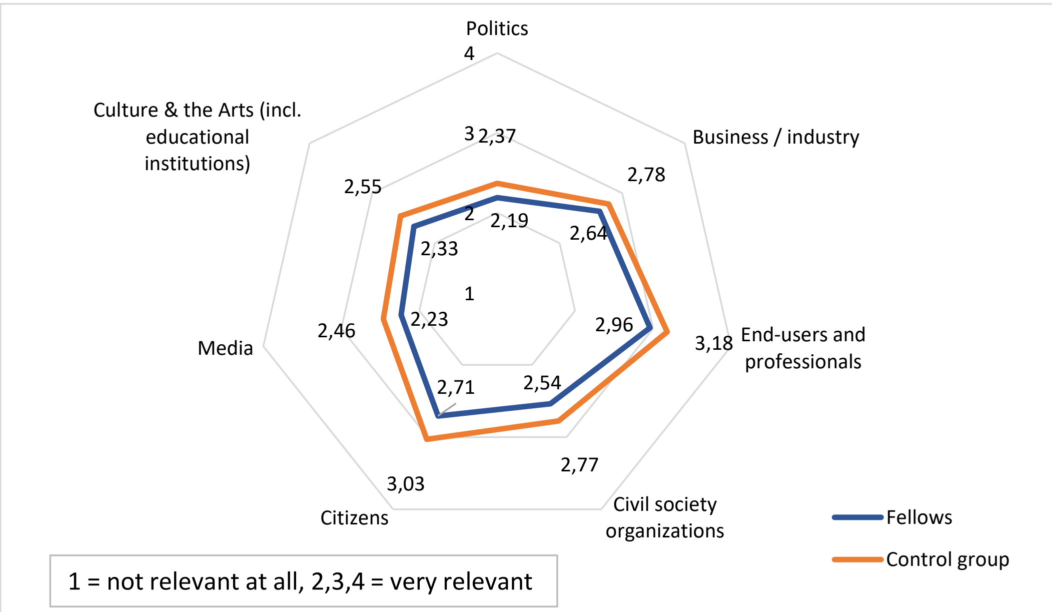


Figure 1: "To which areas of society outside of science is your research relevant?"

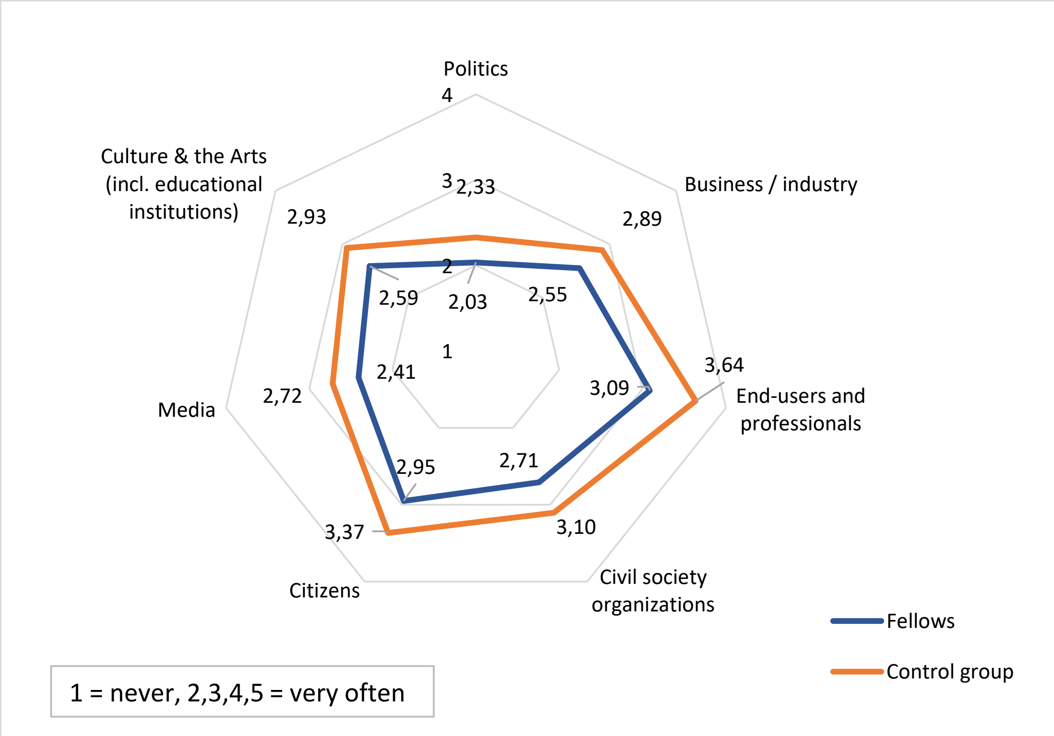


Figure 2: "Have you had any interaction with this group/these groups at any stage, before, during, or after the research process?"

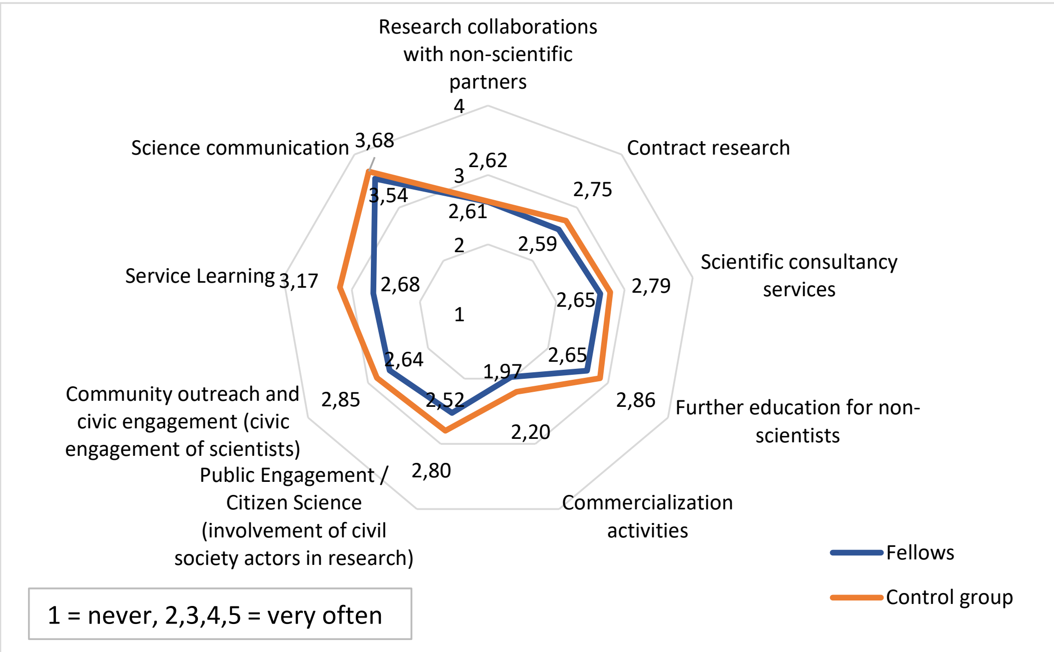


Figure 3: "What kind of transfer channels do you use to make research results accessible out-side of academia?"

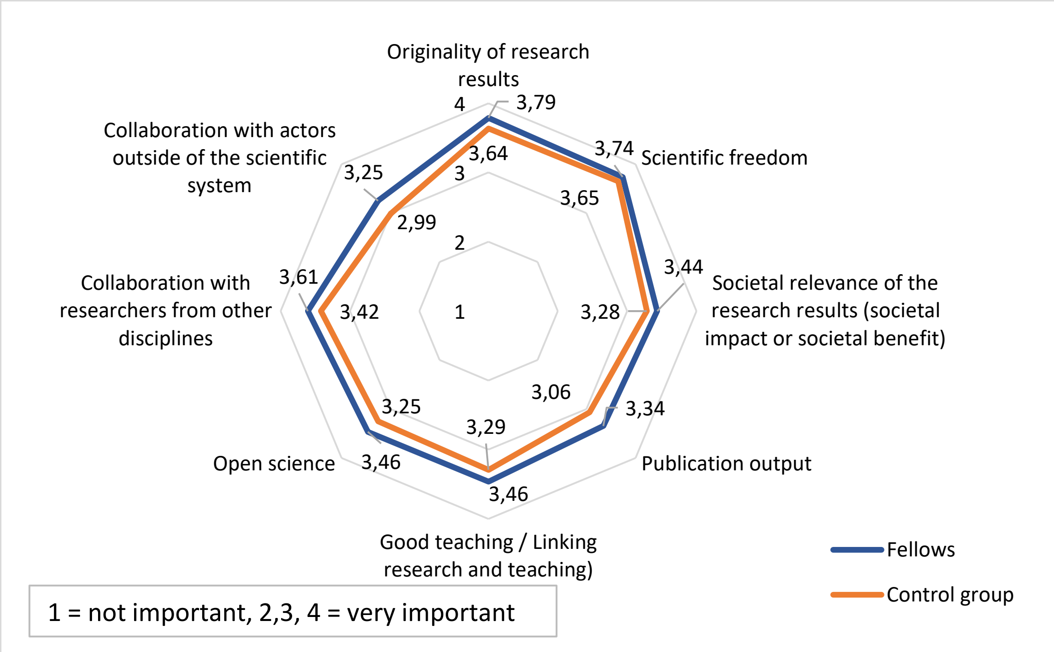


Figure 4: "What priority do you think the following goals should have in the scientific system?"

In addition, we asked scientists about their attitudes toward the role of science in society and the relationship between science and society. Here, the participants in both surveys were each asked to locate themselves between two opposing positions (semantic differential):

- Position 1a: "Science should retain a high degree of autonomy with regard to the demands of society." Position 1b: "Science should be at the service of society and provide solutions for social problems."
- Position 2a: "Scientists should actively engage in public debates." Position 2b: "Scientists should refrain from participating in public debates".
- Position 3a: "Scientists should only make statements about their own research." Position 3b: "Scientists should also contribute to topics beyond the scope of their own work."

Here too, it became obvious that the funded research fellows see science more as an important lever to promote change within society and to generate development-relevant impact. For example, the fellows were more vocal about the need for science to provide solutions to social problems, for scientists to be active in social debates and for more engagement in general. The differences are, however, statistically not significant.

Overall, the control group approach shows that the funded research fellows are more open to an active role for science in society. However, this does not manifest itself in a higher level of commitment or in a stronger engagement in transfer or exchange activities with social stakeholders.

6. DISCUSSION

This paper addresses the questions of how research funding programmes can generate both, social and scientific impact and to what extent change agent characteristics of the applying researchers can be integrated into research assessment processes. A triangulative methodological approach was used to analyze the selection process of a german research funding program targeting excellent scientists from the Global South. The funding program enables these scientists to conduct a research stay of up to two years in Germany. The effects of this research are intended to benefit the scientists' home countries. Consequently, the proposed research projects are expected to have a social and/or development relevance.

As the experience from this funding program shows, it is possible to combine funding objectives such as scientific excellence on the one hand and social impact or - as in this case - developmental relevance on the other. When implementing these funding objectives in the research assessment process, a prioritization and operationalization of the selection criteria appears necessary. In this program, development relevance or foreseeable social benefits is a necessary condition for funding and an initial selection criterion. Subsequently, the scientific excellence of the applicant and the proposed research project is evaluated. It seems particularly expedient to contextualize both selection criteria and to operationalize them as broadly as possible in order to comply with the heterogeneity of the researchers and their individual and region-specific conditions. In this program, the applicants were assessed in comprehensive form. Single criterions were rarely decisive for the assessment. Instead, all selection criteria - academic career, scientific productivity, key publications and developmental relevance - were considered with regard to the individual context. Characteristics such as age, gender, origin as well as private life circumstances were explicitly taken into account.

At the same time, the conceptualization of social and/or developmental relevance remains vague. Almost all presented and discussed applications in the research assessment process were considered to have social and/or development relevance in the proposed research projects. However, the assessment of development relevance often appeared to be subjective. A scheme for assessing development relevance based on clear indicators appears to be necessary (e.g. relevance to which SDGs or which social problems? Relevance for which social actors or area?). Nevertheless, similar to scientific relevance, a contextualized and broad understanding of development relevance seems to be of central importance here as well. For example, to narrow of a focus on topics of classical development cooperation - in the discussion among research assessment jury members, reference was often made to typical "bridge and well projects" in poor countries - would fall short. Also, a strong focus on immediate social outputs and outcomes would have the negative effect of promoting predominantly application-oriented research. More fundamentally-oriented research or research projects from certain scientific disciplines (e.g. from humanities or natural sciences) would tend to be disadvantaged. This can contribute to a negative trend, as research in the Global South, compared to the wealthier countries in the North, is already highly application-oriented. Rather, the developmental relevance of basic research should also be appreciated in the research assessment process.

As a further step, so-called change agent characteristics of the applicants were evaluated in this funding programme. As the analyses show, clearer criteria here are also important with respect to what exactly is meant by change agent characteristics and what their significance is for the selection process. As observed in the research assessment jury sessions, change agent characteristics were mostly equated with special scientific or (civic) social engagement. It was positively emphasized in several applications. However, the lack of such an engagement was not evaluated negatively. However, the surveys showed that change agent characteristics are a possible approach to increase the likelihood of realizing social impact through research funding and to connect scientific excellence with developmental and/or social relevance. According to the control group survey, the funded scientists had significantly stronger change agent attitudes than the non-funded researchers with regard to the question of what role science should play in society. This refers in particular to the statement (i) science should be at the service of society and provide solutions for societal problems, (ii) scientists should actively engage in public debates and (iii) scientists should contribute to topics beyond the scope of their own work. These positions received higher approval ratings among the funded research fellows than among the non-funded control group. This indicates that the funding programme is quite successful in selecting those scientists who support an image of science that is actively involved in society and shapes social developments. However, this did not manifest itself in higher engagement or transfer activities, which were used as indicators for change agency in this study. Nevertheless, it is important in this context to point out the limitations of the study design, especially the methodological approach to generating the comparison group using bibliometric twins. It was not possible to form a fully representative comparison group, as for some of the research fellows no identical bibliometric twin could be identified. This also shows the possibilities and limitations of control group approaches in many real evaluation contexts, which are increasingly demanded by the evaluation community when analysing science, technology and innovation funding measures (EFI 2024). Due to the individual design of the programme and the numerous influencing factors and characteristics of the treatment group that need to be considered, it is difficult to identify a representative control group - at least in this case.

At the same time, the application of selection criteria in research funding processes that go beyond purely scientific aspects of the application or the scientific performance of the researchers and consider personal characteristics or engagement activities should be handled with great caution. They can also harbour risks, especially if they are not clearly conceptualized and defined. For example, negative incentives could arise for scientists to overload their research proposals with unfounded and completely exaggerated expectations regarding societal impact. Reviewers could be distracted from other selection criteria such as originality or scientific standards by too much story-telling, which could result in more scientists being selected who mainly master the right proposal prose. In addition, if there is too much focus on aspects like development potential or change agent criteria, there could be a risk that in the end, mostly strongly application-orientated research is funded to the detriment of basic research, which is primarily interest-orientated rather than application-orientated., as was also critically noted in the discussion among the jury members. Future research should therefore focus more on the analysis of change agent characteristics among scientists, as there is a considerable need of empirical and conceptual work in this area.

REFERENCES

Alberts, B.; Kirschner, M. W.; Tilghman, S.; Varmus, H. (2014) Rescuing US biomedical research from its systemic flaws. In *Proceedings of the National Academy of Sciences* 111/16: 5773–7.

Beckhard, R. (1969) *Organizational development: strategies and methods*. Boston: Addison - Wesley.

Belcher, B. M.; Rasmussen, K. E.; Kemshaw, M. R. et al. (2016) Defining and assessing research quality in transdisciplinary context, *Research Evaluation* 45/1: 1–17.

Bonn, N. A.; Bouter, L. (2023) Research Assessments Should Recognize Responsible Research Practices. Narrative Review of a Lively Debate and Promising Developments. In: Erick Valdés, Juan Alberto Lecaros (eds.) *Handbook of Bioethical Decision. Volume I: Decision at the Bench*, pp. 441–72, Cham: Springer International Publishing.

Boon, W.; Edler, J. (2018) Demand, Challenges, and Innovation. Making Sense of New Trends in Innovation Policy*, Science and Public Policy* 45/4: 435–47.

Bornmann, L. (2013) What is social impact of research and how can it be assessed? A literature survey. In *Journal of the American Society for Information Science and Technology* 64/2: 217–33.

Caldwell, R. (2005) Things fall apart? Discourses on agency and change in organizations. In *Human Relations* 58/1: 83–114.

Curry, S.; Rijcke, S. de; Hatch, A.; et al. (2020) The changing role of funders in responsible research assessment: progress, obstacles and the way ahead. Working Paper. Research on Research Institute (RoRI).

D'Este, P.; Ramos-Vielba, I.; Woolley, R.; Amara, N. (2018) How do researchers generate scientific and societal impacts? Toward an analytical and operational framework, *Science and Public Policy* 45/6: 752–63.

Donovan, C.; Hanney, S. (2011) The 'Payback Framework' Explained, *Research Evaluation* 20: 181–3.

DORA (2021) *Reimaging academic assessment: Stories of innovation and change*. Available online at https://sfdora.org/dora-case- studies/.

Edwards, M. A.; Roy, S. (2017) Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of perverse Incentives and Hypercompetition, *Environmental Engineering Science* 31/1: 51–61.

EFI - Commission of Experts for Research and Innovation (2024): Report on Research, Innovation and Technological Performance in Germany 2024, Berlin: EFI.

Ferretti, F.; Guimaraes Pereira, D. V.; Hardemann, S. (2018) Research excellence indicators: time to reimage the 'making of', *Science and Public Policy* 45/5: 731–41.

Gingras, Yves (2014) *Bibliometrics and Research Evaluation. Uses and abuses.* Cambridge: MIT Press.

Hagstrom, Warren O. (1975) *The Scientific Community*. Carbondale: Southern Illinois University Press.

Hicks, D.; Wouters, P.; Waltman, L.; Rijcke, S. d.; Rafols, I. (2015) The Leiden manifesto for research metrics, *Nature* 520, 429–31.

Hutton, David W. (1994) *The Change Agents' Handbook*. Milwaukee: ASQC/Quality Press.

Joly, P. B.; Gaunand, A.; Colinet, L.; Larédo, P.; et al. (2015) ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization, *Research Evaluation* 24/4: 440–53.

Kawulich, B. B. (2005) Participant Observation as a Data Collection Method, *Forum Qualitative Sozialforschung: Forum Qualitative Social Research* 6/2.

Kok, M. O.; Schuit, A. J. (2012) Contribution mapping: A method for mapping the contribution of research to enhance its impact, *Health research policy and systems* 10/1: 1–16.

Kraemer-Mbula, Erika; Tijssen, Robert; Wallace, Matthew L.; McLean, Robert (2020) *Transforming Research Excellence. New Ideas from the Global South*. Cape Town: African Minds.

Lauronen, J.-P. (2020) The dilemmas and uncertainties in assessing the societal impact of research, *Science and Public Policy* 47/2.

Lunenburg, F. C. (2010) Managing Change. The Role of the Change Agent, *International Journal of Management, Business, and Administration* 13/1: 1–6.

Matt, M.; Gaunand, A.; Joly, P.-B.; Colinet, L. (2017) Opening the Black Box of Impact - Ideal-Type Impact Pathways in a Public Agricultural Research Organization, *Research Policy* 46: 207–18.

Mazzucato, M. (2018) Mission-oriented innovation policies: challenges and opportunities, *Industrial and corporate change* 27/5: 803–15.

Merton, R. K. (1957) Priorities in scientific discovery: A chapter in the sociology of science, *American Sociology Review* 22: 635–59.

Moed, H. F. (2020) Appropriate Use of Metrics in Research Assessment of Autonomous Academic Institutions, *Scholarly Assessment Reports* 2/1.

Moher, D.; Naudet, F.; Cristea, I. A.; Miedema, F.; et al. (2018) Assessing scientists for hiring, promotion, and tenure, *PLoS Biology* 16/3.

Moran, J.; Brightman, B. (2000) Leading organizational change, *Journal of Workplace Learning* 12: 66–74.

Ottaway, R. N. (1983) The Change Agent: A Taxanomy in Relation to the Change Process, *Human Relations* 36/4: 361–92.

Reale, E.; Avramov, D.; Canhial, K.; Donovan, C.; et al. (2017) A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research, *Research Evaluation* 27/4: 298–308.

Rogers, E. M.; Shoemaker, F. F. (1971) *Communications of innovations: A cross-cultural appraoch*. New York: The Free Press.

Smit, J. P.; Hessels, L. K. (2021) The production of scientific and societal value in research evaluation: A review of societal impact assessment methods, *Research Evaluation* 30/3: 323–35.

Spaapen, J.; van Drooge, L. (2011) Introducing 'Productive Interactions' in Social Impact Assessment, *Research Assessment* 20/3: 211–18.

Tijssen, R.; Winnink, J. (2022) Global and Local Research Excellence in Afrika: New Performance Assessment and Funding, *Science, Technology and Society* 27/3: 368–87.

UN General Assembly (2015) *Transforming our world. The 2030 Agenda for Sustainable Development*. 21 October 2015. A/RES/70/1. New York.

Wilsdon, J.; Allen, L.; Belfiore, E.; Campbell, P.; et al. (2015) *The metric tide: Report of the independent review of the role of metrics in research assessment* *and management*. Available online at https://re.ukri.org/documents/hefce-documents/metric-tide- 2015- pdf/.

Zuckerman, H.; Merton, R. K. (1971) Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system, *Minerva* 9/1: 66–100.

AUTHOR

HENDRIK BERGHÄUSER
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Straße 48
D-76139 Karlsruhe
E-Mail: Hendrik.Berghaeuser@isi.fraunhofer.de
ORCID: 0000-0002-3884-3404

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e4, 1-25
DOI: 10.22163/fteval.2025.696

INDICATORS AND METRICS IN SSH RESEARCH:

HOW SCHOLARS VALUE PUBLICATION PRACTICES IN THE FACE OF EPISTEMIC CAPITALISM

FLORIAN BAYER
DOI: 10.22163/FTEVAL.2025.696

ABSTRACT

Situated at the level of individual researchers, this paper extends empirical research on indicators, metrics and other forms of quantification in everyday (research) practice to the social sciences and humanities (SSH). I draw on 46 qualitative in-depth interviews with senior researchers and early career researchers in history, political science and area studies to trace how SSH scholars value publication practices and outputs. Building on approaches from valuation studies (Helgesson & Muniesa, 2013) three registers of valuing (Heuts & Mol, 2013) are identified: an epistemic, a reputational and an institutional register of valuing publication practices. By exploring overlaps and relations between the three registers, folded valuations (Helgesson, 2016) and their mobilization in different valuation constellations (Waibel et al., 2021), I investigate the role of indicators, metrics and other forms of quantification. Results show that epistemic practice in SSH fields is permeated with indicator use. On the one hand indicators and metrics are a means to denote relevance across the three registers in everyday practice. On the other hand, output-oriented research cultures rely on socio-technical practices of quantification to promote "research quality" and "research excellence", as such practices are closely related to the epistemic and organizational practices that constitute epistemic capitalism (Fochler, 2016). The paper concludes with implications for reforms of research assessment (CoARA, 2022) and how this relates to reforming contemporary research cultures more generally.

Keywords: indicator use, social sciences and humanities (SSH), valuation studies, research quality, CoARA, epistemic capitalism

INTRODUCTION

For more than two decades, scholars have observed and scrutinized auditing and evaluation practices in Western societies (Power, 1997; Dahler-Larsen, 2012), including their role in the governance of scientific systems and the resulting implications for academic subjectivities (Strathern, 2000; Shore, 2008). The rise of evaluation and assessment procedures in science has been accompanied by the definition of performance targets and benchmarks as well as an increasing output orientation of funding systems (Hicks, 2012; Gläser & Laudel, 2016) and the emergence of new methods and practices of quantification. While the quantification of scientific output with a special focus on scholarly communication originates in the late 1950s and early 1960s (Garfield, 1955, 1964), bibliometric indicators and metrics gained momentum in scientific management around the year 2000, when computerized information and communication technologies (ICTs) allowed for their increasing dispersal and application (Burrows, 2012). In the meantime, growing concerns about the use of quantitative indicators to evaluate scientific performance have been voiced repeatedly, calling for the assessment of *"research on its own merits rather than on the basis of the journal in which the research is published"* (American Society for Cell Biology (ASCB), 2012), for more responsible use of publication-based metrics in assessment procedures (Hicks et al., 2015; Wilsdon et al., 2015) as well as for more general reforms of research assessment with a focus on qualitative assessment, supported by the responsible use of quantitative indicators (CoARA, 2022). At the same time scholarship has started to empirically investigate the role of indicators and metrics in research practice, observing their use in and effects on epistemic practices at the level of research groups and individual researchers (de Rijcke et al., 2016).

This paper contributes to these lines of work by presenting empirical results on the role of indicators and metrics in the social sciences and humanities (SSH). Situated at the level of individual researchers, I explore how SSH scholars perceive and reflect the conduct of research and its outputs in everyday practice. Developing an empirical approach based on valuation studies (Helgesson & Muniesa, 2013) I draw on qualitative in-depth interviews with senior researchers and early career researchers in history, political science and area studies to analyze how SSH researchers value publication practices and outputs in everyday practice and investigate the role of indicators, metrics and other forms of quantification in this regard.

The resulting empirical observations help to unpack opaque notions such as “research quality” and “research excellence”, offer new insights for ongoing debates and efforts to reform research assessment (CoARA, 2022), and contribute to the analysis and critique of output-oriented, competitive academic research cultures in terms of epistemic capitalism, namely a cultural configuration of organizing and practicing research based on the entrepreneurial management of careers, publications and grant portfolios (Fochler, 2016).

EMPIRICAL APPROACH, METHODS AND ANALYSIS

Building on empirical work conducted as part of a Ph.D. project in science & technology studies (STS), this paper extends empirical research on indicators and metrics in (research) practice (de Rijcke et al., 2016) to SSH fields. Situated at the level of individual researchers, I aim at identifying how SSH scholars mobilize indicators, metrics and other forms of quantification in everyday (research) practice based on a case study at the University of Vienna.

Throughout the last years Austrian science policy has increased the utilization of performance-based indicators in funding universities, assigning 20 percent of the budget in teaching and research based on at least one competitive indicator in each domain.[1] Within this framework research performance is addressed based on grant income and the number of employed PhD students per academic year.[2] In resulting budget negotiations with the Federal Ministry for Education, Science and Research universities commit to meet benchmarks for revenues from R&D projects and for employing Ph.D. students. The national funding model is reflected within universities, e.g. in funding its different faculties, the University of Vienna is calculating a portion of their budgets based on these performance-based indicators. However, scholars with management experience report that the University of Vienna is not defining,

1 Bundesgesetz, mit dem das Universitätsgesetz 2002 geändert wird, Bundesgesetzblatt, BGBl. I Nr. 8/2018, published 4 April 2018. URL: https://www.ris.bka.gv.at/eli/bgbl/I/2018/8 (Accessed 07.03.2025).

2 Erläuterungen, Regierungsvorlage: Bundesgesetz, mit dem das Universitätsgesetz 2002 geändert wird, 31.01.2018. URL: https://www.parlament.gv.at/dokument/XXVI/I/10/fname_679289.pdf (Accessed: 07.03.2025); Verordnung des Bundesministers für Bildung, Wissenschaft und Forschung über die Umsetzung der kapazitätsorientierten, studierendenbezogenen Universitätsfinanzierung, BGBl. II Nr. 202/2018, published 4 August 2018. URL: https://www.ris.bka.gv.at/eli/bgbl/II/2018/202 (Accessed 07.03.2025).

assessing and economically incentivizing quantified benchmarks at the level of departments, research groups or individual scholars. Instead, publication output and grant income are simply discussed on a regular basis between the rectorate and faculties and are thoroughly assessed in a seven-year evaluation cycle of faculties. Furthermore, publication output and grant income play a central role in qualification agreements for tenure track positions and in the evaluation of newly hired professors, typically after five years. In relation to international examples such as Poland, Australia or the UK, the degree of institutional metricization and economization (Kulczycki, 2023) can therefore be described as low.

Indicator-related incentives have been observed to trickle down to institutional practices targeting individual SSH scholars in the Norwegian case, which is also characterized by loose coupling (Aagaard, 2015). Since Aagaard observed a high degree of informal indicator use in the SSH, especially in career-related governance practices such as salary negotiations, it seemed promising to empirically study indicator use in relation to career trajectories more generally. More recently scholarship on the impact of quantification on the social sciences in the United Kingdom (Pardo-Guerra, 2022) as well as research on strategic decision making in relation to the academic labor market in general (Gläser & Laudel, 2015; Laudel & Bielick, 2018, 2019) has convincingly utilized a concept of academic careers based on the works of the Chicago School of Sociology.[3] Such a conceptual focus on career trajectories enables us to empirically trace how SSH scholars negotiate and navigate the nexus between individual and collective research practices, between the academic self and the community it is embedded in, as well as between individual action and structural requirements and expectations.

Mobilizing disciplinary differences as a comparative lens, the study was confined to one research institution. With about 10,700 employees and more than 85,200 students, organized in 20 faculties and centers, the University of Vienna is Austria's biggest institution for education and research, where SSH fields are well represented.[4] The fields history, political science and area studies were chosen for empirical analysis, because each of them is

3 Transcending earlier narrow conceptualizations of careers as a sequence of jobs or professional statuses, by more broadly and firmly grounding the concept in various social settings, authors like Hughes and Goffmann had stressed the capacity of this notion to conceptualize the nexus between the personal and the collective, individual action and social structure (Hughes cf. Barley, 1989, p. 46; Goffman, 1961, pp. 127–128).

4 Universität Wien, Zahlen, Daten & Broschüren. URL: https://www.univie.ac.at/ueber-uns/auf-einen-blick/zahlen-daten-broschueren/ (Accessed: 07.03.2025).

institutionalized in a different faculty. History and political science are both traditional disciplines, usually represented at universities covering the SSH. Both have strong disciplinary traditions and institutions within Austria and at a global level. Area studies is an explicitly interdisciplinary research field, bringing together different perspectives and traditions – ranging from social and cultural anthropology, economics, political science to cultural studies, linguistics and literary studies – based on a regional focus, e.g. japanese studies or chinese studies.

Qualitative in-depth interviews with senior researchers (SR) were conducted to explore the characteristics of these research fields by obtaining the perspectives of scholars, who have led and shaped departments, who have experience in different types of hiring processes, and who have experience in leading and mentoring pre- and postdoctoral researchers. The next step was to focus the interviews on early career researchers (ECR) at different stages of their careers, starting with late predoctoral researchers who had had their first experiences with academic publishing, researchers in postdoctoral positions, as well as more advanced and established scholars, who were about to receive or had recently received tenure.

In total 46 interviews were conducted with 44 researchers (23 male, 21 female) between September 2018 and December 2023. Starting with historians in the first wave (9/2018 – 4/2019), I moved on to interview SR across the three research fields in a second wave (11/2019 – 2/2020). Building on the results of waves 1 and 2, I conducted a third wave of interviews (1/2023 – 12/2023) with ECRs in all three fields. All interviewees gave written and oral consent. Conversations – ranging from one hour up to three hours, 40 in German language, 6 in English – were recorded, transcribed and imported to Atlas.ti.

The first round of coding was conducted in parallel with data collection, following the principles of Grounded Theory (Strauss & Corbin, 1998; Bryant & Charmaz, 2010), so that emerging categories and themes could inform ongoing sampling and data collection based on a zig-zag approach (Rivas, 2018). The coding process was restarted and restructured several times during data collection waves 1 and 2. These repeated iterations and adaptations were related to moving away from traditional Grounded Theory towards Abductive Analysis (Timmermans & Tavory, 2012; Tavory & Timmermans, 2014), which offered a good way into structuring the material and developing a coding scheme based on the process of alternative casing (Tavory & Timmermans, 2014, pp. 58–61).

Analyzing SSH scholars' reflections on research and publication practices when reporting about their personal experiences throughout their careers –

e.g. in applying for academic positions, the ways they had (not) planned or prepared for certain career steps and the role they attribute to different research and publication practices in their everyday working routines – I turned to valuation studies to make sense of the material. In particular, Heuts and Mol's (2013) approach to valuation as a practice enabled me to further sensitize quality judgments as a central aspect of research practice. Moving away from looking into certain qualities of things, they are *"foregrounding 'valuing'"* as an activity by identifying and closely looking into different *"registers of valuing"*. These registers *"indicate a shared relevance, while what is or isn't good in relation to this relevance may differ from one situation to another"* (Heuts & Mol, 2013, p. 129). Speaking of "*valuing*" also highlights how assessment, judgment, valuation, evaluation, improvement are practices that "*slide over into each other*" (p. 130). Coding the material of waves 1 and 2, three registers of valuing publication practices emerged, transposing Gläser & Laudel's (2015) model of the simultaneous pursuit of three different careers in academia: an epistemic register, a reputational register, and an institutional register. Throughout the coding process emerging codes and categories were grouped along research and publication practices and the three registers alike.

For the final wave of interviews with ECRs, a new interview guideline was developed based on insights from previous waves. ECRs were invited to report and reflect their academic career so far, before they were asked to choose the five most important research outputs on their CV and recall their practical histories from the very beginning. The guideline was accompanied by a set of cards to actively confront the interviewees with quotes and insights from previous interviews along the emerging codes and categories of the coding process. The cards were used to jump-start reflections, to mobilize disciplinary differences, and to offer orientation on what aspects and levels of abstraction to focus on, enabling the interviewees to relate and position their own sense-making and experiences to that of others.[5]

5 Card-based methods have been used to render non-debatable issues debatable in focus group discussions on nano-technologies (Felt et al., 2019), and to enable researchers to reflexively discuss matters of responsible research and innovation (Felt et al., 2018) and research integrity (Felt & Frantz, 2022).

VALUING PUBLICATION PRACTICES IN THE SSH

This empirical approach based on valuation studies enables a new perspective on how SSH scholars value publication practices in everyday contexts. Due to the output orientation of contemporary academia, publications are no longer just a means of scholarly communication. They have become essential for signaling "research quality" or "research excellence" in a variety of settings and contexts. As a result, all sorts of considerations in epistemic practice are related to publication practices. The analysis and description of the registers of valuing publication practices provide sensitivity and orientation in this regard. Heuts and Mol (2013, p. 129) emphasize how valuation practices are messy and complex. The registers of valuing drawn upon overlap and are sometimes in tension with each other. In a similar vein, Helgesson (2016) has suggested investigating the ways in which multiple valuations and different valuation practices are folded into each other:

> *"Looking into the nooks and crannies of a conglomeration of interrelated valuation practices further provides a glimpse of a politics beyond the singular valuation practice; this is the politics of how valuation practices are folded on to one another, and how these folds are characterized"* (Helgesson, 2016, pp. 100–101)

Other scholars highlight that not only the multiplicity of valuing as a practice as such needs to be accounted for, but also the multiple contexts in which valuations take place. Waibel and colleagues speak of valuation constellations to reflect the positions and relations between the valuee, the valuator, and the audience, and to include the role of valuation rules and infrastructures (Waibel et al., 2021).

The results presented and discussed here are based on the analysis of interview material concerned with research outputs from the perspective of their production process. As indicated above, this focus originates in an interview design focusing on relations between publication practices and career trajectories. Instead of studying moments of assessment or asking researchers what constitutes a good monograph or journal article, scholars were encouraged to recall the practical histories of their own publications: e.g. starting with the context of the respective research, the associated research agenda, project idea or research questions, to reflect the research process, to recall initial ideas for manuscripts, up to the drafting of the manuscript and the review and production process of publications. Analyzing how SSH scholars value research and publication practices from this perspective reflects overlaps and relations between the identified registers of valuing publication practices, considers specific foldings of different kinds of valuations, and how they are related to experienced, imagined, and anticipated valuation constellations.

THREE REGISTERS OF VALUING PUBLICATION PRACTICES

In a first step empirical analysis uncovered how SSH scholars draw on three registers of valuing publication practices: an epistemic register denoting qualities in relation to processes of knowledge creation and production; a reputational register expressing relevance based on the capacity to contribute to reputation building in the scientific community; and an institutional register denoting quality in relation to meeting demands and expectations from institutional settings and contexts. Throughout this section I will present these registers in more detail by showcasing how SSH scholars value publishing as a research practice in general terms; then I will move on to illustrate the mobilization of the three registers with regard to monographs and journal articles, which are described as the two most relevant output types in relation to career trajectories.

The three registers of valuing publication practices are accessible not only in relation to individual research outputs or types of outputs, but also in relation to publishing as a research practice itself. Scholars' reflections on their most important research achievements and outputs often came along with valuing publishing in general. Across the three research fields there was no clear difference between producing research results and communicating them through publication practices. In many instances scholars describe the writing process as an integral part of "doing research".

When explaining their motivations and aims for publishing, researchers indicate epistemic relevance of publishing as a research practice with regard to several dimensions: Time spent on publishing is described as valuable, if it enables the development or pursuit of research interests embedded in or relating to an overall research agenda. Another central motivation for publishing is contributing to certain debates/bodies of knowledge. Along this dimension, publishing is not an end in itself. Rather it requires to have something substantial to say to begin with: *"[...] it is not very meaningful to simply publish boring stuff or more of the same"* (I23-1, P97). In the words of a SR: *"You have to go for the questions where you can actually make a contribution."* (I16, P77).

Even though the focus is on academic debates, relevance for the public and societal concerns are also mobilized to denote epistemic relevance. Finally, there is also an individual dimension in the epistemic register: publishing is denoted as "good" when enabling personal learning in epistemic terms: "[With regard to publishing] *you also have to understand: you are not the same person at the beginning and at the end of the predoc phase* [...] *over the years your experience grows, your competencies expand"* (I25, P216).

In the reputational register, publishing is denoted as an important research practice based on its capacity to contribute to reputation building. In most general terms this means that scholars think publishing is important because it enables and facilitates the recognition of the self as an academic, researcher, and scholar in the first place. A SR recalling career advice by their own mentor expresses this dimension most vividly: *"... because* [mentor] *had told me: That's not possible, not publishing during your Ph.D. That is terrible, then you are not a person"* (I17, P33). Similarly, publishing is denoted as relevant, when it contributes to the development of a scholarly profile and to the shaping and formation of one's academic personality. Publishing practices are constructed as "good" or "bad" based on the different kinds of positioning work they enable, including the signaling of academic maturity or facilitating topic-name recognition.

The institutional register allows researchers to frame the relevance of publishing in relation to demands and expectations originating from institutions. In scholars' narrations on the relevance of publishing as a research practice, they reflect institutional settings, contexts and practices as mediators for the important role attributed to publishing. Scholars refer to, draw on and mobilize regulations, policies and practices at the level of departments, faculties or the university (e.g. regulations for paper-based dissertations or qualification agreements for tenure track positions), funding institutions (from application guidelines to hearsay on review processes and decision making), and the academic job market (from experiences with applications, job talks to participation on hiring committees) in order to express the importance of publishing. This way institutional settings, contexts and practices do not only exert pressure on scholars by articulating and enacting demands and expectations; at the same time all of these moments can be used as argumentative resources to denote publishing as an important, if not *the* most important research practice.

The institutional register also offers a whole repertoire of motivations and goals for publishing, that scholars can draw on to denote relevance. The most dominant one would be the overall goal of developing a publication portfolio that ticks the relevant boxes in institutional settings and contexts. In the words of an ECR:

> *"I really do not know for sure if say like the-, for the department when they hire a new professor if they like, tick boxes how many or in what category it's published. But I feel that these are all altogether to be considered. Yeah (..) and I think for the younger scholars, it's perhaps also an effective way of like presenting yourself publicly or to the universities, of course."* (I28, P108)

Other dimensions in the institutional register comprise the capacity to demonstrate experience, performance, and productivity through publishing, while publishing itself is also described as the core task of being an academic.

VALUING MONOGRAPHS

When recalling the practical histories of individual publications, interviewees repeatedly describe publishing monographs as enabling specific forms of thought and research processes, due to the relative absence of limitations in form and space. In contrast to other formats, monographs offer the most freedom for authors to choose the structure and form of presentation, usually there is also no word limit. For this reason, researchers value the publication of monographs because of the freedom to present their research as they see fit and as appropriate with respect to the object of their research, instead of following editorial guidelines. In these valuations the writing process is described as an integral part of the research process. The practice of publishing a monograph is valued as an act of doing research rather than an act of communicating research results produced in an earlier, separate phase of the research process. In these terms publishing monographs is denoted as good, because it enables specific epistemic practices, such as thoroughly discussing and appropriating theories, or working with and integrating multiple approaches. As an ECR in history recalls one mentor explaining: *"Well you can somehow argue anything in an article, but you can only see whether an argument really works in a book"* (I45, P167).

Drawing on the reputational register, scholars value publishing monographs as the traditional format or research output that comes with high status gains and prestige. Scholars in history and area studies repeatedly argue that monographs are still important in their fields. So even though the status of the monograph is seen as challenged, it is at the same time described as the gold standard. This resonates in ECRs' reflections on how doing a paper-based dissertation was only a theoretical option, which many of them did not even know about, while others were explicitly advised against it by their supervisors. In area studies the status of the monograph is challenged based on researchers' multiple, transdisciplinary identities. Scholars with backgrounds in the social sciences tend to value publishing monographs as less important or as challenged, while scholars working in the tradition of the humanities emphasize its prestige. In political science the importance of the monograph is described differently depending on subfields and empirical orientation. Scholars with a focus on theory and qualitative research tend to value monographs as traditionally prestigious, while political scientists with a quantitative empirical orientation see a declining role of the monograph in reputation building.

Beyond being valued as the traditional format for reputation building, the monograph is valued for its capacity to represent a scholar's academic personality. Drawing on the reputational register, scholars denote publishing monographs as "good", because doing so forms researchers, as writing a research monograph shapes one's personality as an academic. Scholars do not only express pride in having written books themselves. They also express admiration and respect for others who did so. In this sense, a mythical dimension is ascribed to the monograph:

> *"*[…] *it's difficult to write a book* […]*, therefore it's part of a maturing process and so on.* […] *I think it's a lot symbolic and not necessarily broken down into individual criteria as to why monographs and books are so important."* (I22, P77)

Finally, scholars also draw on the institutional register to denote publishing monographs as relevant and important. Field specific differences in these valuations correspond to those in the reputational register: While historians describe the publication of monographs as a necessity in the competition for academic jobs, political scientists with a quantitative empirical orientation argue this is not the case at all. These valuations in the institutional register also find an expression with regard to tenure: A typical academic career trajectory in history is reported to require the publication of two research monographs, one being the Ph.D. thesis and the second one for a tenured position. Scholars in area studies express the need to plan for at least one monograph in the long run to maintain a competitive profile ticking all the relevant boxes. Political scientists with a quantitative empirical orientation argue that monographs are no longer a prerequisite for tenure in their field. However, in certain sections of the academic job market, especially in the German speaking context, not having published a monograph at all might become an obstacle in the competition for full professorships.

VALUING JOURNAL ARTICLES

Valuations of publishing journal articles drawing on the epistemic register relate to the research and the writing process. In their most basic form such valuations revolve around the content, topic, argument or the empirical material or substance of the respective manuscript or article. Likewise, journal articles are denoted as good, by relating the aforementioned aspects to ongoing debates in the respective discipline, (sub)field or community. Regarding the writing process, the journal article is repeatedly described as a specific form or craft. Scholars express how writing and presenting results in the form of journal articles is very much shaped by formal guidelines, established or expected structures and the more general style and form of the journal, including the need to meet the imagined, anticipated, or in due course clearly expressed expectations of editors and reviewers. Learning and mastering the craft of conceptualizing, drafting, and getting journal articles accepted for publication is described as a key personal learning. Across the board, scholars vividly recounted and recalled how publishing their first article was an intense and exciting epistemic experience in exchange with mentors, editors and reviewers.

Valuations of individual journal articles drawing on the epistemic register typically go hand in hand with epistemic valuations of the journal the article was published in. Similar to individual articles, journals are valued in the epistemic register in terms of the discipline, (sub)field or community they are addressing. Correspondingly, the epistemic relevance of journals is expressed in terms of the debates, topics, and works published in them. Also, individual experiences with the editorial and review process are mobilized in the epistemic register of valuing journals. Even though scholars repeatedly express how astonishingly contingent review processes are, past experiences are mobilized as an argumentative resource to assess and express the quality of journals in epistemic terms.

With regard to reputation building, the folding of valuations concerning individual journal articles and their journals turns upside down, as reputational valuations mostly revolve around the outlets, rather than the individual articles themselves. This means that an individual journal article contributes to reputation building based on the status the outlet is ascribed in the discipline, (sub)field or community. In these terms the prestige of the journal is related to its audience: the most prestigious journals are described as broad and general flagship journals of disciplines and research fields (requiring also contributions of general and broad importance in epistemic terms). In contrast, contributions to more specialized subfield journals are ascribed less relevance, hence they contribute less to reputation building. Successfully publishing an article in a respective journal is not only part of positioning work with regard to the academic profile, it also comes with being recognized as a scholar in the field: *"With a text like the one I wrote* [...] *you are accepted as a participant in the debate, so to speak, or you can at least suggest that you want to have a say, let's put it that way, right?"* (I45, P59).

In the institutional register, the relevance of publishing journal articles is predominantly denoted by emphasizing the role of articles in getting jobs in academia. Journal articles are described as the decisive element on CVs when it comes to pursuing an academic career trajectory. Articles are valued as "good" because they enable scholars to distinguish themselves and to make it to the next round of hiring procedures for postdoctoral and tenure track positions, as well as full professorships. Also in the institutional register, the valuation of individual articles is folded into valuations of the publication venue. Across all three research fields, interviewees emphasize how publishing articles in internationally recognized, peer reviewed journals is key in pursuing an academic career. As a political scientist recalls the first accepted manuscript:

> *"How do I remember that? Well, of course with a lot of sweat and fear about what would come out of it, because I thought that would be ideal if it worked out.* [...] *I had submitted it somewhere else before and it was closely rejected* [...] *and it was also submitted very high.* [...] *I really, I think, revised it very, very fundamentally, changed very fundamental things* [...] *the paper itself is certainly one of the most polished and well thought-out things I have ever done, simply because so much time went into it.* [...] *I think it was simply because of the high stakes, that was the first publication where I thought to myself, this will really help me now for my future career. And I want to do anything but mess the whole thing up."* (I32, P197)

Publishing journal articles is also denoted as important with regard to funding applications, or expectations at the level of departments, faculties or the university. In the most general terms, the practice of producing peer reviewed journal articles is valued as important by describing it as the core task of being an academic.

THE ROLE OF INDICATORS, METRICS AND QUANTIFICATION

By delving into overlaps and relations of the three registers of valuing publication practices, we can disassemble vague singulars like "research quality" and "research excellence" into a broad variety of dimensions, notions and their context-specific mobilization. We can strip notions like "quality" of their opaqueness, by breaking them down into different kinds of qualities, namely the multiple and heterogeneous characteristics of publication practices and outputs. In doing so we can also trace how citation-based indicators and metrics are mobilized together with other forms of quantification in valuing publication practices.

First, the empirical analysis presented illustrates how SSH scholars make sense of their own publication practices and outputs in relation to varying concrete, imagined, and anticipated valuation constellations (Waibel et al., 2021): Publications are a means of producing, presenting and communicating research and its results, they are a means for recognition in the community, and they are a means in the competition for institutional reward. We can observe how multiple aspects and dimensions are conflated in practice, by studying overlaps and relations between the three registers in different valuation constellations. Usually, we assume these to be neatly separated, following a linear process of abstraction: research quality is based on solid and careful epistemic practice, which is consequently recognized, criticized and acknowledged through peer review and finally rewarded by institutions. Analyzing how SSH researchers value publication practices clearly indicates that the situation is more complex, as all of these dimensions come to matter already in the research process, very often simultaneously.

This also applies to the mobilization of indicators and metrics, which become relevant not only through institutional assessment procedures and evaluations. Indicator use is spread across all three registers of valuing as well as all sorts of practices and contexts, as is exemplified by an ECR elaborating on the role of citation-based information in literature research:

> *"Generally speaking, citation numbers. This is one of the easiest ways I can look at whether or not this is a good journal. It doesn't (.) it's this peer review, DOI number kind of thing. Yeah, that's great. That's the very basic that you want to have but then you need to look at citation numbers. How often has this been cited? And then you start looking at that.* [...] *I look at certain articles and then*

> *it tells me these are journal articles that are constantly being used.* [...] *This tells me this is an important and incredibly important journal in* [subfield] *that has to be used. I had to look at it, not just at this issue. So, I actually go into the database of that journal and then look at other issues they have come up with and then see what they're doing."* (I29, P47)

Reflections of this kind demonstrate how citation-based indicators and metrics have permeated epistemic practice. Scholars use and mobilize them as one piece of information among others in finding, assessing and appropriating the work of other scholars, as much as in decision-making about their own publication practices. They are mobilized to denote relevance in the epistemic register, in the reputational register and in the institutional register alike. By reflecting relations between the three registers of valuing we can trace how different aspects and dimensions are folded into one another (Helgesson, 2016) to construct meaning and relevance. This kind of analysis clearly indicates, how status and prestige – often combined with citation-based indicators and metrics – are not only mobilized in institutional assessment procedures, but play a crucial role in making sense of, strategically planning and carrying out epistemic practice. This has important implications: The use of indicators in epistemic practice must not itself be conceptualized as necessarily inappropriate, nor as the result of external pressures exerted by evaluation procedures. Rather, these observations suggest that *"thinking with indicators* [...] *inform*[s] *research* as *it is being conceived and conducted"* (Müller & de Rijcke, 2017, p. 161). Following Dahler-Larsen (2014), we might speak of constitutive effects of indicator use in epistemic practice.

Second, the empirical analysis of how the three different registers are folded into one another facilitates a better understanding of how SSH scholars plan, structure and carry out epistemic practice in output-oriented academic research cultures. Doing so enables the observation and articulation of specific dynamics that are brought about or mediated by competition. In everyday practice SSH scholars anticipate competitive editorial and review processes for limited publication space with journals and publishers. Similarly, they anticipate how their publication output will be valued by hiring committees, which are assessing large numbers of applicants trying to identify the best candidates that deserve a closer look in the second round. This kind of competition fosters abstractions to enable the comparison of candidates, e.g. by focusing on a limited spectrum of publication venues. As a SR points out in relation to hiring for postdoctoral positions:

"I can't read the entire oeuvre of 60 people [...] *on the assumption that there are golden nuggets in some papers, yes. So, you need criteria to go by and, of course, ranked journals are an important indicator. They've managed to jump over this hurdle, yes, that's no guarantee that it's* [...] *great research* [...] *in the sense that it revolutionizes the world or really brings something fundamentally new to the discipline or reorganizes the subfield or anything, but it is the criterion that there is at least solid, solid work that is good enough to get into the respective journal."* (I16, P117)

For individual scholars the strong emphasis on output orientation and competition in contemporary research cultures implies that pursuing a career in academia resembles a quest for return on investment by managing publication portfolios (Fochler, 2016; Rushforth et al., 2019). This is reflected in scholars' strategic decision-making on what publication practices to spend time on. An ECR recalling their strategy for developing a competitive portfolio illustrates how the different registers are mobilized in anticipation of future hiring committees:

"I knew at that time, and I know now that publishing in Q1 journals, it's valued much more than publishing in lower-ranked journals. And especially respected journals, sometimes even if it's not Q1, but it's a respected journal. I don't know whether it is true, to be honest with you, but this is also something that my supervisor had told me at that time. I think it's more important to have one good publication in a very good journal rather than having many in low-ranked journals or not respected journals, because the committee will not look at them anyway. It will be a bit of a waste of time. It's also good to show that you can publish on different outlets and talk to different audiences. But for me, the respect of the journal and the Q tire was among the most important decisions besides the topic fitting." (I35, P57)

Both examples demonstrate how output-oriented, competitive valuation constellations for positions or resources depend on abstractions in order to compare profiles and portfolios of individual scholars, even if procedures are based on peer review and are carried out with rigor and dedication to the highest standards. In other words, competition based on publication portfolios takes part in bringing about abstract goals and criteria, emphasizing the reputational and institutional register rather than the epistemic register. Also in this regard indicators and metrics serve as means to denote relevance. In assessing publication portfolios indicators are frequently mobilized to value publication venues, which is not necessarily inappropriate, as they are not used to signify immediate epistemic relevance for individual articles. In many cases they serve as a means in venue-related reputational valuations. Scrutinizing overlaps and relations between the different registers of valuing illustrates that indicators and metrics play a crucial role in such valuation constellations, but the overall dynamics originate in the orchestration of output-oriented competition. Criticizing the presence and use of quantitative indicators might not be enough, because the phenomena at hand are dynamics resulting from output-oriented, competitive production and valuation.

This leads to a third observation. Studying overlaps and relations of the three registers of valuing publication practices enables more fine-grained and detailed investigations into the role of indicators, metrics and other sociotechnical practices of quantification in contemporary academic cultures. Fochler (2016) has conceptualized the entrepreneurial management of careers, publications and grant portfolios in terms of the more general cultural configuration of epistemic capitalism. Epistemic capitalism denotes *"a particular cultural dynamic in knowledge production"*, that is based on *"the accumulation of capital as worth made durable, through the act of doing research"* (p. 924). In this perspective, the examples above illustrate how SSH research is practiced and organized in specific forms and thereby subjected to this very particular dynamic. Indicators and metrics play a central role in enacting markets *"based* [...] *on the strategic competition for the best asset position in relation to others."* (p. 927). In that sense ECRs' strategic reflections on developing their publication portfolios illustrate how they are *"concerned with accumulating indicators of their own worth in terms of their future employability"* (p. 934).

By delving into the different registers of valuing publication practices in SSH research, we can better understand how markets are enacted and how scholars are compared to each other based on publication output. Again, indicators and metrics are not necessarily used in naive or inappropriate ways in these processes. Many SSH scholars have profound knowledge of indicators, metrics, and their problems. Consequently, they are also hesitant and careful in mobilizing them to assess and compare individual scholars based on publication portfolios. A good example is the following quote by a SR, specifying how it's not simply about more publications being better:

> *"I don't mean it like that.* [...] *(sighs) it's rather, averages* [...]*, if I have two, three, four people and keep all other factors stable, which is of course difficult, then, under these conditions, the person who has a higher-ranked publication or a longer list of publications would prevail."* (I22, P169)

Drawing on the reputational and the epistemic register, the interviewee goes on emphasizing how assessing the overall profile of the respective scholars and the content of the publications in relation to the job announcement are more important. Even though these “other factors” are given priority and point to the high relevance that is attributed to epistemic valuations, this kind of reflection demonstrates how output-oriented competition is enacted based on rather abstract comparisons. Through competitively assessing publication portfolios individual publications are related to each other in an abstract, commensurable from that requires quantitative terms to express difference. In other instances, quantitative comparisons are a means to construct equivalencies. E.g. a SR reflecting the role of habilitations in book form for full professorships: *“So when it comes to someone with ten peer-reviewed journals and someone with four plus habilitation, I would say: that’s seven. That’s also my argument, explicitly”* (I19, P167).

Both examples demonstrate how competition based on publication portfolios is enabling and facilitating quantitative reasoning in valuing publication practices, and that this kind of quantitative reasoning is not external or opposed to peer review. Of course, this does not imply that such kinds of valuations do not involve qualitative aspects or content at all. On the contrary, qualitative aspects usually precede or accompany these kinds of valuations, as researchers draw on the three registers of valuing publication practices to denote what kind of outputs are to be considered and to what extent in negotiating equivalencies and differences, i.e. to argue what should count at all and how much.

The three registers of valuing publication practices have offered an analytical perspective to scrutinize in more detail how scholars make use of citation-based quantitative infrastructures, also in contexts relying on peer review. This has also highlighted how abstraction and quantification do not only originate in indicator-related performance goals or assessment but are brought about by and deeply inscribed in epistemic and organizational practices of contemporary research cultures. Empirically studying the three registers of valuing publication practices, the folding of respective valuations, and their mobilization with regard to multiple valuation constellations illustrates how planning, doing and presenting research – also in SSH fields – need to take very specific forms in order to enable the production and accumulation of epistemic capital (Fochler, 2016).

CONCLUSION

Empirical analysis of how SSH scholars value publication practices and outputs in everyday practice has highlighted the diversity and multiplicity of how publishing and publication outputs are valued in different contexts and settings. The inquiry differentiated three registers of valuing publication practices and outputs, namely an epistemic, a reputational and an institutional register. Looking into these registers and their relations not only enabled a more detailed articulation and understanding of publication practices in SSH research. Doing so, helped to unpack valuations behind notoriously vague notions like "research quality" and "research excellence" in the context of output-oriented academic research cultures. We have observed how epistemic practice is imbued with all sorts of valuations regarding publication practices, which includes the mobilization of indicators and metrics across the different registers. We should be hesitant to deplore valuations of this kind in advocating for purely epistemic valuation and judgment. Because research and scholarship are social practices organized based on recognition and reward (Stephan, 2012, pp. 17–34), epistemic practice cannot be imagined as a sphere prior to or isolated from these issues. Since *"thinking with indicators"* (Müller & de Rijcke, 2017) has also permeated epistemic practices in SSH fields, we need to examine in which cases, at which moments, and to what extent this is the case, rather than assuming that it is the result of assessment and evaluation procedures alone. We can do so by empirically tracing how valuations mobilizing the identified registers are folded into one another in specific settings and contexts, and how different valuation constellations and their anticipation shape epistemic practices.

However, the results presented also have implications for ongoing efforts to reform research assessment. For instance, the Coalition for the Advancement of Research Assessment is calling for more responsible assessment by focusing procedures "*primarily on qualitative evaluation for which peer review is central, supported by responsible use of quantitative indicators*" (CoARA, 2022, p. 5). The agreement suggests to *"move towards research assessment criteria that focus primarily on quality, while recognizing that responsible use of quantitative indicators can support assessment"* (ibid.). This recommendation implicitly juxtaposes peer review as a procedure to assess quality and the use of quantitative indicators, neglecting the role of the latter in valuations constructing and defining "research quality" in everyday practice. The empirical results presented in this paper show how researchers make context-specific use of quantitative indicators in valuing publication practices across the three identified registers of valuing. In other words, quantitative indicators have become one means among many for constructing and articulating "research quality" in everyday research practice as well as in peer review. In light of this observation, it seems important to emphasize that the focus on quality through peer review does not in itself contradict the use of quantitative indicators to construct and assess "research quality".

This is closely related to a second implication regarding CoARA's core commitment #3, which is calling to abandon inappropriate forms of indicator use such as *"assessing outputs based on metrics relating to publication venue, format or language"* (p. 6) Again, the empirical results show how publication venues are an important means of valuing individual journal articles in many settings and contexts. Looking at these valuations in more detail, we have explored how the mobilization of indicators and metrics, as well as status and prestige, are not necessarily considered inappropriate either. Scholars express awareness of the limits of such valuations and are carefully folding the valuation of articles and outlets into each other. If the complexities and multiplicities of such valuations are not adequately addressed and acknowledged, efforts to reform assessment procedures risk failing, as actors with different positions are unlikely to find common ground to discuss what is considered (in)appropriate or (il)legitimate and why. Empirically scrutinizing how publication outputs are valued based on the three registers of valuing publication practices offered clarity and orientation in this regard. Thus, the implementation of CoARA requires to create spaces for collective deliberations on what should constitute "research quality" and why, taking the observed complexity and multiplicities of valuations into account.

Finally, these observations also reflect how debates for reforms of research assessment are not only a central arena to discuss assessment and evaluation procedures. They are at the same time normative negotiations on what should constitute "research quality", "research excellence", and how research cultures should be organized. This way questioning indicator use in research assessment is implicitly tied to more general ideas and assumptions about whether and how research should be driven by recognition, reward and competition. The empirical results presented suggest, that indicator use is a central element in the construction and evaluation of "research quality", because contemporary research cultures are characterized by output-oriented competition and the entrepreneurial management of publication and grant portfolios (Rushforth et al., 2019). Doing research in the face of epistemic capitalism (Fochler, 2016) implies that the products of research have to take specific social forms to produce value, which in turn needs to be accumulated as epistemic capital, e.g. through successful career accumulation. Under these circumstances producing research results "of high quality" is not confined to carrying out and conducting research with rigor and based on the highest standards. Because of the overall output-orientation of research the production and accumulation of epistemic capital is dependent on repeated communication in proper form. I.e. publication outputs need to address the requirements of multiple valuation constellations by drawing on the epistemic, reputational and institutional register of valuing publication practices alike, in order to facilitate the production and accumulation of epistemic capital.

Considering this, the results presented above suggest that the role of quantification in valuing and assessing research outputs does not merely originate in the presence and use of bibliometric indicators. Rather, the organization of research in the form of epistemic capitalism implies to treat publications in abstract terms to negotiate "quality" and "excellence" in competitive settings. Comparing publication output in abstract forms requires expressing differences (and equivalences) in quantitative terms. As a result "more" is usually better, but only in the context of negotiating what counts as "excellent" or "high quality" and how much. Further inquiry along these lines can help us understand how this feeds into dynamics that are central to any capitalist configuration: *"an orientation towards attaining ever more capital* [...] *as an end in itself"* (Fochler, 2016, p. 929) and how this contradicts moving towards more heterodox and diverse economies of value and worth in academia (Fochler, forthcoming). The observations presented in this paper echo calls to question the organization of contemporary academia in the form of individualized competition (Kulczycki, 2023, pp. 188–191). Because abstraction and quantification originate in epistemic and organizational practices related to epistemic capitalism, actively challenging these phenomena would require establishing alternatives to output-oriented competition by seeking a new nexus between the production of individual researchers and the overall, collective achievements that constitute scholarship.

REFERENCES

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. Science and Public Policy, 42(5), 725–737. https://doi.org/10.1093/scipol/scu087

American Society for Cell Biology (ASCB). (2012). The San Francisco Declaration on Research Assessment (DORA). https://sfdora.org/

Bryant, A., & Charmaz, K. (2010). The SAGE Handbook of Grounded Theory: Paperback Edition. SAGE Publications.

Burrows, R. (2012). Living with the h-index? Metric assemblages in the contemporary academy. The Sociological Review, 60(2), 355–372. https://doi.org/10.1111/j.1467-954X.2012.02077.x

CoARA. (2022). Agreement on Reforming Research Assessment. Coalition for the Advancement of Research Assessment. https://coara.eu/agreement/the-agreement-full-text/

Dahler-Larsen, P. (2012). The Evaluation Society. Stanford University Press.

Dahler-Larsen, P. (2014). Constitutive Effects of Performance Indicators: Getting beyond unintended consequences. Public Management Review, 16(7), 969–986. https://doi.org/10.1080/14719037.2013.770058

de Rijcke, S. de, Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use - A literature review. Research Evaluation, 25(2), 161–169. https://doi.org/10.1093/reseval/rvv038

Felt, U., Fochler, M., & Sigl, L. (2018). IMAGINE RRI. A card-based method for reflecting on responsibility in life science research. Journal of Responsible Innovation, 5(2), 201–224. https://doi.org/10.1080/23299460.2018.1457402

Felt, U., & Frantz, F. (2022). RESPONSE_ABILITY A Card-Based Engagement Method to Support Researchers' Ability to Respond to Integrity Issues. Science and Engineering Ethics, 28(2), 14. https://doi.org/10.1007/s11948-022-00365-6

Felt, U., Schumann, S., & Schwarz-Plaschg, C. G. (2019). IMAGINE: A Card-Based Discussion Method. In P. Liamputtong (Ed.), Handbook of Research Methods in Health Social Sciences (pp. 1167–1182). Springer. https://doi.org/10.1007/978-981-10-5251-4_9

Fochler, M. (2016). Variants of Epistemic Capitalism. Science, Technology, & Human Values, 41(5), 922–948. https://doi.org/10.1177/0162243916652224

Fochler, M. (forthcoming). How is a good academic? Narrative responsibility in the academic value economy. In K. Asdal & L. Doganova (Eds.), The value economy. Mattering Press.

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, 122(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Garfield, E. (1964). "Science Citation Index" - A New Dimension in Indexing. Science, 144(3619), 649–654.

Gläser, J., & Laudel, G. (2015). The Three Careers of an Academic. 35, 1–40. http://www.laudel.info/wp-content/uploads/2015/12/35_2015discussion_paper_Nr_35_Glaeser_Laudel.pdf

Gläser, J., & Laudel, G. (2016). Governing Science: How Science Policy Shapes Research Content. European Journal of Sociology, 57(1), 117–168. https://doi.org/10.1017/S0003975616000047

Helgesson, C.-F. (2016). Folded Valuations? Valuation Studies, 4(2), Article 2. https://doi.org/10.3384/VS.2001-5992.164293

Helgesson, C.-F., & Muniesa, F. (2013). For What It's Worth: An Introduction to Valuation Studies. Valuation Studies, 1(1), Article 1. https://doi.org/10.3384/vs.2001-5992.13111

Heuts, F., & Mol, A. (2013). What Is a Good Tomato? A Case of Valuing in Practice. Valuation Studies, 1(2), 125–146. https://doi.org/10.3384/vs.2001-5992.1312125

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251–261. https://doi.org/10.1016/j.respol.2011.09.007

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. Nature, 520(7548), 429.

Kulczycki, E. (2023). The Evaluation Game: How Publication Metrics Shape Scholarly Communication. Cambridge University Press. https://doi.org/10.1017/9781009351218

Laudel, G., & Bielick, J. (2018). The Emergence of Individual Research Programs in the Early Career Phase of Academics. Science, Technology, & Human Values, 43(6), 972–1010. https://doi.org/10.1177/0162243918763100

Laudel, G., & Bielick, J. (2019). How do field-specific research practices affect mobility decisions of early career researchers? Research Policy, 48(9), 103800. https://doi.org/10.1016/j.respol.2019.05.009

Müller, R., & de Rijcke, S. (2017). Thinking with indicators. Exploring the epistemic impacts of academic performance indicators in the life sciences. Research Evaluation, 26(3), 157–168. https://doi.org/10.1093/reseval/rvx023

Pardo-Guerra, J. P. (2022). The Quantified Scholar: How Research Evaluations Transformed the British Social Sciences. In The Quantified Scholar. Columbia University Press. https://doi.org/10.7312/pard19780

Power, M. (1997). The Audit Society: Rituals of Verification. Oxford University Press.

Rivas, C. (2018). Finding themes in qualitative data. In C. Seale (Ed.), Researching Society and Culture (pp. 431–453). SAGE Publications.

Rushforth, A., Franssen, T., & de Rijcke, S. (2019). Portfolios of Worth: Capitalizing on Basic and Clinical Problems in Biomedical Research Groups. Science, Technology, & Human Values, 44(2), 209–236.

Shore, C. (2008). Audit culture and Illiberal governance: Universities and the politics of accountability. Anthropological Theory, 8(3), 278–298. https://doi.org/10.1177/1463499608093815

Stephan, P. E. (2012). How economics shapes science. Harvard Univ. Press.

Strathern, M. (2000). Audit Cultures: Anthropological Studies in Accountability, Ethics and the Academy. Routledge.

Strauss, A., & Corbin, J. (1998). Basics of Qualitative Research. SAGE Publications.

Tavory, I., & Timmermans, S. (2014). Abductive Analysis. University of Chicago Press.

Timmermans, S., & Tavory, I. (2012). Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis. Sociological Theory, 30(3), 167–186. https://doi.org/10.1177/0735275112457914

Waibel, D., Peetz, T., & Meier, F. (2021). Valuation Constellations. Valuation Studies, 8(1), 33–66. https://doi.org/10.3384/VS.2001-5992.2021.8.1.33-66

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. https://doi.org/10.13140/RG.2.1.4929.1363

ACKNOWLEDGEMENTS

Research for this paper received financial support from the Vienna University Library, I thank Christian Gumpenberger for continuous support. I thank my supervisor Ulrike Felt and the community at the Department of Science & Technology Studies for hosting my Ph.D. project and their continuous engagement with my work. Special thanks to Paul Trauttmansdorff for commenting the manuscript.

AUTHOR

FLORIAN BAYER
University of Vienna, Department of Science & Technology Studies,
Universitätsstraße 7, 1010 Wien
Email: florian.bayer@univie.ac.at
ORCID: 0009-0002-9298-0730

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e5, 1-19
DOI: 10.22163/fteval.2025.697

ADVANCING RESEARCH IMPACT EVALUATION IN THE DIGITAL ERA:

INSIGHTS FROM EU-FUNDED RARE DISEASE PROJECTS

IOANNA GRYPARI, SERGIO DI VIRGILIO, HARIS PAPAGEORGIOU, ARIS FERGADIS AND DIMITRIS PAPPAS
DOI: 10.22163/FTEVAL.2025.697

ABSTRACT

This study presents a data-driven methodology for evaluating the impact of publicly funded research, addressing the growing complexity of research and innovation landscapes. By integrating diverse data sources (including publications, clinical trials, and company websites) and leveraging advanced analytics such as natural language processing (NLP) and deep learning workflows, this approach overcomes traditional limitations in research impact evaluation. A case study on rare diseases demonstrates how the methodology uncovers pathways linking research outputs to societal benefits while balancing automation with expert validation to ensure accuracy and relevance. These findings underscore the strategic importance of robust, data-driven insights for aligning research priorities with evolving societal imperatives.

Keywords: Research Impact Evaluation, Rare Diseases, Natural Language Processing, Publicly Funded Research, Horizon 2020

1. INTRODUCTION

Evaluating the societal impact of publicly funded research is a significant challenge, often constrained by extended timelines and the complex, interconnected nature of research landscapes. Research outcomes frequently require years or even decades to translate into societal or economic benefits, involving diverse actors, disciplines, and outputs. This misalignment between the extended timelines of research impact and the shorter cycles of policy evaluation underscores the need for innovative approaches. Traditional evaluation methods, reliant on structured data and statistical indicators, provide a baseline understanding but often fail to capture the intricate pathways through which research drives societal change. For instance, foundational knowledge from a project may indirectly influence innovations years later, connections that are difficult to trace without advanced tools.

To address these challenges, this study presents a methodology that integrates diverse data sources with advanced artificial intelligence (AI) approaches, including text mining, natural language processing (NLP), and machine learning (ML). By linking datasets such as projects, publications, and corporate activities, the methodology uncovers connections between research outputs and societal impacts, a key for evidence-based policymaking. Combining scalable automation with expert human oversight, it ensures both accuracy and contextual relevance, adapting to the complexity and diversity of modern research. Expert-informed interpretation ensures that subtle or long-tail pathways of influence are recognised.

The methodology is demonstrated through a case study on rare diseases, a domain of significant societal importance that exemplifies the need for collaborative and long-term research efforts. Rare diseases, while individually uncommon, collectively affect from 27 to 36 million people in the European Union.[1] Between 2014 and 2020, the EU invested more than €2.9 billion in over 600 rare disease research and innovation (R&I) projects under FP7 and Horizon 2020.[2] Despite this substantial investment, understanding how research outputs translate into tangible societal benefits remains a critical challenge.

1 https://research-and-innovation.ec.europa.eu/research-area/health/rare-diseases_en

2 European Commission, 2021

A central component of the methodology is the use of big data analytics to process vast amounts of structured and unstructured information. Recent advancements[3] in tools such as knowledge graphs and ensemble algorithms enable the extraction of meaningful insights from diverse datasets, mapping the lifecycle of research activities and providing policymakers with actionable intelligence for targeted interventions. Moreover, the inclusion of innovative indicators, which extend beyond traditional metrics, facilitates a richer understanding of research impact. These indicators can capture contextual dimensions, such as alignment with global priorities like the United Nations' Sustainable Development Goals (SDGs)[4] or relevance to specific health challenges.

While this approach represents a step forward in research impact evaluation, it is not without limitations. Indirect and nuanced pathways often require qualitative insights that cannot be fully automated, underscoring the importance of expert validation to ensure analytical robustness. Moreover, the interconnected nature of research landscapes introduces additional complexity: multiple developments frequently occur simultaneously, influenced by diverse actors, external events, and evolving societal needs. Even with advanced tools and methodologies, it is often impossible to definitively attribute specific outcomes to individual projects or interventions. This highlights the need for cautious interpretation and an appreciation of the dynamic, multifaceted nature of research impact. Even if definitive attribution is elusive, partial, or probabilistic, insights are invaluable for shaping policy decisions.

This paper is organised as follows: Section 2 details the methodology underpinning the evaluation framework, Section 3 presents the rare diseases case study, Section 4 presents the results, and Section 5 concludes with a discussion of the broader implications of this methodology for research assessment and evidence-based policymaking.

3 European Commission, 2023a

4 https://sdgs.un.org/goals

2. METHODOLOGY IN BRIEF

This study employs a multifaceted methodology for evaluating R&I activities, bringing together AI techniques and domain expertise. While this study applies the framework to rare diseases, the methodology is designed to be research-theme agnostic and can be adapted to various fields, including energy, climate, and digital technologies. The approach builds on prior work conducted in IntelComp[5] and Data4Impact,[6] which explored AI-driven frameworks for assessing research impact through large-scale data integration and advanced analytics.[7] While this section provides an overview, a forthcoming technical paper will elaborate on the specific workflows and computational models in greater detail.

The framework is designed to accommodate large-scale and heterogeneous data sources, producing policy-relevant indicators while maintaining robust oversight through expert validation. By blending automated and human-driven processes, the approach aims to strike a balance between scalability and interpretative accuracy.

GUIDING PRINCIPLES

Several guiding principles shape this methodology. First, it adopts a 360° view of data, integrating a broad range of R&I information, including publications, patents, industry records, and policy documents, among others. This holistic perspective is enriched by standardised frameworks, such as the SDGs and the International Classification of Diseases (ICD),[8] which situate research outputs within broader societal and policy contexts. Second, the workflow is modular and end-to-end, covering data cleaning, information extraction, integration, and final analysis. Third, it embodies the expert-in-the-loop paradigm, recognising that AI-generated outputs require human review and domain contextualisation to ensure validity and alignment with policy objectives. Finally, openness and transparency guide all activities, from data handling (e.g. adherence to FAIR principles) to methodological documentation, fostering trust and replicability.

5 Horizon 2020 project, with grant ID 101004870, https://cordis.europa.eu/project/id/101004870

6 Horizon 2020 project with grand ID 770531, https://cordis.europa.eu/project/id/770531

7 Grypari et al., 2020; Stanciauskas et al., 2020

8 https://www.who.int/standards/classifications/classification-of-diseases

DATA SOURCES AND PREPARATION

A core strength of the methodology lies in its capacity to merge structured and unstructured data reflecting multiple stages of the research lifecycle. Project databases offer foundational information on objectives, consortium structures, and funding levels. Scientific outputs, particularly publications, serve as an initial measure of research activity and dissemination. Patents, clinical trials, and other innovation-related data provide indicators of technology transfer and translational progress, whereas industry data, such as company websites, illuminate commercialisation pathways. The framework incorporates broader societal elements, such ESG metrics, regulations, policies, and human resources (skills demanded vs supplied), to create a comprehensive view of how research may impact economic, environmental, and societal imperatives. Finally, ontologies and standards, including the ICD and SDGs, facilitate semantic enrichment and consistent categorisation.

DATA PREPARATION

Each dataset undergoes a comprehensive cleaning, disambiguation, and deduplication process, removing inconsistencies and redundancies. Structured metadata, such as project IDs and publication DOIs,[9] are reconciled with unstructured content (e.g. abstracts, company websites) to form a unified database. In many cases, semantic linking is applied, mapping disease mentions or similar references to standardised terminologies (e.g. ICD codes). This ensures that subsequent analyses operate on harmonised, context-rich data.

EXTRACTION

Building on this curated dataset, machine learning (ML) and natural language processing (NLP) techniques extract and categorize relevant entities. Named Entity Recognition (NER) models identify diseases, technologies, and other key entities, while topic modelling detects thematic structures and helps capture how research priorities evolve over time.

To better capture relationships between research outputs, we apply semantic similarity analysis to identify connections between different publications, projects, patents and so on. Additionally, co-occurrence analysis helps detect recurring associations between key terms, providing insight into emerging research directions. These extracted entities and relationships are structured

9 https://www.doi.org/the-identifier/what-is-a-doi/

into knowledge graphs, which link research topics to relevant stakeholders, funding programs, and translational applications such as clinical trials and industrial uptake.[10] By organizing research impact pathways in this structured manner, the methodology enables downstream analysis to assess how publicly funded research contributes to long-term innovation and societal benefits.

INTEGRATION

After extraction, the framework integrates these varied data streams to illuminate broader connections. ML classifiers categorise research outputs according to established taxonomies (e.g. SDGs) to ensure alignment with recognised global priorities. In parallel, impact pathway analysis uses ML models to trace how early-stage findings (e.g. publications) transition into tangible applications (such as clinical trials, patents, or commercial products). Through this process, the methodology generates novel metrics – for example, gauging how far an idea has progressed from fundamental research to real-world implementation. As the database grows in size and quality, the framework's analytical precision improves, allowing for more reliable impact assessments across disciplines.

SYNTHESIS

In the final stage, inference methods evaluate the R&I ecosystem from a holistic standpoint. Citation networks depict the longevity and influence of foundational work, highlighting how discoveries spread across disciplines. Industry uptake scores measure how publicly funded research permeates ongoing industrial R&D, indicating potential commercialisation pathways. Beyond standard indicators like citation counts, contextualised measures (e.g. thematic momentum) provide a more nuanced understanding of research impact, one that is directly relevant to policymakers responsible for guiding future funding and innovation strategies.

TECHNOLOGICAL INFRASTRUCTURE

This methodological design operates within a cloud-native, modular architecture capable of supporting computationally demanding NLP and ML workflows:

10 In this paper, 'knowledge' refers to structured information about research outputs and their interconnections, derived from multiple data sources (e.g., publications, patents, clinical trials, and company websites).

- High-Performance Computing (HPC) environments enable large-scale data ingestion and batch-processing tasks, ensuring the timely analysis of extensive, heterogeneous datasets.
- Containerisation (e.g. Docker) and continuous integration/continuous deployment (CI/CD) pipelines facilitate rapid, iterative model development, allowing the framework to evolve in tandem with emerging analytical tools.
- Microservices and distributed infrastructure provide scalability, adaptability, and efficient resource utilisation, making it feasible to integrate additional modules (e.g. new entity classes or ontologies) without disrupting the overall pipeline.

By integrating diverse data sources, applying advanced NLP and ML techniques, and embedding expert validation throughout, this methodology presents a transparent and adaptable framework for R&I evaluation.

3. RARE DISEASES AS A CASE STUDY

Rare diseases (RD) pose a pressing public health challenge in the European Union (EU). They are defined as affecting no more than one person in every 2,000 individuals. Collectively, however, these conditions impact approximately 36 million people across the EU, and encompass 6,000 to 8,000 distinct disorders.[11] They are frequently characterised by high unmet medical needs, significant variability in clinical presentations, and limited treatment options, necessitating substantial collaboration at both European and international levels. Such collaboration draws on diverse expertise, from clinical practice to biotechnology and health policy, underscoring the inherent complexity of rare diseases and the need for robust, cross-sectoral approaches. Recognising the societal and economic implications of rare diseases, the EU has made them a key focus of research and innovation activities. Approximately 70% of these disorders manifest in childhood, often leading to diagnostic delays, challenging care requirements, and profound long-term impacts on patients and families.

11 European Commission, 2021

TAILORING THE METHODOLOGY TO RARE DISEASES

While the methodology outlined in Section 2 is broad enough to evaluate large-scale R&I activities across disciplines, our use case on rare diseases serves as a focused illustration. The framework itself is research-theme agnostic and can be applied to other domains. Rather than attempting a fully comprehensive analysis, we have selectively applied certain data sources to highlight how the framework can be adapted to specific domains. This narrower scope underscores its flexibility and capacity to generate meaningful insights across varying scales. A central task for the case study is establishing a rare disease project portfolio that adequately represents both the breadth of EU investments and the depth of targeted research activities. To achieve this, a multi-layered approach was adopted:

1. **Extended Portfolio**: Natural Language Processing (NLP) and probabilistic models were used to scan a wide range of EU-funded projects, capturing direct and indirect references to rare diseases (about 10,000 projects).

2. **Core Portfolio**: From this broad set, additional filters were introduced to isolate projects explicitly addressing rare disease topics, ensuring higher specificity (about 1,700 projects).

3. **Curated Portfolio**: Finally, manual review by domain experts confirmed a subset of projects with a primary and direct focus on rare diseases (about 400 projects).

By combining automation with expert input, this three-tiered structure enables flexible analyses: one can examine thematic diversity in the extended dataset while zeroing in on more specialised findings in the curated list. While expert validation was applied here in the context of rare diseases, this approach is adaptable to other fields by incorporating domain-specific expertise at key validation stages, ensuring accuracy and relevance regardless of the research area.

To navigate the multifaceted nature of rare disease research, the study relies on an array of NLP and graph-based approaches. For instance, entity extraction and topic modelling identify critical diseases, and thematic clusters within project documentation[12], publications,[13] clinical trials,[14] and company websites. Citation graphs and knowledge graphs trace how research outputs are interlinked and how these relationships evolve over time. These techniques collectively address two prominent challenges: the fragmentation of data across scattered sources and the difficulty of following a project's influence through multiple, often indirect, pathways.

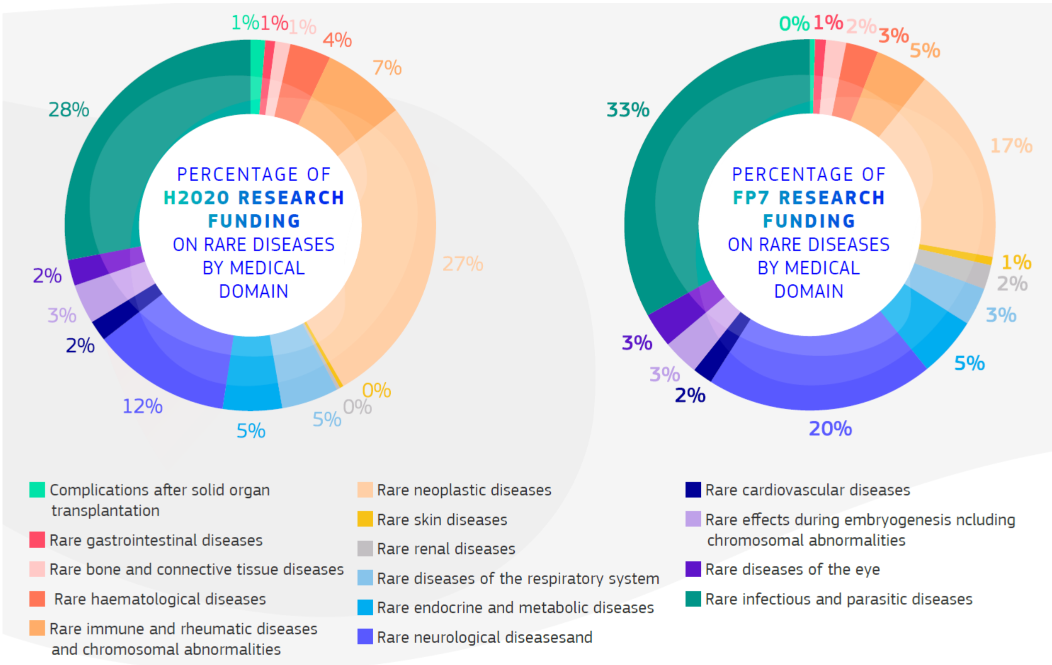


Figure 1: Distribution of Rare Disease Funding Across ICD-10. Source: European Commission, Directorate-General for Research and Innovation, Collaboration: a key to unlock the challenges of rare diseases research, Publications Office, 2021

A summary published by the European Commission (Figure 1) illustrates the percentage of rare disease funding allocated to different ICD-10[15] categories under FP7 and Horizon 2020. Compared to FP7, there appears to be a relative increase in funding for nervous system disorders and congenital anomalies under Horizon 2020. These shifts in emphasis offer a preliminary snapshot of how EU research priorities in rare diseases evolved between the two programmes.

12 Data from CORDIS https://cordis.europa.eu/

13 Data from the OpenAIRE Graph https://graph.openaire.eu/

14 Data from PubMed https://pubmed.ncbi.nlm.nih.gov/, and ClinicalTrials.gov https://clinicaltrials.gov/

15 https://icd.who.int/browse10/2019/en

Results and insights drawn from applying our adapted methodology to the curated project portfolio will explore patterns in greater depth, examining how the thematic distribution, collaboration networks, industrial uptake, and clinical links collectively shape the rare disease research ecosystem.

4. RESULTS & INSIGHTS

This section presents the principal findings obtained by applying the data-driven framework (described in Section 2) to our curated portfolio of rare disease projects under FP7 and Horizon 2020.

EMERGING THEMES IN RARE DISEASE RESEARCH

We begin by applying topic modelling to the full texts of project descriptions and scientific publications. This unsupervised approach clusters thematically related documents and labels them according to human expert input, enabling us to identify which disease areas or research themes gained (or lost) prominence between FP7 and Horizon 2020.

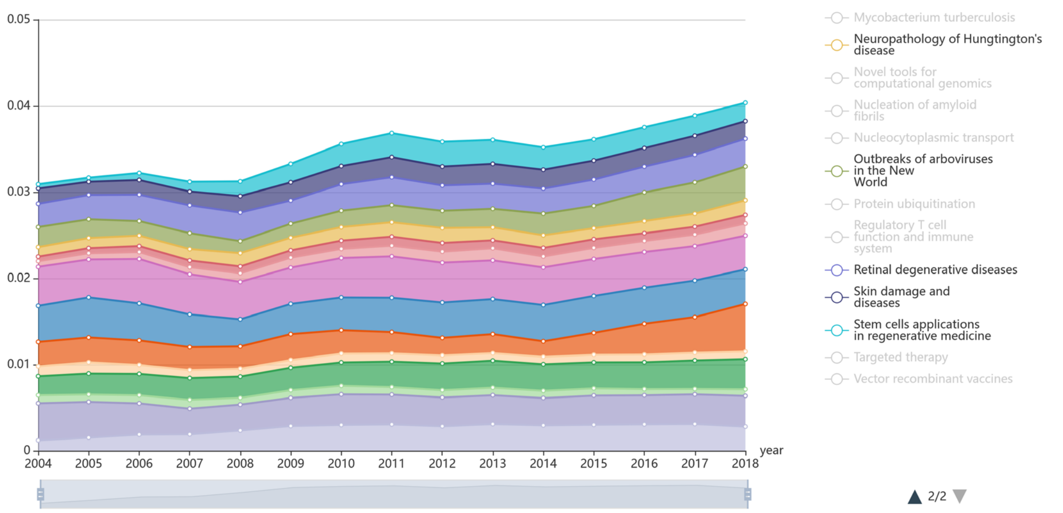


Figure 2: Evolution of Select Topics in Rare Disease Portfolio Over Time. Sources: OpenAIRE Graph, CORDIS.

Figure 2 displays examples of topic evolution across time. Notably, the 'Outbreaks of Arboviruses in the New World' topic rises markedly under Horizon 2020, coinciding with heightened global concerns over Zika and dengue, which have been particularly prominent in Latin America. In contrast, while malaria remains one of the most EU-funded research areas, its topic momentum in H2020 is lower—despite its persistent burden, especially in Sub-Saharan Africa.

It is important to interpret these results in context. The increased focus on arboviruses can be seen as an illustration of the EC's capacity to address urgent crises. However, it does not necessarily imply diminished attention to other high-burden diseases, including malaria and leishmaniasis. In several calls, the natural draw of scientists and public health actors specialising in arboviruses, many located or collaborating in Latin America, led to a proportional rise in projects on these topics. As shown and explained in Figure 4 below, the EC's investments in both arboviral and malaria/leishmaniasis research still outpaces the broader health field's average proportion. Nonetheless, the data highlight how rapid shifts in global health needs can shape which research themes gain traction in any given funding cycle.

COLLABORATION PATTERNS AND REGIONAL DISPARITIES

To examine cross-organisational partnerships and co-publications, we constructed graphs linking projects, participating organisations, and the resulting publications. The visualisations below offer an "user friendly" birds' eye perspective on how project participants and coauthors collaborated over time in rare disease projects.

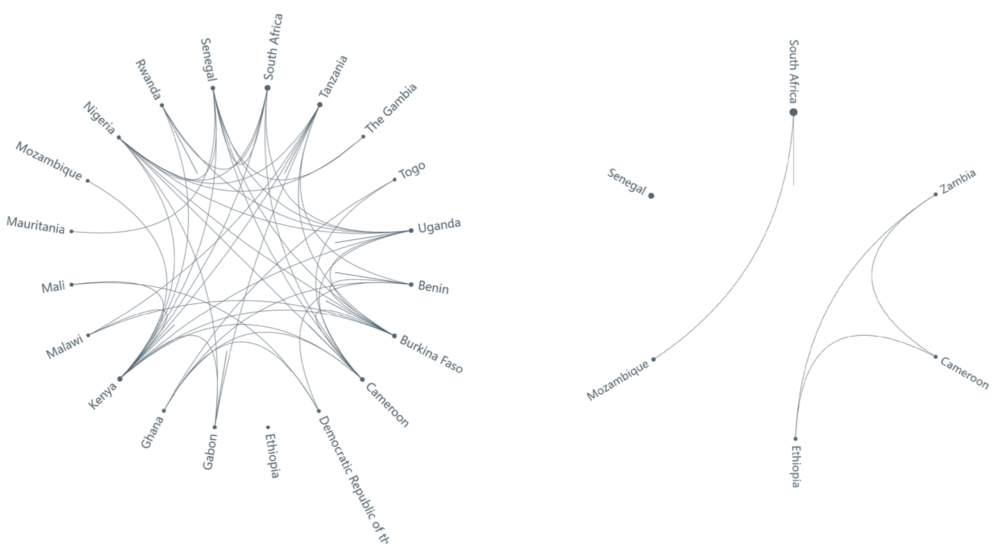


Figure 3: Project & Publications Collaborations in FP7 (left) and H2020 (right) in Sub-Saharan Africa. Sources: OpenAIRE Graph, CORDIS.

Figure 3 illustrates the collaboration network for sub-Saharan Africa organisations, indicating that African organisations formed stronger co-participation ties under FP7. During Horizon 2020, the data show a pronounced surge in collaborations between Latin American partners (not shown above), consistent with the rise of arbovirus-related topics. Crucially, this reorientation does not necessarily mean sub-Saharan Africa received fewer resources in absolute terms; rather, the number of joint publications and grants involving African partners decreased.

The figure below uses each topic's share of publications as a proxy measure of resource allocation and research focus. While this metric helps illustrate relative emphasis, it does not represent a precise accounting of budgets or project-level expenditure. Nonetheless, it supports the broader finding that the EC-funded topics depicted often exceed the global field's average proportion, indicating a deliberate policy to address areas requiring public-sector intervention. As shown, both "Outbreaks of Arboviruses in the New World" and "Malaria and Leishmaniasis" command disproportionate investment, reflecting the EC's strategic emphasis on these research challenges.

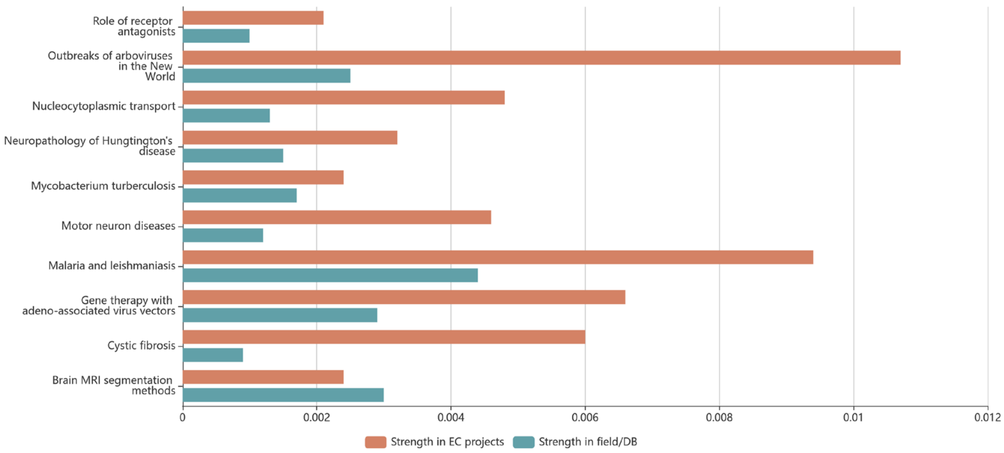


Figure 4: EC vs the Field, Investment in Topics as a Share of Total Publications. Sources: OpenAIRE Graph, CORDIS

INDUSTRIAL UPTAKE AND R&D CONTINUITY

A major objective of EU-funded research is to stimulate industrial innovation. To assess whether companies continued working on project-related topics, we applied a deep learning method to analyse text from company websites, measuring semantic similarities with each company's past project deliverables. This yielded an R&D Uptake Score, which quantifies how closely a firm's current activities resemble its earlier, publicly funded research.

In Figure 5, companies closer to the outer edge of the circle (score ~1) exhibit strong continuity between their present-day research and the innovations they pursued under EC projects, whereas those nearer to the centre (score ~0) appear to have shifted focus. A low score does not imply low impact; firms may pivot strategically in response to market signals or integrate project methodologies into different domains. Conversely, a high score suggests thematic consistency but does not guarantee successful product development. Additional data, such as patent portfolios, licensing records, or clinical trial sponsorship, would further enrich assessments of how public investments translate into commercial outcomes.

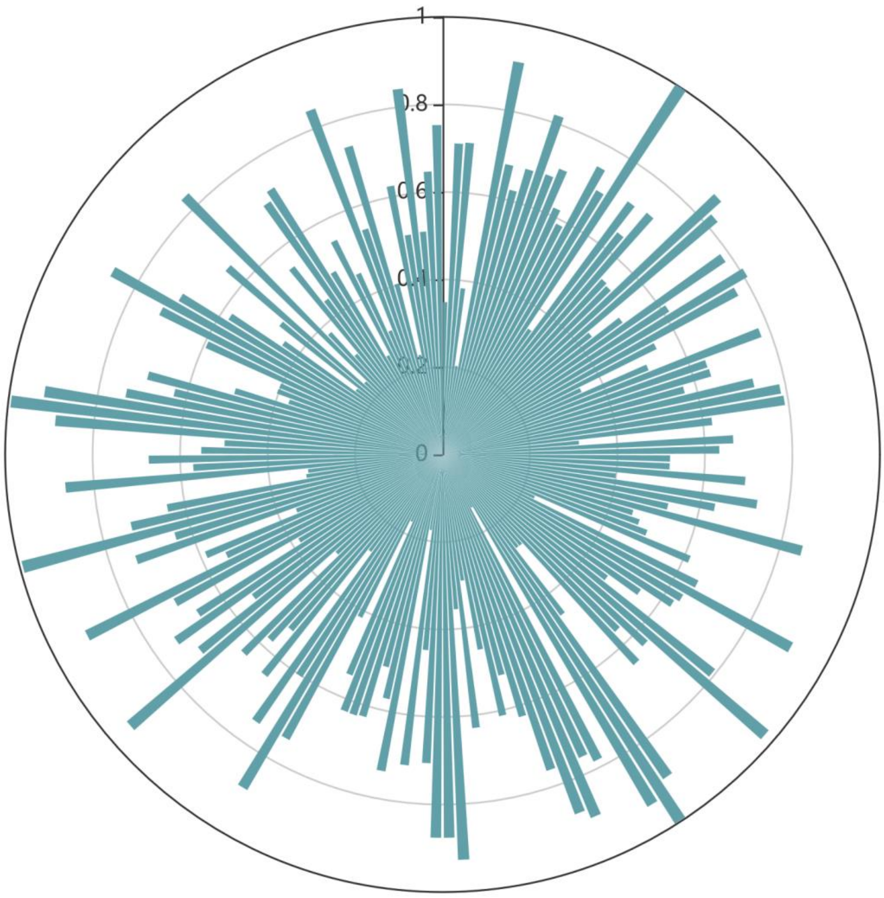


Figure 5: Company Uptake Score for Rare Disease Portfolio. Sources: OpenAIRE Graph, CORDIS, Company Websites. Each column represents a company, with column height reflecting the uptake score, a measure of how closely a firm's current research aligns with its past EC-funded projects. The score is calculated using AI-driven text similarity analysis, comparing company website content with project publications. Higher scores (taller columns) indicate continued work in the same field, while lower scores suggest a shift in focus.

LINKING TO HEALTH OUTCOMES VIA CLINICAL TRIALS AND GUIDELINES

A final cornerstone of our assessment tracks whether project outputs feed into clinical trials, a critical juncture between scientific discovery and patient outcomes. Our analysis revealed over 1,800 trials citing publications linked to the rare disease portfolio, including 843 trials in which at least one original project participant was directly involved. This engagement demonstrates the continuity from research funding to the clinical testing phase.

Nevertheless, progress from publication to improved patient care can be slow. 100 clinical guidelines were found to reference the curated projects' publications, and in six cases, the guidelines included a direct mention of the project in their metadata. Some guidelines emerged more than three years after the project's end date, illustrating the iterative and often protracted path from funded research to real-world application. Even "unsuccessful" trials can shape best practices or refine methodological approaches, serving as stepping stones toward future breakthroughs.

To capture these indirect but critical contributions, additional contextual indicators could offer a more complete understanding of how public research investments unfold in clinical practice. As the use case below demonstrates, interpreting the full chain of evidence demands careful triangulation among multiple data sources.

A CLINICAL TRIAL USE CASE: ANALYSING PATHWAYS TO SOCIETAL IMPACT

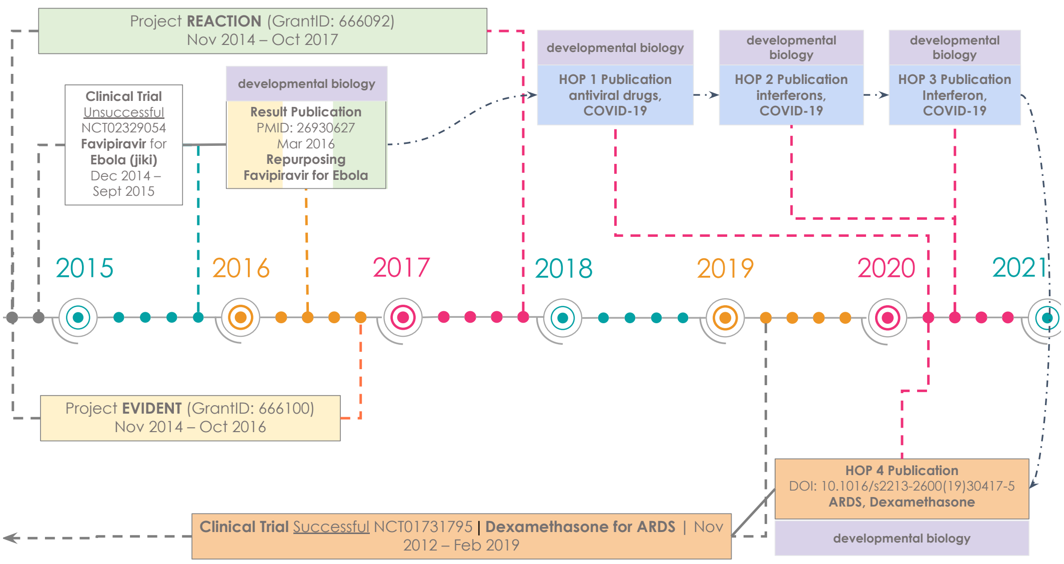


Figure 6: Flowchart of a Clinical Trial Use Case

In the domain of health research, clinical trials act as a pivotal bridge between laboratory innovations and tangible societal benefits. A compelling illustration of these research-to-impact pathways arises from two Horizon 2020–funded projects, REACTION and EVIDENT, which jointly produced a publication (PMID: 26930627) summarising the outcomes of the "jiki" trial (NCT02329054). This trial investigated the repurposing of Favipiravir for Ebola treatment and ended unsuccessfully at Phase 2. However, advanced citation analysis revealed a four-hop linkage connecting the "jiki" trial to a successful clinical trial (NCT01731795) on the use of Dexamethasone for Acute Respiratory Distress Syndrome (ARDS), as depicted in the flowchart above.

This multi-hop chain illustrates the cumulative nature of scientific discovery: even halted or "unsuccessful" trials can contribute knowledge that informs later breakthroughs. The "jiki" trial was cited in 190 subsequent publications, nine of which were associated with other trial efforts that themselves ended prior to commercial success. Techniques such as semantic similarity analysis, and topic evolution tracking helped uncover these indirect pathways. Yet, human expertise remains essential for validating weak or ambiguous connections, ensuring contextual accuracy, and mitigating the risk of over-attribution.

Although the path from Ebola research to an ARDS breakthrough might appear tenuous, tracing through multiple layers of citations and knowledge diffusion, the example underscores the incremental, interwoven nature of health research. While direct attribution is difficult, the original Ebola trial contributed to a growing body of knowledge, influencing subsequent studies that may have played a role in shaping later breakthroughs.

5. CONCLUSION

This study demonstrates how a data-driven, AI-augmented methodology can illuminate the often intricate and indirect pathways that link publicly funded research to societal outcomes. By integrating large-scale datasets, advanced NLP and ML tools, and graph-based analyses, the approach uncovers patterns and relationships that traditional methods might overlook. However, our findings also highlight the inherent complexity of attributing research impacts, given the diversity of actors, the interwoven nature of developments, and the delayed emergence of tangible benefits.

Despite its strengths, the framework requires careful interpretation. AI-driven analyses may oversimplify complex relationships or introduce spurious correlations, making expert validation essential to ensure meaningful insights. Additionally, not all types of impact, such as policy influence, are easily captured with structured data alone, highlighting the need for complementary qualitative assessments.

For researchers and policymakers looking to apply this framework in other disciplines, its scalability and adaptability are key advantages, but expert knowledge remains critical for ensuring results are contextually valid. In fields with fewer structured indicators, AI's role may shift from directly identifying patterns to helping generate hypotheses, requiring an iterative process between automated insights and expert interpretation. Ensuring reliability in AI-driven findings demands cross-validation across multiple data sources, transparency in methodological assumptions, and active monitoring for biases in both data and model design.

From a policymaking perspective, this synergy between AI-driven analytics and expert validation provides a powerful tool for evidence-based decision-making. This approach enables stakeholders to better allocate resources, support high-impact collaborations, and track emerging research priorities, while remaining aware of the limitations of purely algorithmic methods. Finally, new impact metrics generated through this methodology help address gaps in traditional assessment frameworks, but their value depends on continuous refinement, interdisciplinary validation, and engagement with the broader research community. By maintaining a balance between advanced analytics and expert oversight, research investments can better align with societal goals, maximizing their long-term impact.

REFERENCES

Adam, G. P., Pappas, D., Papageorgiou, H., Evangelou, E., & Trikalinos, T. A. (2022). A novel tool that allows interactive screening of PubMed citations showed promise for the semi-automation of identification of Biomedical Literature. Journal of Clinical Epidemiology, 150, 63–71. https://doi.org/10.1016/j.jclinepi.2022.06.007

European Commission: Directorate-General for Research and Innovation, Collaboration – A key to unlock the challenges of rare diseases research – February 2025, Publications Office of the European Union, 2025. https://data.europa.eu/doi/10.2777/8029727

European Commission: Directorate-General for Research and Innovation, Collaboration – A key to unlock the challenges of rare diseases research, Publications Office, 2021. https://data.europa.eu/doi/10.2777/249334

European Commission: Directorate-General for Research and Innovation and Denham, S., Tracking of research results – Specific contract No 2018/RTD/TRR003 – Final data and data sources report, Denham, S. (editor), Publications Office of the European Union, 2023. https://data.europa.eu/doi/10.2777/57692

European Commission: Directorate-General for Research and Innovation and Denham, S., Tracking of research results – Specific contract No 2018/RTD/TRR003 – TRR final report, Denham, S. (editor), Publications Office of the European Union, 2023. https://data.europa.eu/doi/10.2777/128148

Fergadis, A., Pappas, D., Karamolegkou, A., & Papageorgiou, H. (2021, November). Argumentation mining in scientific literature for sustainable development. In Proceedings of the 8th Workshop on Argument Mining (pp. 100-111). https://doi.org/10.18653/v1/2021.argmining-1.10

Grypari, I., Pappas, D., Manola, N., & Papageorgiou, H. (2020, May). Research & Innovation Activities' Impact Assessment: The Data4Impact System. In Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov) (pp. 22–27).

Kotitsas, S., Pappas, D., Manola, N., & Papageorgiou, H. (2023). SCINOBO: a novel system classifying scholarly communication in a dynamically constructed hierarchical Field-of-Science taxonomy. Frontiers in Research Metrics and Analytics, 8, 114983. https://doi.org/10.3389/frma.2023.1149834

Saeidnia, H. R., Hosseini, E., Abdoli, S., & Ausloos, M. (2024). Unleashing the power of AI: a systematic review of cutting-edge techniques in AI-enhanced scientometrics, webometrics and bibliometrics. Library Hi Tech. https://doi.org/10.1108/LHT-10-2023-0514

Stanciauskas, V., Grypari, I., Nelhans, G., Papageorgiou, G., & Demiros, I. (2020). Policy report on new indicators and approaches for assessing the societal impact of research and innovation activities: Big Data approaches for improved monitoring of re-search and innovation performance and assessment of the societal impact in the Health, Demographic Change and Wellbeing Societal Challenge.

Stavropoulos, P., Lyris, I., Manola, N., Grypari, I., & Papageorgiou, H. (2023, December). Empowering Knowledge Discovery from Scientific Literature: A novel approach to Research Artifact Analysis. In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023) (pp. 37–53). https://doi.org/10.18653/v1/2023.nlposs-1.5

ACKNOWLEDGEMENT

This work was partially funded by the European Union through the Horizon 2020 projects Data4Impact (Grant Agreement ID: 77053) and IntelComp (Grant Agreement ID: 101004870). The views and opinions expressed herein are solely those of the authors and do not necessarily reflect those of the European Union or the relevant granting authorities. Neither the European Union nor the granting authorities can be held responsible for them. Special thanks to Christina Kyriakopoulou for her meticulous curation of the list of rare disease projects and other project outputs and to Tsveta Schyns-Liharska (EC Blue Book Trainee) for offering their valuable insights in the earlier stages of this work.

AUTHORS

IOANNA GRYPARI (Corresponding Author)
Email: ioanna.grypari@opix.ai
OPIX P.C., https://www.opix.ai
Athena Research Center, https://www.athenarc.gr/en
OpenAIRE AMKE, https://www.openaire.eu/
ORCID: 0000-0002-7485-1591

SERGIO DI VIRGILIO
DG R&I, European Commission, 1049 Brussels, Belgium
ORCID: 0000-0001-5751-9661

HARIS PAPAGEORGIOU
OPIX P.C., https://www.opix.ai
Athena Research Center, https://www.athenarc.gr/en
ORCID: 0000-0002-7352-2403

ARIS FERGADIS
OPIX P.C., https://www.opix.ai

DIMITRIS PAPPAS
OPIX P.C., https://www.opix.ai
ORCID: 0000-0001-5784-0658

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e6, 1-28
DOI: 10.22163/fteval.2025.698

EVALUATING TRANSFORMATIVE INNOVATION POLICY INSTRUMENTS

THE EXAMPLE OF THE AUSTRIAN PROGRAMME FOR THE PROMOTION OF FEMALE RESEARCHERS "INNOVATORINNEN"

VERENA RÉGENT AND BRIGITTE ECKER
DOI: 10.22163/FTEVAL.2025.698

ABSTRACT

This paper focuses on the evaluation of transformative innovation policy (TIP) instruments through the case of the Austrian programme INNOVATORINNEN which promotes women in applied research and innovation. To contextualise the empirical analysis, the paper reviews the evolution of innovation policy, focusing on instruments designed to support female researchers and advance gender equality in Austria, as well as their evaluations. The analysis positions INNOVATORINNEN as a transformative innovation policy instrument.

Using deductive content analysis, the evaluation concept and practice of the INNOVATORINNEN programme are assessed against the requirements for TIP evaluations outlined in recent literature. Key aspects examined include evaluation strategy, the role of evaluation, theory of change, and methodology. The findings reveal a strong alignment between the programme's evaluation and TIP evaluation criteria, particularly in fostering programme learning, reflection, and evidence-based development. Notably, the evaluation incorporates content-oriented, co-creative processes that actively engage programme participants and programme owners, resulting in a "knowledge-triangle" driving the continuous improvement of the programme. The paper concludes by reflecting on key lessons from the evaluation process,

emphasizing the importance of openness, flexibility, a willingness to learn, trust and mutual respect among all involved parties.

Keywords: transformative innovation policy (TIP), TIP evaluation, promotion of female researchers, application-oriented research programme, deductive content analysis

1. INTRODUCTION

Innovation policy has seen a shift when it comes to the main targets of investment in research and innovation – an evolution that has been discussed in recent academic literature, amongst others by Weber and Rohracher (2012), Schot and Steinmüller (2018) or Joly and Matt (2022). Transformative innovation policy (TIP) is an emerging generation of innovation policies, reorienting public science funders' and innovation policy professionals' efforts for initiating or contributing to societal change on a broad scale (Ghosh et al., 2021). TIP becomes most evident when it comes to governmental responses to recent global policy agendas, such as the *Sustainable Development Goals* (SDGs), the *Paris Climate Agreement*, the European Union *Green Deal,* and the 2020 World Economic Forum agenda on "Fixing inequality" (Ghosh et al., 2021). In the context of TIP, new requirements for the evaluation of policy instruments arise, as outlined e.g. in Molas-Gallart et al. (2020 and 2021) and Boni et al. (2019).

This paper focuses on the Austrian programme for the advancement of women in application-oriented research and innovation (R&I) "INNOVATORINNEN" by the Austrian *Ministry of Labour and Economy* (BMAW) and the *Austrian Research Promotion Agency* (FFG). This paper argues that INNOVATORINNEN is a case of a transformative innovation policy instrument. Particular attention is given to the accompanying evaluation of the programme: The purpose of this contribution is to provide insights into the evaluation, which – next to a set of "traditional" policy evaluation methods – uses novel evaluation elements based on strong interaction with programme management and participants. In this light, this contribution also aims to elaborate on how far the evaluation concept and practice qualify as TIP evaluation. To achieve this, the evaluation concept is analysed in the face of the requirements of TIP evaluations as proposed by Molas-Gallart et al. (2020 & 2021), Wise et al. (2022), TIPC (2019), Boni et al. (2019), and Ghosh et al. (2021).

The paper is structured as follows: Section 2 explains the theoretical and empirical background, focusing on the development of innovation policy, its underlying understanding of innovation and its evaluation, as well as on the development of innovation programmes for advancing women in R&I in Austria. Section 3 presents the INNOVATORINNEN programme as the empirical context of the paper and provides arguments for its characterisation as potential TIP instrument. Moreover, it sheds light on the accompanying evaluation of the programme. Section 4 details the research question and methodological approach; section 5 deals with the findings of the analysis of the evaluation concept, and section 6 provides a conclusion.

2. EXPLANATION OF THEORETICAL AND EMPIRICAL BACKGROUNDS

This section delves into the theoretical and empirical foundations of the chapter. The first sub-section examines recent developments in innovation policy, situating the emergence of instruments designed to promote female researchers in R&I within this broader context. The second sub-section explores the evolution of evaluation approaches that have accompanied the progression of innovation policy over time. Finally, the third sub-section highlights key innovation policy instruments aimed at supporting female researchers and innovators in applied R&I in Austria.

2.1 THE WAY TOWARDS TRANSFORMATIVE INNOVATION POLICY

In line with a changing understanding of innovation and its effects, innovation policy has changed throughout the past decades: This development has been traced in recent academic literature (e.g. Weber & Rohracher, 2012; Schot & Steinmüller, 2018, Joly & Matt, 2022) and can be summarised in three frames, which are outlined in the following based on Wise et al. (2022), Carayannis & Campbell (2009, 2012), Schot et al. (2019) and Wroblewski & Schaller-Steidl (2023). Figure 1 provides an overview of the three frames of innovation policy, also pointing to the underlying notion of innovation and innovation policy for the advancement of women in R&I.

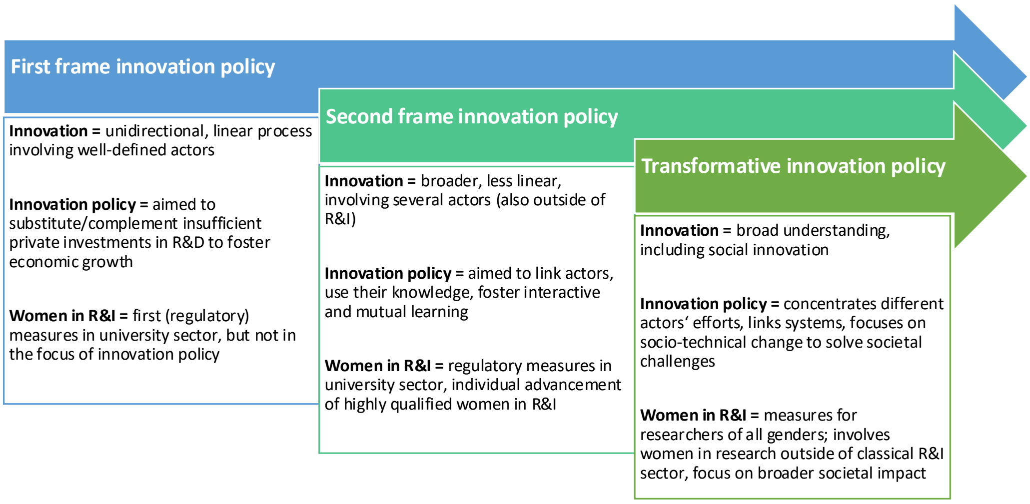


Figure 1 Overview of innovation policy frames. (Source: Own illustration, based on Wise et al. (2022, p. 273), Carayannis & Campbell (2009, 2012), Schot et al. (2019) and Wroblewski & Schaller-Steidl (2023))

In the *first frame*, prevalent between the 1970s and 1980s, innovation was considered a unidirectional, linear process from development to commercialisation, involving well-defined actors from the R&I sector (Carayannis & Campbell, 2012, p.3), and as means to foster economic growth. Accordingly, innovation policy primarily aimed at solving the market failure of insufficient private investments in R&I. Policy instruments in use were aimed at stimulating knowledge generation, developing regulatory and educational policy as well as raising awareness of the importance of innovation and technological advancement (Wise et al., 2022, p. 272). The first measures for advancing female researchers in Austria appeared in the 1970s, however, they were mainly limited to the university sector, while the area of applied research close to industry in large parts remained unregulated. Promoting women was not the primary focus of the *first frame*-innovation policy; this is not solely due to the then-prevailing notion of innovation but is primarily attributable to the societal conditions and practices that were dominant at that time.

In the *second frame*, from the 1980s up until today, the understanding of innovation has become broader and less linear – for example, Etzkowitz' & Leydesdorff's (1995) triple helix, the concept of open innovation (e.g. Chesbrough, 2003) and Carayannis & Campbell's (2009) quadruple helix suggested to also involve actors outside the R&I sector, such as government, civil society and industry, into the innovation process. Accordingly, innovation policy has been based on the notion of better linking and using the knowledge of different actors alongside fostering mutual learning. Policy instruments under this frame aim at stimulating and facilitating linkages and coordination between actors to foster interactive learning, knowledge utilisation, innovation and entrepreneurship, which, in turn, stimulate economic growth (Wise et al., 2022). In the run of this *second frame,* promotion programmes for the advancement of women in science, research and innovation started to be established more systematically in Austria and Europe. Early programmes in the 1990s primarily focused on regulations on equal opportunities in the university sector; in the early 2000s, policy measures in Austria were extended to women in applied and industrial research, mainly comprising the individual advancement of highly qualified women from the R&I sector (Wroblewski & Schaller-Steidl, 2023; details see in section 3.3 of this paper).

The *third frame* of innovation policy is currently emerging and summarised under the term TIP, which is based on an extended understanding of innovation, including social innovation. The notion of the term transformation implies a change of socio-technical systems to solve complex societal challenges, such as the climate crisis, growing inequality, or a socioeconomic health crisis in the aftermath of the COVID-19-pandemic (Schot & Steinmüller, 2018; Ghosh et al., 2021). Focusing on this transformation, TIP aims to concentrate different actors' efforts, coordinating with other policy sectors and fostering new connections between systems. Consequently, TIP instruments focus on missions, challenge competition, or challenge-driven innovation programmes stimulating experimentation and co-production (Schot et al., 2019, p. 22–23). Within this *third frame*, inequalities in different areas of life are addressed as societal challenges, thus equal opportunities in R&I can be considered a major concern in TIP. In recent years, the measures for the advancement of women were partly redesigned to address equal opportunities for researchers of all genders in their early stages, while concrete objectives within these programmes are supposed to assure wide female participation (Wroblewski & Schaller-Steidl, 2023, chapter 3). At the same time, programmes such as INNOVATORINNEN – the empirical context of this paper – were introduced, aiming at promoting women in applied R&I based on an extended understanding of innovation, and at yielding broader societal impact (see section 4 of this paper).

2.2 EVALUATING TRANSFORMATIVE INNOVATION POLICY

With the broadening understanding of innovation and the development of innovation policy, the evaluation of policy instruments has been confronted with new requirements. While evaluation of innovation policies under the *first frame* was primarily focused on statistical measures of R&I inputs (e.g. funding sources, performers, personnel) and outputs (e.g. published articles, patents), the extended framing of innovation policy to system level (*second frame*) was accompanied by new evaluation strategies. In particular, survey methods and qualitative research methods were used to complement existing statistical approaches. The focus was shifted to new aspects of innovation, such as innovation capabilities or linkages between actors in the innovation process (Wise et al., 2022, p. 273).

In the context of TIP, the need for new evaluation strategies is highlighted by several authors (e.g. Molas-Gallart et al., 2020 and 2021, Boni et al., 2019). Next to the traditional purposes of (formative and summative) evaluation – assessing efficiency, effectiveness, and the relevance of policy programmes (Peersman, 2015) – Boni et al. (2019), Schot et al. (2019), Ghosh et al. (2021) and Molas-Gallart et al. (2021) call for a new evaluation strategy that comprises monitoring transformative outcomes and "signs of change" (in the shape of changes in behaviour, emerging constellation or relationships or activities among people, groups and organisations, evolution of strategic aims), and informing the direction of the pursued systemic change process. In line with Molas-Gallart et al. (2021, p. 435), these authors stress the integration of evaluation as strategic dimension of the given programme with the aim of enhancing reflexivity and learning.

According to Wise et al. (2022), TIP evaluation is strongly rooted in sustainability transitions literature and multi-level perspectives on socio-technical transitions. The Transformative Innovation Policy Consortium (TIPC, 2019) developed a "formative approach to TIP evaluation" that differs from traditional (formative, summative) and developmental (see e.g. Patton, 2016, p. 28) evaluation in several respects. In particular, it stresses mixed methods, participatory approaches, and the integration of evaluation as a formative and strategic dimension of a programme to support learning and to inform strategic choices over time. Both data gathering and analysis involve participating actors and try to include a variety of perspectives. The results are used as "food for thought" and guide the adjustment of the envisaged transformation path (Wise et al., 2022). Table 1 summarises the integrated characteristics of TIP evaluations.

Table 1 Characteristics of TIP evaluations (Source: Own illustration based on Patton (2006), Molas-Gallart et al. (2020 & 2021), Wise et al. (2022), TIPC (2019), Boni et al. (2019), Ghosh et al. (2021))

Evaluation strategy	▪ Monitoring transformative outcomes and signs of systemic change in real time (behavioural changes, emerging relationships / activities / constellation / interactions between actors, evolution of strategic aims) ▪ Informing the direction of the change process
Role of evaluation	▪ Evaluation integrated as strategic dimension of the programme to enhance reflexivity and learning ▪ Informing strategic choices concerning the programme over time ▪ Evaluation results used as “food for thought”, guide the adjustment of the envisaged transformation path, help to refine the transformation process
Theory of change	▪ Flexible, revisited, and refined throughout the evaluation process ▪ Nested approach to assess multiple levels
Methodology	▪ Mixed methods ▪ Participatory approaches in data gathering and analysis ▪ Include a variety of perspectives

2.3 DEVELOPMENT AND EVALUATION OF PROGRAMMES TO PROMOTE WOMEN IN APPLICATION-ORIENTED RESEARCH AND INNOVATION IN AUSTRIA

Austria looks back to a relatively long tradition of measures to promote the equality of women in research, science, and innovation. The first measures date back to the 1970s; however, it was not until the 1990s and thus far into the *second frame* of innovation policy, that an entire set of instruments can be identified. These instruments, next to the promotion of women, aimed at antidiscrimination and the establishment of women and gender studies as dedicated disciplines. The bulk of these early policies was concentrated on the publicly-financed university sector and facilitated the institutionalisation and professionalisation of equality approaches. Comparable efforts in the broader higher education sector (comprising universities of applied sciences (UAS), private universities, and universities for teacher education) only followed in the past decade (Wroblewski & Schaller-Steidl, 2023, p. 11ff).

The sector of applied research close to industry does not know any comparable regulations but benefitted from programmes that addressed female researchers’ individual career cycles and the consideration of the gender dimension in research, which were launched at the turn of the millennium (*second frame* of innovation policies). One prominent example is the programme fFORTE, which was recommended by the *Austrian Council* *for Research and Technology Policy* (*Rat für Forschungs- und Technologiepolitik,* RFTE) in 2001. With this intersectoral programme, the former *Ministry of Education, Science and Research* (BMBWF[1]) and the *Ministry of Traffic, Innovation and Technology* (BMVIT) addressed female researchers' entire educational and career cycles; a set of structurally effective measures was realised in subsequent years under the umbrella of fFORTE, such as doctoral colleges at two Technical Universities or the scholarship programme DOC-fFORTE of the *Austrian Academy of Sciences* (ÖAW). In 2005, the *Ministry of Labour and Economy* (BMAW) joined the initiative and financed measures under the programme w-fFORTE (short for "economic impulses from women in research and technology" [translation from German[2]]). The programme focused, amongst others, on the promotion of mixed teams for improving the quality of research and innovation (Dorr et al., 2020, Wroblewski & Schaller-Steidl, 2023) and explicitly addressed the aspect of economic growth, which still used to be dominant in the prevailing notion of innovation of that time (see section 3.1).

Around the same time as w-fFORTE, the programme FEMtech was established by the BMVIT under the umbrella of fFORTE. The aim of the programme was to foster the embedment of the gender dimension in research content as well as women's careers in technical and scientific areas. It was divided into three main elements: (1) FEMtech research projects (projects with a gender dimension in their research contents, mainly aimed at awareness raising), (2) FEMtech internships (for young female scientists to gain ground in applied research) and (3) FEMtech career (the programme supports organisations in employing more women in the fields of science and technology, e.g. via the FEMtech Career Check for SMEs; Grasenick et al., 2011). In 2024, FEMtech was continued as "Diversitec"[3], focusing more broadly on aspects of diversity, equality, and inclusion in R&I.

In 2009, the impulse programme "Laura Bassi Centres of Expertise" was established as a lighthouse project in the frame of the BMAW's w-fFORTE programme. Its aim was to address the problem of female underrepresentation, particularly in those areas of research in which research and development (R&D)-expenditures used to be highest, as well as in top positions. The impulse programme was embedded in the w-fFORTE

1 Nowadays Federal Ministry of Women, Science and Research (BMFWF)

2 *"Wirtschaftsimpulse für Frauen in Forschung und Technologie"*

3 For more information see https://www.diversitec.at/

programme and served as a one-time funding pilot initiative. Its main objective was to "highlight excellent female research performance at the intersection between science and industry" (Heckl & Dörflinger, 2014, p. 36). Participation was limited to women in classical areas of R&I.

In 2020, the programme "w-fFORTE Innovatorinnen" was launched, aiming at supporting women in site-relevant R&I in a targeted way, and rendering them more visible. Highly qualified female researchers were encouraged and empowered to develop their ideas, expand their professional networks, and gain greater creative freedom and opportunities for professional growth (Alber et al., 2021). The programme was the Leadership-pilot of INNOVATORINNEN, the empirical focus in this chapter (see section 4).

From the perspective of Schot & Steinmüller (2018) and Wise et al. (2022), all of the above-mentioned innovation policy instruments show signs of frame two-innovation policies: They are or were based on the aim to seize different actors' knowledge, to link these actors and foster their mutual learning – also cross-sectionally. The overarching aim of these policies was to stimulate, support and highlight excellence, and to foster economic growth through the promotion of female researchers. All these programmes underwent evaluations in the past decade (Grasenick et al., 2011, Heckl & Dörflinger, 2014, Alber et al., 2021). It was constated throughout the evaluation studies that all the mentioned programmes enjoyed an excellent reputation, both in Austria as well as internationally (ibid.). Methodologically, the evaluations comprised mixed-methods designs and participatory approaches. Most of them counted on different types of document analyses, (statistical) data analysis, case studies (e.g. of funded projects), quantitative (online) surveys with beneficiaries, qualitative interviews and/or focus groups/workshops, e.g. with experts and/or the owners of the programme.

3. THE INNOVATORINNEN PROGRAMME

The Austrian research promotion programme specifically for application-oriented female researchers, INNOVATORINNEN, initiated by the BMAW in cooperation with the *Austrian Research Promotion Agency* FFG in 2022, emerged from its predecessor programme w-fFORTE (2005-2021), in particular from the pilot "w-fFORTE Innovatorinnen" (2020-2021) – see section 3.3 – and findings of its evaluation. It comprises (1) a so-called "Leadership programme" (more details see below), (2) an alumnae network[4] and (3) the "INNOVATORINNEN Club"[5]. Moreover, resulting from the evaluation subject to this chapter, a fourth line, the "INNOVATORINNEN Lab" was established which took place as pilot in 2024[6].

The programme specifically addresses female researchers, innovators, R&I entrepreneurs, and practitioners regardless of their disciplines, affiliations, or career levels. Its principal aim is to support women in their designing and shaping roles in R&I and to increase their visibility. To enter the Leadership programme, candidates are asked to apply with individual "R&I-missions", for instance establishing themselves in a new research field, realising a new project, exploiting their research results, founding an enterprise, reaching out to relevant stakeholders, or similar. Candidates are supposed to argue the expected economic, societal, or ecological impact of their missions, which is a decisive criterion for being selected into the programme by an independent jury. In the Leadership programme, successful applicants are supported throughout a period of 10 months in working on their missions based on intense exchange with mentors, trainers, and their peer group. An essential element of this work is to identify and interact with relevant (non-academic) stakeholders and communicate own ideas to others. Moreover, a focus is set on activities for personal empowerment and acquiring new innovation and cooperation competences.

For the first round of the Leadership programme starting in February 2022, 18 participants were selected and formed a heterogenous peer group: Successful candidates came from seven (out of nine) Austrian states (*Bundesländer*); 39% were affiliated with private companies, start-ups or were in the process of founding their own enterprises. Around one-third of the participants came from non-university research institutions, and another third from universities. Moreover, the group was characterised by different professional phases and different age groups (in a range between 25 and 55 years), whereby a majority of 56 % ranged between ages 25 and 34. With regards to disciplines, classical topics of applied research were represented, such as wood technology,

4 The alumnae network comprises all former Leadership and Lab-participants; there are dedicated events and trainings for alumnae, and alumnae are regularly involved in activities of ongoing Leadership programme courses.

5 INNOVATORINNEN Club is an open format for female researchers and innovators, offering different types of events and trainings.

6 INNOVATORINNEN Lab aimed at support female researchers in their dissemination and exploitation visions. For more information see e.g. https://www.ffg.at/sites/default/files/2024-02/Leitfaden__INNOVATORINNEN_LAB_final_4.pdf

biomedical analytics, micro-mechanics, material sciences, and digitalisation, but also missions[7] in the area of development cooperation, humanitarian aid, or theatre & digitalisation.

While the framework of the Leadership programme and the alumnae network had been well defined at the onset of the programme in 2022, the INNOVATORINNEN Club was still in development and fed by findings from the evaluation underway. In its current state, it is open to all female researchers and innovators and offers networking events and training.

The results of the evaluation of the first round of INNOVATORINNEN underlined the positive effects of the programme on its participants and their missions (Régent & Ecker, 2024). In the past, the pilot programme w-fFORTE Innovatorinnen had already reached international recognition: It was cited as one out of 15 best practice examples in a study by the German *Stifterverband für die Deutsche Wissenschaft* as a format that fosters "competences for openness and a culture of enabling" (Leimüller et al., 2021).

3.1 INNOVATORINNEN AS TIP INSTRUMENT

Considering current definitions of transformative innovation policy, INNOVATORINNEN aligns with the characteristics of a TIP instrument for several reasons. First and foremost, the programme aims at tackling the important and ongoing societal problem that female researchers and innovators still rarely assume a shaping role in R&I. Therefore, INNOVATORINNEN aims at the explicit promotion of innovation emerging from non-male life realities. The continuing underrepresentation of women in leading roles in science and innovation is a challenge that disadvantages a major proportion of the population (e.g. Wroblewski, 2022, Greussing et al., 2016, OECD, 2016, Klapfer & Moser, 2022, Wisenöcker et al., 2021) and holds far-reaching consequences for society at large: Recent studies suggest that when female scientists have freedom of shaping research, both the contents of and the approaches to research topics change, as was illustrated in the frame of the programme Laura Bassi Centres of Expertise and the w-fFORTE Innovatorinnen programme (see both in section 3.3; Wroblewski & Schaller-Steidl, 2023). As found in the frame of the evaluation study subject to this chapter, survey results from close to 280 respondents suggested that, if female researchers had more decisive power, they would more strongly pursue research projects to solve social and ecological problems and work towards

7 For selected examples of R&I-missions pursued during the Leadership programme, see https://www.ffg.at/content/how-she-did-it

changing work conditions and collaborative practices (Régent & Ecker, 2024, p. 37-59). Enabling women to shape innovative processes and to realise innovation from R&I-outcomes of their interest is thus an important societal effect that is fostered by INNOVATORINNEN.

With a view to pertinent definitions of TIP, Schot et al. (2019, p. 21) describe TIP as instruments that aim at fostering new connections between systems, providing spaces for experimentation, and co-creating solutions for broader socio-technical system change. Unlike traditional programmes that are aimed at promoting women in research and science, INNOVATORINNEN adopts an explicitly interdisciplinary and intersectional approach, bringing together participants from diverse sectors, disciplines, and affiliations to form peer groups. In doing so, it acknowledges that scientific careers are increasingly non-linear and often do not follow the "typical" academic path. The creation of new knowledge, developments, and products increasingly happens at the intersections of research and innovation in other sectors. Many important impulses for tackling the grand challenges and working towards the SDGs come from areas that are not traditionally rooted within the academic sector, such as education, creative industry, or the social sector.

This said, INNOVATORINNEN adopts an approach centred on pursuing research careers guided by individual missions. In contrast to former or comparable programmes, the focus is not primarily on female researchers' individual careers (even though there are effects for participants on the personal level), but on missions with arguably strong and broad impacts. Societal effects of R&I-missions and a sound illustration of the expected impacts are an essential selection criterion for candidates of the Leadership programme; moreover, throughout the programme, a strong focus is given to working out impact pathways (Régent et al., 2023). As participants stated in the evaluation, the mission- and impact-oriented nature of INNOVATORINNEN is a convincing factor of the programme, even for women who are usually not attracted by programmes explicitly addressed to an exclusively female target group (Régent & Ecker, 2024). For the (thematically open) programme in total, this means that much rather than focusing on aspects such as the gender dimension or the excellence of the research, as was the case in the previous programmes (see section 3.3), the broader societal impact is in the front. As it was found in a survey among programme participants in the frame of the evaluation subject to this chapter, most participants were aiming at impacts in the area of health and wellbeing (SDG 3), measures for climate protection (SDG 13), and sustainable consumption and communities (SDG 12) (ibid.).

Moreover, one of the key elements of INNOVATORINNEN is its focus on fostering impactful exchanges between participants and stakeholders from diverse fields relevant to their research. For instance, researchers in health or medical sciences engage with representatives of the health system, facilitating the transfer of their research into practice. The Leadership programme provides participants with structured opportunities to share and refine their ideas through interactions with actors from various sectors, including science, civil society, government, and industry. These exchanges take place in carefully designed co-creation and experimentation workshops. Moreover, Leadership participants are tasked with reaching out individually to relevant stakeholders to discuss and advance their ideas. In this regard, the programme's training extends beyond conventional science communication, equipping participants with the knowledge and tools necessary to translate R&I results into real-world applications. This includes collaborating with practitioners in relevant systems and positioning themselves as experts within relevant fields of application. Co-creation is a cornerstone of the programme's approach, reflected in its emphasis on engaging stakeholders and the integration of dedicated co-creation and experimentation workshops. These workshops not only encourage participants to connect with stakeholders but also provide a structured environment for collaborative exploration and innovation.

Finally, also the FFG is breaking new ground with INNOVATORINNEN: The programme is situated in the strategy department (much rather than in the classical funding administration) which experiments with new formats and target group-specific offerings. The aim of the FFG strategy department is to systematically collect learning experiences for taking up new roles as a funding agency with a view to the implementation of transformative innovation policy instruments. In the area of non-monetary support, the programme INNOVATORINNEN tries to strengthen the impact of R&I for the grand societal challenges and SDGs in a target group-oriented way and with novel networking formats and systemic innovation processes.

While INNOVATORINNEN aligns with many aspects of transformative innovation policy as illustrated in this section, there are also considerations regarding its scope and level of impact that merit reflection. INNOVATORINNEN indeed fosters important societal effects by empowering female researchers, however it does not aim to directly address organisational structures. Strengthening individual agency is a key step towards transformation, but at some point, lasting systemic change may also require complementary measures at the institutional level. This, however, is not within the focus and scope of the programme and can potentially be fulfilled by complementary programmes that aim to influence broader organisational and structural shifts to widen the transformative potential.

Since its pilot, the INNOVATORINNEN programme has been undergoing an accompanying evaluation. As Molas-Gallart et al. (2021, p. 4) state, evaluation of TIP instruments should be integrated as strategic dimension to enhance reflexivity and learning. This aspect can be found in INNOVATORINNEN – it has emerged from its predecessor programme w-fFORTE Innovatorinnen and its evaluation; Furthermore, INNOVATORINNEN fundamentally builds on its accompanying evaluation, comprising collaborative approaches between the evaluators and the programme management. More details to the evaluation can be found in section 3.2.

3.2 EVALUATION OF THE INNOVATORINNEN PROGRAMME

The authors of this article were commissioned with the accompanying evaluation of the programme INNOVATORINNEN. The evaluation was started in May 2022 und continued through December 2023. This way, two entire cycles of the Leadership programme could be accompanied. The evaluation study was built on two modules focusing (1) on the Leadership programme and the alumnae network, and (2) on the INNOVATORINNEN Club. Details on the respective research questions and methodological steps can be seen in Table 2.

<table>
<tr><th>Module</th><th>1: Accompanying evaluation of the Leadership programme cycles 2022 and 2023</th><th>2: Research for the INNOVATORINNEN Club</th></tr>
<tr><td>Research questions</td><td>What are the participants' characteristics (in terms of age, discipline, affiliation, missions, intended impacts, role in projects, care responsibilities, etc.)?
How suitable and effective do participants consider the programme? (e.g. with a view to changes in their own self-perception, confidence and working style, personal progress towards empowerment and their individual mission, experience with peer group, etc.)
Which effects are reported by programme alumnae and how does the network among them evolve?</td><td>What are the characteristics of the INNOVATORINNEN Club-target group?
What do female researchers' life and work realities look like?
How strongly are women involved in the development of new projects and cooperation?
Which support can the INNOVATORINNEN Club provide women?
What would women change in R&I (structures and processes, research topics, target groups, etc.) if they had full decisive power?</td></tr>
<tr><td>Methodological steps</td><td>Quantitative surveys among all participants of the first and second cycle of the Leadership programme (2022 and 2023)
Qualitative interviews with selected participants of the first and second cycle of the Leadership programme with the aim to gain profound knowledge on participants' personal stories linked to their participation in the programme (5 interviews per yearly course)
Open participative observation of selected programme elements
Participation of evaluators in co-creation workshops, co-productive sessions together with programme participants, the owners of the programme, and other external stakeholders with a focus on participants' missions
Focus groups with programme owners, selected participants and alumnae with a focus on programme characteristics and learning points for the programme</td><td>Quantitative survey among former female beneficiaries of FFG-funding (n = 277)
Profound reflection on the underrepresentation of women in R&I and the role of the INNOVATORINNEN programme between the programme owners and the evaluators, resulting in a published book chapter[8]
Reflexion workshop with programme owners and external stakeholders
Qualitative interviews and focus groups with selected users of INNOVATORINNEN Club-offerings</td></tr>
</table>

Table 2 Evaluation of the INNOVATORINNEN programme (Source: Own illustration)

8 Régent et al. (2023)

In total, the evaluation concept comprises a set of traditional (e.g. surveys, interviews) and non-traditional (e.g. co-creation) evaluation methods, including close collaboration between evaluators and programme owners. In the following sections, the analysis of the evaluation concept and practices shall be presented.

4. METHODOLOGICAL APPROACH

Next to providing insights into the accompanying evaluation of the INNOVATORINNEN programme, the aim of this contribution is to assess its evaluation concept and practice against the requirements of TIP evaluations as articulated in relevant literature. The analysis focuses on the accompanying evaluation of INNOVATORINNEN during 2022 and 2023[9], with the results published in early 2024 (Régent & Ecker, 2024). Of particular interest is whether and how the various elements of this evaluation align with TIP evaluation requirements, as outlined in the literature.

The evaluation elements under investigation include the research design, evaluation methods, and evaluation practices. These were analysed using a deductive content analysis framework. The requirements for TIP evaluations proposed by Wise et al. (2022) and other key authors (summarised in Table 1) served as the coding scheme. This scheme emphasises core aspects such as evaluation strategy, the role of evaluation, theory of change, and methodological rigor.

To ensure a systematic and comprehensive analysis, a two-cycle coding approach was applied, following the procedures suggested in Creswell & Creswell (2018). In the first cycle, the data was coded according to the predefined categories derived from the TIP evaluation literature. The second cycle of coding focused on refining these patterns and synthesising them into broader insights, allowing for a nuanced understanding of how the evaluation aligns with or deviates from TIP evaluation standards.

9 The authors of this paper were evaluators of INNOVATORINNEN in 2022 and 2023. Moreover, they evaluated the pilot programme in 2021 and are currently evaluating the second round of the programme (2024-2026). The analysis presented in this contribution is limited to the years of 2022 and 2023.

5. FINDINGS: DOES THE INNOVATORINNEN EVALUATION QUALIFY AS TIP EVALUATION?

This section presents the findings of the deductive content analysis, offering insights into how the evaluation of the INNOVATORINNEN programme aligns with the requirements of TIP evaluations. Table 3 provides an overview of the key results, serving as a concise summary of the analysis. These findings are further elaborated in the subsequent discussion, where each aspect of the evaluation is explored in more detail.

Aspect	Elements of TIP evaluations	INNOVATORINNEN evaluation
Evaluation strategy	▪ Monitoring transformative outcomes and signs of systemic change in real time (behavioural changes, emerging relationships/ activities/constellation/ interactions between actors, evolution of strategic aims) ▪ Informing the direction of the change process	▪ Evaluation as strategic dimension of the programme. ▪ Monitoring "signs of change" via surveys and interviews with participants, alumnae and further representatives of the target groups; participative observation at several instances. ▪ Informing change process through co-creation workshops with focus on participants' missions; workshops and co-authored publications with programme owners.
Role of evaluation	▪ Evaluation integrated as strategic dimension of the programme to enhance reflexivity and learning ▪ Informing strategic choices concerning the programme over the time ▪ Evaluation results used as "food for thought", guide the adjustment of the envisaged transformation path, help to refine the transformation process	▪ (Interim) results are regularly reported and considered 1) throughout the Leadership programme, 2) from one cycle of the Leadership programme to another, 3) for developing the INNOVATORINNEN Club. ▪ Evaluation is considered a strategic dimension of the programme with a view to 1) programme development, and 2) contribution to the participants' missions (co-creation).
Theory of change	▪ Flexible, revisited and refined throughout the evaluation process ▪ Nested approach to assess multiple levels	▪ Evaluation concept is adjusted to programme needs in real-time: 1) Data gathering instruments are created in close collaboration with programme owners, 2) in module 2, methodological steps are used flexibly, depending on the developing state of the INNOVATORINNEN Club.
Methodology	▪ Mixed methods ▪ Participatory approaches in data gathering and analysis ▪ Include a variety of perspectives	▪ Mixed methods and participatory approaches in data gathering and dissemination. ▪ No participatory approaches in data analysis. ▪ External stakeholders' view only rudimentarily considered.

Table 3 TIP evaluation elements in the evaluation of the INNOVATORINNEN programme

5.1 EVALUATION STRATEGY

The programme owners explicitly consider INNOVATORINNEN a “learning programme” based on controlled trial and testing (Alber et al., 2021). Already, the creation of the INNOVATORINNEN programme as such was inspired and co-determined by evaluation results of its predecessor programme (ibid.). In total, INNOVATORINNEN is conceptualised in a way to assure regular feedback of (interim) results to the programme owners with the aim to inform the programme’s further development. To achieve this, workshops between the evaluators and programme owners were held on a regular basis to share the latest developments and results and to discuss their implementation into the further run of the programme underway. This way, the programme was developing in real time based on evaluation results. These referred, on the one hand, to participants’ views on the Leadership programme – their behavioural and interactional changes were monitored in the evaluation (Module 1) through regular surveys, interviews, and open participative observation; on the other hand, the wider perspective of women in R&I, their work and life realities as well as potential obstacles to leadership in R&I were considered (Module 2). The results of both modules were used to further develop the Leadership programme and to develop the INNOVATORINNEN Club.

The INNOVATORINNEN evaluation also appears in line with TIP evaluations intending to inform the direction of the change process. Two elements are particularly noteworthy in this context. First, the evaluators actively participated in co-creation workshops within the Leadership programme. In terms of the evaluation, these workshops served a dual purpose: they provided an opportunity for data collection through open, participatory observation, while also positioning the evaluators as external stakeholders and experts alongside other external contributors. In this role, the evaluators engaged with participants to discuss and further develop their individual missions. This approach allowed the evaluators to contribute their expertise as social and economic scientists and, more significantly in this context, to apply the knowledge gained during the ongoing evaluation. By integrating these insights into a co-creative setting, the workshops facilitated the refinement and advancement of participants’ missions.

Second, the evaluators engaged in reflexive processes with the programme owners to examine the broader issue of women’s underrepresentation in research and innovation (R&I), particularly in industry-related contexts and in top positions. A notable outcome of these discussions was a co-authored book chapter, published in June 2023, which integrated insights gained during the evaluation. This contribution not only advanced the evaluation process, but also enriched the public and academic discourse on the topic.

Both of these aspects are uncommon in policy evaluations and highlight the collaborative, egalitarian approach between programme owners, participants, and evaluators. Hence, this partnership underscores a mutual commitment to learning and co-creation throughout the evaluation process.

5.2 ROLE OF THE EVALUATION

As outlined in sub-section 5.1, the accompanying evaluation played a significant and strategic role for INNOVATORINNEN. In the frame of Module 1, data was gathered from the participants of the Leadership programme via surveys, interviews, and observations regularly. Findings were reported to the programme owners who primarily used them as information source for developing the subsequent cycle of the Leadership programme. Partly, adjustments within the same cycle of the Leadership programme were made.

An even more decisive role of the evaluation could be seen in the development of the INNOVATORINNEN Club (Module 2) – apart from an initial anchor concept, programme owners flexibly designed and adapted the major components of the Club in line with evaluation results. Particular importance was given to the survey with close to 280 female researchers all over Austria (the sample was drawn from women who had received FFG funding in a shaping or leading role in the ten years prior to the survey) which aimed at eliciting their needs with a view to a supportive network under the umbrella of the INNOVATORINNEN Club. In autumn 2022, the Club started with its first events and offerings.

5.3 THEORY OF CHANGE

As discussed in the previous sub-sections, the evaluation concept was implemented in a flexible and adaptive manner. At the onset of the evaluation, the evaluators developed an anchor concept that served as a guiding framework. This concept identified the data required for the evaluation, as well as the methods for its analysis (see Table 2). It is important to note that the development of data-gathering instruments was carried out in close collaboration with the programme owners to ensure alignment with the programme's objectives and context.

Reflecting principles of the „theory of change" in TIP evaluations, the evaluation concept incorporated iterative and adaptive elements to respond to emerging insights and evolving needs. For instance, certain methodological steps, such as the evaluators' participation in co-creation workshops, were introduced mid-process. This adaptation was informed by the realisation that integrating the evaluators into these workshops could provide dual benefits: generating richer data through participatory observation and offering programme beneficiaries valuable feedback based on insights gathered through the ongoing evaluation.

This iterative approach aligns with the theory of change by ensuring that the evaluation not only assesses outcomes, but also actively contributes to achieving the programme's transformative goals. By enabling real-time adjustments and fostering learning among stakeholders, the evaluation process itself became an integral part of driving the programme's mission forward. Such an approach underscores the importance of flexibility and collaboration in TIP evaluations to ensure they remain responsive and impactful.

5.4 EVALUATION METHODOLOGY

Even though containing non-traditional methodological steps, the evaluation methodology applied in the accompanying evaluation of INNOVATORINNEN differs in two aspects from that requested for TIP evaluations. Indeed, the INNOVATORINNEN-evaluation is based on a mixed methods research design (see Table 2) as well as on participatory approaches in data gathering and, as outlined in section 5.1, in the dissemination of results; however, in contrast to the methodological elements of TIP evaluations, the element of data analysis was done in an utterly non-participatory manner by the evaluators.

Moreover, deviations from TIP evaluations can be observed regarding the inclusion of diverse (including external) perspectives. While programme owners', participants' and female researchers' views on a broader scale were included, further external stakeholders, such as representatives of the innovation system, were not considered in this evaluation, apart from representatives of the Austrian BMAW in a reflection and validation workshop (that said, BMAW is the funder of the programme and thus not external in the strict sense).

6. CONCLUSIONS AND LEARNINGS

This paper deals with the Austrian programme for the advancement of female researchers INNOVATORINNEN. The paper argues that INNOVATORINNEN is an example of a potentially transformative innovation policy. It differs from previous related programmes in several respects which are elaborated on throughout the paper. In particular, it aims to enable innovation coming from non-male life realities and thereby fostering societal, economic, or ecological impact. This comprises supporting female researchers in taking a shaping role in R&I and increasing their visibility. Thus, impacts on a personal level are achieved through the programme; however, they are considered a side-effect that results from orientation to individual F&Is missions. That said, INNOVATORINNEN does not aim to directly influence the institutional level. Envisaging definitions of TIP as articulated in recent literature, this paper argues for INNOVATORINNEN, overall, to be an example of a TIP instrument.

The primary aim of this paper was to explore whether the accompanying evaluation of the INNOVATORINNEN programme meets the criteria for a TIP evaluation. A deductive content analysis has shown that most of the requirements of TIP evaluations (see Table 2) can be found in the accompanying evaluation of the INNOVATORINNEN programme: Evaluators and commissioners act as equal partners in a collaboration that is clearly focused on the content-related development of the programme, which is strongly responsive to evaluation results underway. Notably, INNOVATORINNEN can be considered an example of a "learning programme" based on controlled trial and testing (Alber et al., 2021). The evaluation process is characterised by mutual learning and knowledge transfer in both directions, which not only provide a basis for informing and refining the development of the programme, but also for shaping the discourse of empowerment and visibility of female researchers and innovators from an intersectional perspective. An example of the responsive nature of the INNOVATORINNEN programme is the development of the INNOVATORINNEN Lab based on evaluation findings that suggested participants' major interest in the implementation, dissemination, and exploitation of their R&I results. The pilot of INNOVATORINNEN Lab took place in 2024.

Comparing the INNOVATORINNEN evaluation with TIP evaluations, two essential elements stand out in a particular way: (1) The evaluators' participation in co-creation workshops with programme participants, aiming at benefitting participants' missions from knowledge generated in the run of the accompanying evaluation; and (2) reflection processes between evaluators and programme owners with the aim, amongst others, to shape the public discourse on the topic of advancing female researchers. Both elements underline the egalitarian collaboration and mutual learning between evaluators, programme owners, and participants, building up a knowledge-triangle that ultimately enriches the further development of the programme and contributes to its objectives.

In contrast, concerning evaluation methodology, two key elements of TIP requirements were not fully met in the accompanying INNOVATORINNEN evaluation. These elements include the use of a participatory approach to data analysis and the incorporation of a broader range of external stakeholders' perspectives, such as representatives from the Austrian innovation system apart from the funding ministry.

This contribution also provided insights into the evaluation of the INNOVATORINNEN programme as well as its challenges and learning points. In contrast to traditional evaluations, the accompanying INNOVATORINNEN evaluation demanded high flexibility among all involved parties: evaluation experts in the role of independent external knowledge gatherers, the evaluators had to be ready for constant shifts from the original methodological concept and flexibly assessed and implemented evaluation requirements raised by the programme owners. They, in turn, had to demonstrate the same level of flexibility in their programme design and show openness to an evolving and open-ended evaluation process. In addition, they needed to engage with the methodological steps – an area that is not necessarily subject to their work –, while respecting the independent nature of the evaluation. Finally, the quality of the evaluation was in large parts dependent on programme participants' openness and flexibility with regards to their engagement with the evaluation and the evaluators. In total, both evaluators and programme owners were required to create and maintain an atmosphere of trust and mutual respect throughout the entire duration of the evaluation that went far beyond what is required in more traditional evaluations.

This paper aimed to illustrate a case of an accompanying evaluation incorporating several non-traditional elements that align closely with the principles of TIP evaluations. However, the study comes with a set of limitations. First, the historical analysis of programmes aimed at advancing female researchers is confined to the Austrian context, limiting the broader applicability of its findings. Second, due to the qualitative nature of the analysis, it is important to note that the authors of this paper have been directly responsible for the accompanying evaluation of INNOVATORINNEN during its pilot phase, as well as its first and second rounds. Consequently, the analysis presented here should be viewed as a reflective examination of the authors' own work. While this enables the integration of tacit knowledge gained throughout the evaluation process, it also precludes an external or more objective perspective. Finally, this research, based on a single case study, is not embedded in a broader empirical analysis of TIP evaluations. Further empirical research is required to gather diverse examples and practices systematically, which would help to further extend and substantiate the theoretical and empirical foundation of TIP evaluations.

REFERENCES

Alber, C., Dusl, L., Ecker, B., Pohoryles-Drexl, S. (2021). Erfahrungen und Ergebnisse aus der begleitenden Erhebung zum Pilot w-fFORTE Innovatorinnen. URL: https://repository.fteval.at/570/1/w-fFORTE%20Innovatorinnen_Erfahrungen_Pilot_BMDW_FFG_WPZResearch.pdf

Boni, A., Giachi, S., and Molas-Gallart, J. (2019). Towards a Framework for Transformative Innovation Policy Evaluation. Transformative Innovation Policy Consortium (TIPC) Research Report (April 2019)

Carayannis, E. G., Campbell, D. F. J. (2009) "Mode 3' and 'Quadruple Helix': toward a 21st century fractal innovation ecosystem', Int. J. Technology Management, Vol. 46, Nos. 3/4, pp.201–234.

Carayannis, E. G., Campbell, D. F. J. (2012). The Quintuple Helix innovation model: global warming as a challenge and driver for innovation. J. Innov. Entrep. 2012, 1, 2.

Chataway, J., Daniels, C., Kanger, L., Schot, J., and Steinmueller, E. (2017). 'Developing and Enacting Transformative Innovation Policy'. Paper presented at the 8th International Sustainability Transitions Conference, Gothenburg, Sweden, 18–21 June.

Chesbrough, H. (2003). Open Innovation: The New Imperative for Creating and Profiting from Technology. Harvard Business Press)

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

Dorr, A., Heckl, E., Kaufmann, J. (2020). Evaluierung des Förderschwerpunkts Talente. Endbericht. URL: https://repository.fteval.at/id/eprint/549/1/KMU_Evaluierung_Talente_2020_Endbericht_BF.pdf

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix—University-Industry-Government Relations: A Laboratory for Knowledge-Based Economic Development.

Ghosh B., Kivimaa P., Ramirez M., Schot J., Torrens J. (2021) 'Transformative Outcomes: Assessing and Reorienting Experimentation with Transformative Innovation Policy', Science and Public Policy, 48: 739–18.

Grasenick, K., Kupsa, S., Warthun, N. (2011). Evaluierung des Programms FEMtech. Endbericht. URL: https://repository.fteval.at/id/eprint/166/1/2011_Evaluierung%20des%20Programmes%20FEMtech.pdf

Greussing, E., Schott, M. (2016). Austrian University Female Founders Report 2016. WU Wien, URL: https://www.wu.ac.at/fileadmin/wu/d/cc/gruenden/Report_Female_Founders_16.pdf

Heckl, E., Dörflinger, A. (2014). Begleitende Evaluierung der Impulsaktion „Laura Bassi Centres of Expertise". Endbericht. URL: https://repository.fteval.at/id/eprint/87/1/Begleitende%20Evaluierung%20der%20Impulsaktion%20Laura%20Bassi%20Centres%20of%20Expertise_Endbericht.pdf

Joly, P.B., Matt, M. (2022). Towards a new generation of Research Impact Assessment approaches. Journal of Technology Transfer, 2022, 47 (3), pp.621-631. [10.1007/s10961-017-9601-0]. [hal-01784894]

Klapfer, K., Moser, C. (2022). Arbeitsmarktstatistiken. Ergebnisse der Mikrozensus-Arbeitskräfte-Erhebung und der Offene-Stellen-Erhebung. URL: https://www.statistik.at/fileadmin/publications/Mikrozensus-Arbeitsmarkt-2021.pdf

Leimüller, G., Benke, S., Gerbl, B. (2021). Openness in internationaler Wissenschafts- und Innovationspolitik / Was Deutschland lernen kann: https://innosci.de/wp-content/up loads/210617_innOsci_Studie_Openness_interna tional.pdf.

Molas-Gallart, J., Boni, A., Schot, J., and Giachi, S. (2020). A Formative Approach to the Evaluation of Transformative Innovation Policy. Transformative Innovation Policy Consortium (TIPC) Research Report (July 2020).

Molas-Gallart, J., Boni, A., Giachi, S., and Schot, J. (2021) A Formative Approach to the Evaluation of Transformative Innovation Policies, Research Evaluation, 2021: 1–12.

OECD. (2016). Kurzdossier zum weiblichen Unternehmertum. URL: https://www.oecd.org/cfe/leed/Policy%20Brief%20on%20Women%27s%20Entrepreneurship%2DE.pdf

Patton, M. Q. (2006) 'Evaluation for the Way We Work', Nonprofit Quarterly, 13: 28–33.

Peersman, G. (2015). Impact evaluation, in: https://www.betterevaluation.org/themes/impact_evaluation

Régent, V., Alber, C., Ecker, B., Pohoryles-Drexel, S. (2023). Erfahrungen und Ergebnisse aus der begleitenden Erhebung des Programms INNOVATORINNEN, in: Wroblewski et al. (Hg.). Von der Geschlechterpolitik zur diversitätsorientierten Gleichstellungspolitik. Wien

Régent, V., Ecker, B. (2024). INNOVATORINNEN. Begleitforschung des Programms 2022 – 2023. http://repository.fteval.at/id/eprint/705

Schot, J., Boni, A., Ramirez, M., and Alvial-Palavicino, C. (2019). 'Transformative innovation policy and social innovation'. In: Howaldt, J., Kaletka, C., Schröder, A., and Zirngiebl, M. (eds.) Atlas of Social Innovation, ii: A World of New Practices. Munich: oekom Verlag.

Schot, J., Steinmueller, W. E. (2018) 'Three Frames for Innovation Policy: R&D, Systems of Innovation and Transformative Change', Research Policy, 47: 1554–67.

TIPC (Transformative Innovation Policy Consortium) (2019). Guide to: Three Frames of Innovation. http://www.tipconsortium.net/resource/guide-tothree-frames-of-innovation

Weber, K. M. and Rohracher, H. (2012) 'Legitimizing Research, Technology and Innovation Policies for Transformative Change: Combining Insights from Innovation Systems and Multi-Level Perspective in a Comprehensive 'Failures' Framework', Research Policy, 41: 1037–47

Wise, E., Eklund, M., Smith, M., Wilson, J. (2022). A participatory approach to tracking system transformation in clusters and innovation ecosystems – Evolving practice in Sweden's Vinnväxt programme, in: Research Evaluation, 31(2), 2022, 271-287. https://doi.org/10.1093/reseval/rvac006

Wisenöcker, A., Resl, B., Szentgyorgyi, R., Borchardt, K., Seidl, A., Fitzgerald, A. (2021). Frauen in der Medizin 2021. Frauen in Führungs- und Entscheidungspositionen in der Medizin. Was fördert und was hindert die Karriereentwicklung? Karl Landsteiner Institut für Human Factors & Human Resources im Gesundheitswesen. URL: https://www.kli-hr.at/wp-content/uploads/2022/03/Frauen-in-der-Medizin-2021_%C2%A9-KLI-1.pdf

Wroblewski, A. (2022). Leitfaden zur Entwicklung von Gleichstellungsplänen in österreichischen Hochschul- und Forschungseinrichtungen. Studie im Auftrag des Bundesministeriums für Bildung, Wissenschaft und Forschung und des Bundesministeriums für Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie.

Wroblewski, A., Schaller-Steidl, R. (2023). Kulturwandel. in: Wroblewski et al. (Hg.). Von der Geschlechterpolitik zur diversitätsorientierten Gleichstellungspolitik. Wien (forthcoming)

AUTHORS

VERENA RÉGENT
WPZ Research GmbH
Mariahilfer Straße 115/16, 1060 Vienna, Austria
ORCID: 0000-0002-3870-0112

BRIGITTE ECKER
WPZ Research GmbH
Mariahilfer Straße 115/16, 1060 Vienna, Austria
ORCID: 0000-0003-4513-4777

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e7, 1-19
DOI: 10.22163/fteval.2025.699

EVERYONE IS EQUAL IN THE LOTTERY DRUM. MORE OPPORTUNITIES FOR RISKY RESEARCH, POSTDOCS AND FEMALE SCIENTISTS?

DAGMAR SIMON
DOI: 10.22163/FTEVAL.2025.699

ABSTRACT (ENGLISH)

The growing importance of third-party funding for research funding and for the evaluation of research performance has magnified well-known problems of peer review: risk-aversion reviewer overburden, the danger of bias and the Matthew effect ("To him who has shall be given"). The Volkswagen Foundation responded to these problems with its "Experiment!" funding initiative and has funded risky research questions for several years. In addition to peer review, lottery also played a decisive role in the unusual application and selection process. This article presents selected findings from accompanying research, highlighting the initiative's role in exploring new review procedures and creating a space for experimentation that could inspire other funders.

ABSTRACT (DEUTSCH)

Der große Stellenwert von Drittmitteln für die Forschungsfinanzierung und die Bewertung von Forschungsleistungen hat die bekannten Probleme des Peer Review noch einmal erhöht. Dazu gehören insbesondere die Neigung zu einer risikoaversen Begutachtung, die hohe Belastung von Gutachtenden durch aufwändige Verfahren sowie die Gefahren eines Bias und der Matthäus-Effekt („Wer hat, dem wird gegeben"). Die Volkswagen Stiftung hat mit ihrer Förderinitiative „Experiment!" auf diese Problemlagen reagiert und über mehrere Jahre riskante Forschungsfragen gefördert. In dem ungewöhnlichen Antrags- und Auswahlverfahren spielt neben Peer Review auch das Los eine entscheidende Rolle. In diesem Beitrag werden ausgewählte Ergebnisse der Begleitforschung zur Förderinitiative vorgestellt, die einen wichtigen Beitrag zur Erprobung neuer Begutachtungsverfahren geleistet und einen Experimentierraum auch für andere Förderer geöffnet hat.

Keywords: Risky research, partially randomized selection process, peer review, research funding, bias,
Riskante Forschung, teilrandomisierte Auswahlverfahren, peer review, Forschungsförderung, Bias

1. THIRD PARTY FUNDING, THIRD PARTY FUNDING, THIRD PARTY FUNDING !?!

One of the most striking changes in research funding in almost all European science systems – including the German research funding system – over the past two decades is the greatly increased importance of third-party funding both for research funding and as a significant reputational feature for scientists and scientific institutions in evaluations, especially when the distribution of research funds is based on peer review. In a recently published article on the "costs" of third-party funding (Schweiger, Barnett, van den Besselar, 2024), reference is made not only to the "economic costs of competition" but also to the "epistemic costs of competition". Their data show "a negative correlation of 0.3 between efficiency and the degree of competitive funding, indicating that increasing the share of competitive funding tends to reduce the efficiency of the system: a decline of highly cited publications per additional investment in research" (ibid.). In addition, and this is particularly interesting in the context of our study is the "risk-averse" bias of peer review-based decisions in research funding, which can now also be documented in a study on the funding decisions of the European Research Council (Veugelers, Wang, Stephan, 2022)

In addition, the overloading of peer review has now become an internationally observable problem. In addition to the existing reviews, for example of proposals or applications for scholarships and prizes as well as of qualification theses and expert opinions in the context of appointment procedures, third-party funding applications for research projects, study programs, funding programs, research buildings and large-scale equipment as well as minor travel funds are also increasingly being reviewed. Expectations of the reviewers' expertise have continued to grow. New topics such as transfer activities, infrastructures and governance and, last but not least, complex and time-consuming review processes such as those in the Excellence Initiative are being addressed.[1]

Despite all the pressures, peer review continues to be the foundation of scientific assessment. In peer review, the quality expectations and quality standards of the scientific communities are asserted, whose evaluations lend the decisive currency in science, namely reputation. As recognized and without alternative as the procedure is, studies have nevertheless long drawn attention to structural problems and pointed to a lack of agreement between reviewers (reliability) and validity as well as (among other things, gender-specific) bias and the Matthew effect (Neidhardt, 2016). In addition, in recent years in particular, peer review, which is largely discipline-oriented, has increasingly had to deal with inter- and transdisciplinary reviews of journal manuscripts and grant applications and develop criteria and (new) procedures for this (Simon & Knie, 2021).

This article examines how the Volkswagen Foundation's "Experiment!" funding initiative has responded to these problems of review processes, which are now widely discussed in the scientific community and how the funding recipients assess the initiative. Following a presentation of the "Experiment!" funding initiative and the methodological design of the accompanying research, selected problems are discussed: the risk-averse behavior of reviewers, the high burden on researchers due to the increased number of applications, the problem of bias and the Matthew effect in the review process and the topic of diversity. In particular, the question of whether postdocs and female scientists have better chances in the lottery procedure than in peer review is addressed here.

1 The German Council of Science and Humanities (2017) summarizes that "evaluations today are not only in demand for internal use in the scientific community in the sense of the classic functions of quality assurance and filtering (selection and construction), but also for other purposes, such as the orientation of research institutions and universities or their subunits." Ibid. p. 17.

2. FUNDING INITIATIVE "EXPERIMENT!": WHICH PROBLEMS SHOULD BE ADDRESSED?

When the Volkswagen Foundation launched a new funding initiative called "Experiment!" in 2012, the name of this funding line initially alluded to the nature of the topics and issues funded – and a few years later also to elements of the selection process. The aim of the "Experiment!" funding line was to support research projects that dealt with particularly risky and original research questions. Apart from the restriction that the applications should come from the natural sciences, engineering and life sciences, no content requirements or thematic priorities were set. Applicants were expected to already hold a doctorate and be employed at either a university or a non-university research institution in Germany. The people who were finally selected were given a grant of 120,000 euros for a maximum period of 18 months, which could be used flexibly for personnel or material costs.

The Volkswagen Foundation expressly intended to support risky research, which was understood to mean fundamentally new research projects with an uncertain outcome. Its aim was the "exploration of extremely daring research ideas that fundamentally challenge established knowledge, seek to establish unconventional hypotheses, methodologies or technologies or focus on completely new research directions."[2] Unexpected findings and even project failure were accepted as outcomes. Right from the start, all calls for proposals met with a very high level of interest. Of the total of 704 applications in the years 2013–2016, 67 projects were approved. The selection process was fundamentally changed in 2017: In addition to selection by a jury, roughly the same number of applicants were now selected by lot. Since then, the number of people receiving funding has almost doubled, but the number of applications has also continued to rise: In the so-called partially randomized procedure, 117 projects out of 2,748 applications were approved from 2017–2021.[3]

2 VolkswagenStiftung. cf. https://wwww.volkswagenstiftung.de/de/foerderung/foederangebot/experiment-auf-der-suche-nach-gewagten-forschungsideen-beendet, checked on 20.01.2025

3 Information by the Volkswagen Foundation, unpublished.

In addition to the focus on new and risky research ideas, the Volkswagen Foundation also broke new ground in the design of the application and selection process. Applicants were expected to submit a short, maximum three-page text explaining their idea. This idea had to be completely new, so applicants were not expected to have done any preliminary work or even published work by other scientists. In addition, the three-page outline should be completely anonymized. The jury members who later selected the funded projects therefore neither knew the personal data (age, gender, educational and career history, nationality, etc.) nor the scientific institutions in which the applicants had worked in the past or at the time of application. In addition to this proposal, applicants were asked to provide a self-assessment of their project, which was no more than one page in length. External expert opinions were not required. In this way, the above-mentioned risk of a possible bias, which could arise, for example, due to individual personal characteristics of the applicants, their affiliation to a particular institution or their familiarity with the respective scientific community of the jury members, was to be largely excluded. The numerous applications were first checked by the Volkswagen Foundation´s office for minimum standards in terms of content and form before they were discussed in a jury meeting. The jury was made up of internationally recruited researchers and was rather small (eight to ten members) – not least in view of the broad range of subjects and topics in which applications were possible. The jury members were not recruited on the basis of their professional proximity to individual fields or in their capacity as specialists for specific issues, but as generalists for a broad spectrum of new and promising research ideas. The individual jury members remained anonymous, meaning that the application process was double-blind in order to prevent influence as far as possible.

Since the introduction of lottery elements in 2017, the selection process has become significantly more complex: the jury members now not only had the task of selecting the most convincing applications, but were also asked to decide whether the applications for projects that were not initially selected were of high quality and should therefore take part in a selection by lottery. From this pool of all positively assessed applications, further applications were drawn by lot for funding at the end of a jury meeting. Since 2017, around half of the funding recipients have been selected by peer review and half by drawing lots. The “Experiment!” selection process is therefore a partially randomized procedure.

Following the introduction of a lottery procedure in New Zealand (2013), the Volkswagen Foundation is one of the first funding institutions to use a partially randomized procedure to select research projects. The weaknesses of peer review procedures have been criticized for many years, with peer review decisions on publications in scientific journals taking center stage alongside the selection of research proposals. However, it is only in recent years that lottery procedures have also been practiced. The number of funding organizations that have dared to implement these procedures is still manageable. In addition to the Volkswagen Foundation, these institutions include two organizations from New Zealand – the Health Research Council (HRC) and the Science for Technological Innovation (SfTI) – as well as the Swiss National Science Foundation (SNF) and the Austrian Science Fund (FWF). These first experiments with lottery procedures are being received and discussed with great interest in science research, with a particular focus on the questions of whether the various selection procedures are fairer, whether they increase the chances of unconventional research ideas being approved, how scientists assess a selection of their research proposals by lot and whether lottery procedures actually represent an alternative to peer review based on previous experience (Liu, 2020; Barlösius & Philipps, 2021; Osterloh & Frey, 2019; Philipps, 2022; Röbbecke & Simon, 2023; Roumbanis, 2019). In science policy, innovative selection procedures that could contribute to the further development of the review system are welcomed. For example, the German Council of Science and Humanities is in favor of random selection, if it is difficult to justify a decision in the case of heavily oversubscribed funding offers.

3. METHODS OF THE ACCOMPANYING RESEARCH

Some of the key results of the accompanying research[4] for the "Experiment!" funding line are presented below. On the one hand, it dealt with the question of whether, from the perspective of the funding initiative's recipients, it is possible to identify particularly risky research ideas. On the other hand, it aimed to gain insights into the sensible design of partially randomized procedures, their effects and thus also their future use in the scientific system. To this end, online surveys were conducted, for which all grantees from the first four funding rounds with jury decisions (2013 to 2016) were initially contacted in 2018.

4 The accompanying research (2018-2023) took place as part of a third-party funded project by the Volkswagen Foundation. In addition to the author, Martina Röbbecke (Evaconsult), Michael Ploder and Lisa Schön (Joanneum Research) were involved.

The grantees of the following three approval rounds with a partially randomized procedure (2017 to 2019) were contacted approximately one year after being accepted for funding and asked to complete the online questionnaire. A total of 165 people were contacted, 123 of whom took part in the online surveys.

This resulted in a response rate of 75 percent of those funded in the 2013–2019 approval years. The starting point for the design of the questionnaires were literature analyses, a review of the program documents and several exploratory discussions with researchers who had already received funding as part of "Experiment!". The finalized questionnaires were subjected to various pretests with the involvement of selected experts and then revised.

The sample was compiled from the data provided by the Volkswagen Foundation. The results of the online surveys were analyzed in descriptive form. The data were first cleaned and checked for consistency. The collected data were then analyzed for absolute and relative frequencies as well as for cross-correlations and differences based on defined characteristics.

In addition, 37 guided interviews were conducted with selected grantees in the initial phase of their project, including a further retrospective interview with 14 people towards the end of the funding period. They were asked about their understanding of risky research, their assessment of the application and selection process and the effects of the funding. The interviews focused on those researchers who had been selected and funded in the years 2017 to 2019. During the selection process, care was taken to ensure a balanced composition in terms of gender, lottery and jury decisions, postdocs and professors, as well as the subject groups of natural sciences, life sciences and engineering. Around 90 percent of the guideline-based interviews were transcribed and analyzed using qualitative content analysis methods (Mayring, 2015).

4. FUNDING RISKY RESEARCH

Scientific communities play an important role as guardians of (disciplinary) knowledge and quality standards. With regard to the breakthrough of new research ideas and questions – especially those that could mean a paradigm shift in a field of research – a tension is observed with the function of scientific communities (Kuhn 1976, Kuhn 1977).

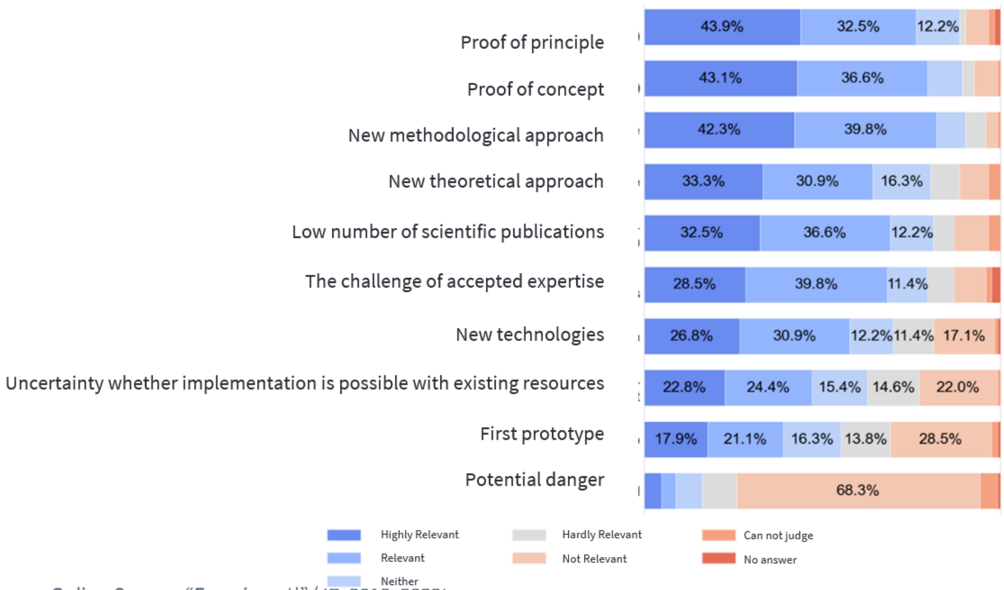
With regard to this tension, one could see a structural dilemma in the fact that, on the one hand, the production of scientific knowledge is oriented towards the “common sense” of knowledge, which is evaluated and assessed by peers, while, on the other hand, the production of knowledge is always dependent on “new”, “original” knowledge (Merton, 1968): “divergent thinking”, “the freedom to go off in different directions” and “rejecting the old solutions and striking out in some new direction” (Kuhn, 1977, p. 226; Kuhn, 1976) increase the chances of discovering new knowledge or understanding, which, however, must be included in the canon of recognized knowledge. In this respect, peer review as the highest evaluation authority is often accused of structural conservatism: What the peers do not know or know how to assess, especially if it even crosses disciplinary boundaries, often has a hard time gaining the necessary recognition. Kuhn points out that normal science often “suppresses fundamental innovations because these necessarily shake its basic positions” (Kuhn, 1976, p. 20), but that “the very nature of normal research guarantees that the new will not be suppressed for very long” (ibid.). Such an anomaly must first be recognized, above all by the paradigm behind it: “The more precise and comprehensive this paradigm is, the more sensitive it is as an indicator for anomalies and thus for a reason for a paradigm shift” (ibid., p. 77).

Recent sociology of science also assumes a tension between new, original knowledge, which can possibly lead to a paradigm shift, and research that sees itself primarily as a further development of the state of art, which cannot be easily resolved: “[The] strategic tension is repeatedly articulated as a dichotomy: in the sociology of science, as reliable ‘succession’ versus risky ‘subversion’ (Bourdieu, 1975) or ‘relevance’ versus ‘originality’ (Whitley, 2000); in the philosophy of science, as ‘conformity’ versus ‘dissent’ or ‘discipline’ versus ‘rebellion’ (Polanyi, 1969); and in the study of innovation, as ‘exploitation’ versus ‘exploration’” (March, 1991). Recent theoretical work supports this broad picture by highlighting the distinctive contributions (Weisberg and Muldoon, 2000) and rewards (Kleinberg and Oren, 2011) associated with traditional versus innovative strategies” (cited in Foster et al., 2015, p. 877).

This tension can affect research funding reviews in different ways. Various analyses on the question of the extent to which the proximity of applications to the reviewers’ research fields – including citing them – has a positive or negative effect on the evaluation (Bourdreau et al., 2016; Li, 2015) come to diametrically opposed conclusions. The findings on new, risky research questions in the funding applications are clear in relation to highly renowned medical funding programs in the USA: “Our second main finding is that more novel proposals are associated with lower evaluations” (Bourdreau et al., 2016, p. 2779).

FIGURE 1: CHARACTERISTICS OF THE FUNDED PROJECT

(relative frequencies per characteristic; Sample n = 123)



Source: Online-Surveys “Experiment!” (JR, 2018–2022)

It can be seen that a high proportion of respondents (over 80 percent) consider new methodological approaches to be particularly relevant for the project. The opportunity to try out new methods was also given high priority in the interviews. In addition, an initial proof of feasibility and an initial proof of principle are attributed high relevance.

According to the funding recipients, the funding initiative opens up the opportunity to actually try out something new that other research funding organizations or other funding programs would not give a chance: The applications would have to be “bent” there so that they succeed in the process.

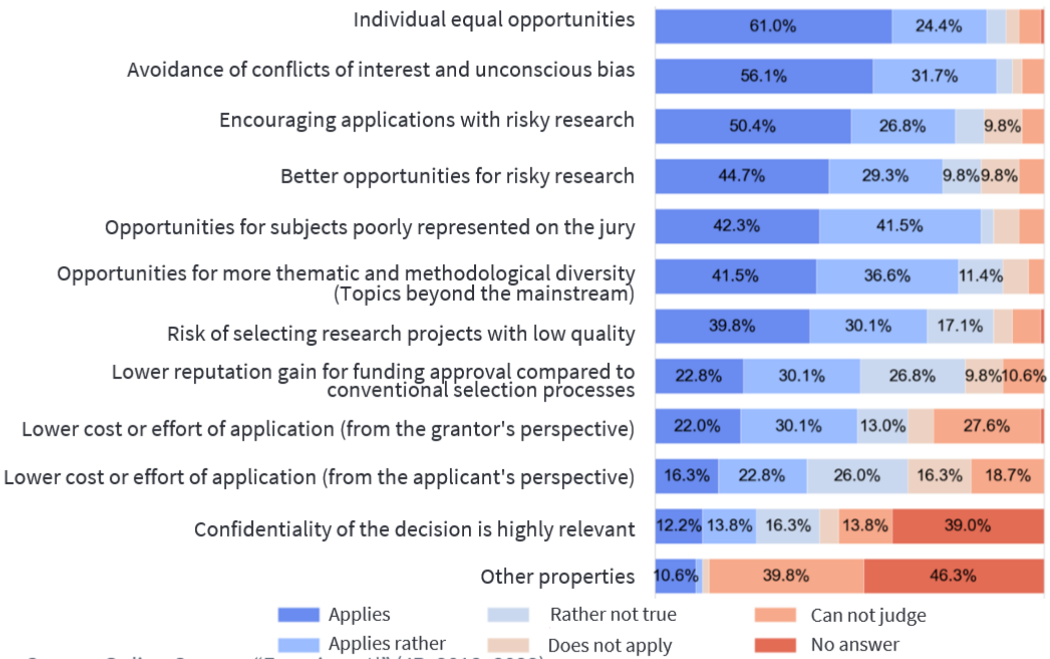
> *“… and that leads to … constructing projects in such a way that they always somehow have a safe component, … Yes, that the research applications do not necessarily correspond to the real intentions of the applicant, I believe. And that you always try to take advantage of the system, so to speak, but actually often apply for something that doesn’t reflect the truth.”[5]*
> *(Senior scientist at a non-university research institution, life sciences)*

5 Translation (ff.) by DS.

In addition to the anonymized selection process, the partially randomized procedure in particular increases the chance of getting unconventional and risky projects off the ground. However, the lottery procedure is not viewed unreservedly positively by the grantees themselves; the online survey shows a mixed picture (see Figure 2). On the one hand, many respondents agree with the assessment that lottery procedures help to avoid conflicts of interest (88%), promote equal opportunities for individuals (85%), encourage applications with risky research (77%) and offer better opportunities for risky research (74%). The fact that lottery procedures also offer opportunities for subjects that are poorly represented in the jury (84%) and for more thematic and methodological diversity (78%) is also rated positively. On the other hand, respondents were critical of the fact that selection by lot could result in a lower reputation gain (53%). More than two-thirds of respondents also fear that lottery procedures could lead to the selection of lower quality research projects (70%).

FIGURE 1: HOW IS LOTTERY ASSESSED?

(relative frequencies per characteristic; Sample n = 123)



Source: Online-Surveys “Experiment!” (JR, 2018–2022)

A mixed assessment of lottery procedures was also evident in the interviews. For the majority of interviewees, a positive assessment of lottery procedures clearly prevailed, not least in view of the weaknesses of peer review procedures. A lottery procedure is particularly suitable for researchers in an early career phase who are not yet well established in the scientific community:

> *"… I also think it's great that this is being established, in order to minimize the bias. It simply has the advantage that (…) – of course, if certain criteria are met – you have a certain chance of getting it. So that also reduces this bias with regard to the promotion of established professors to all opportunities (…) that contributes enormously."*
> *(Professor at a university, life sciences)*

However, some funding recipients regretted that lottery procedures do not allow for personal discussions with the jury members and that there is no expert feedback from the reviewers on the project applied for:

> *"Well, there is also the possibility that you were positively evaluated or (…) got the money through a lottery. (…) I would think it would be nice to find out afterwards. (…) Simply so that you can assess whether – whether the project would have had a chance under normal funding conditions (…) or simply whether the evaluations were quite poor and the lottery procedure then led to the goal. So when developing ideas, as a scientist you don't have that many opportunities to get feedback, honest feedback. (…) Accordingly, I think it would be good to get the evaluation, at least of the proposal."*
> *(Professor at a university, life sciences)*

At the same time, in the interviews, the funding recipients made it clear how possible quality deficiencies could be countered and under what conditions a lottery procedure could be applied. They highlighted two aspects in particular: The quality check of the applications received by the Volkswagen Foundation and the assessment by the jury – as practiced in the "Experiment!" funding initiative:

> *"Definitely positive. So, if it's done like this (.) well, the pure random element wouldn't be good, I don't think, because then you could write something and it would just (.) otherwise it would be pure lottery, but the way the VW Foundation has done it, i.e. pre-screening and then random, and then additionally a jury, so that you really (.) so this half/half, I thought that was very good, so on the one hand you can ensure that there are applications that are actually considered great by experts in the field, that they have a high probability of getting through, but on the other hand that there are applications that are perhaps considered too exotic by the panel of experts, that they also have another good chance. But then, of course, pre-screening is very important."*
> *(Professor at a university, natural sciences)*

Another advantage for the assessment of risky research is seen in the composition of the jury: a small international and interdisciplinary group that is responsible for three major scientific fields – life sciences, natural sciences and technical sciences. Against this background, the jury must focus on overarching questions such as whether the research project is both risky and feasible. In principle, reviewers are considered capable of assessing this.

> *"Yes, that can work. Because I think we are more or less trained to recognize whether an idea is innovative in principle." (Professor at a university, life sciences)*

In addition, another advantage of the special jury composition is seen in the reduced risk of jury members being too close to the applicants in terms of expertise in small or emerging research fields. This type of assessment and jury constellation can therefore also help to counteract bias.

5. SCIENTISTS UNDER PRESSURE WITH APPLICATIONS FOR THIRD-PARTY FUNDING: A LEAN SELECTION PROCESS

As already mentioned, scientists are increasingly under time pressure due to third-party funding applications. This is particularly problematic for postdocs with fixed-term contracts, as a follow-up contract can often only be realized by acquiring third-party funding. The increased financing of research through third-party funding is a worldwide phenomenon and in some cases the volume of third-party funding in other countries is significantly higher than in Germany. In 2012, for example, it was estimated for the Australian science system that researchers had to spend a total of 550 years preparing 3,723 third-party funding applications for the National Health and Medical Research Council (NHMRC), of which only 21 percent were funded. In 2013, the funding rate even fell to 17 percent (Herbert et al., 2014, p. 2). In their study, Herbert et al. also point out the negative effects for applicants in terms of family and health burdens, especially if an application can only be submitted once a year. In the meantime, other funding organizations such as the National Institute of Health in the USA, the Engineering and Physical Sciences Research Council in the UK and the Canadian Institutes of Health Research have made efforts to simplify the application and review process and thus reduce the effort involved (ibid.).

It is therefore not surprising that the application and selection process for the "Experiment!" funding line was predominantly rated very positively by the funding recipients (see Figure 3). The highest level of approval was given to the effort involved in submitting an application, with 89 percent of respondents saying they were very satisfied, and 10 percent satisfied. The comprehensibility of the funding guidelines was also rated positively (83% very satisfied, 17% satisfied). It was also emphasized in the interviews that the time required for the short application was pleasingly low. In this context, numerous interviewees referred to proposals submitted to other funding organizations and in particular to proposals submitted to the German Research Foundation (DFG), which require considerably more time to prepare. On the one hand, it is difficult to combine the time-consuming preparation of research proposals with other tasks in research, teaching and self-administration:

> *"I think the nice thing about the experiment proposal was that it wasn't a 20-page proposal. And I mean, of course research funding also has to be competitive, but sometimes that also leads to a huge waste of resources, at least that's my feeling … So I think I'm quite well funded, but I only have a certain funding quota, … that doesn't always work for me either and so it takes three, four, five applications before one is approved, and then they can only be recycled to a limited extent. In this respect, I thought this aspect of the experiment was excellent, that the cost-benefit ratio made sense, so to speak."*
> *(Professor at a university, life sciences)*

On the other hand, it was emphasized in the discussions that short applications that focus on the research idea are much better suited to funding risky research than those application formats in which the chosen methodology, the expected result and the required time frame, including milestones, must already be set out in detail:

> *"And what is also very pleasant about this concept is that you don't write a 50- to 100-page application, but the idea and the risk of the idea count, and in this respect the description of the idea is in the foreground … which is really very difficult to fulfil in many project applications, especially if you want to do something innovative like this."*
> *(Professor at a university, life sciences)*

Other interviewees emphasized more strongly that in other funding procedures, preliminary work must also be presented, and publications must be proven when submitting an application. This is not only difficult in the case of a completely new research idea, but also means that not only the research project, but also the respective person in their academic environment is assessed:

> *"... yes, I often have the feeling, well, how should I put it, that the person submitting the proposal is also assessed. And in my eyes, for example at the DFG ... if I don't have ten great papers, I think I'll have problems getting the project through."*
> *(Professor at a university of applied sciences, engineering sciences)*

In addition to the well-known advantages of a lean application procedure – such as limited time expenditure – many grantees also emphasized that such a procedure is particularly appropriate for risky research ideas.

6. BIAS AND MATTHEW EFFECT: ANONYMIZATION OF APPLICANTS AND LOTTERY PROCEDURE

As mentioned above, research on peer review addresses, among other things, the problem of bias in the assessment of research proposals where the applicants have not been anonymized, which is the norm in research funding. In particular, a gender-specific bias has been proven in studies.[6] Another phenomenon in this context is the Matthew effect, a term coined by the sociologist of science Merton (1968) with regard to the citation frequency of scientific publications, in which he pointed out that successes (for example in research funding) can be explained by previous successes and less by current achievements. The reference to previous successes is particularly evident in research funding applications in the publication lists and other achievements

6 Cf. Wenneras & Wold, 1997; Kaatz et al., 2015. In a text analysis, Kaatz et al. evaluate the responses of funded and non-funded scientists from the renowned R01 program of the National Institutes of Health: Gender stereotypes lead evaluators to give a woman greater praise than a man for the same performance ... Paradoxically, gender stereotypes also lead reviewers to require more proof of ability from a woman than a man prior to confirming her competence, and greater proof to confirm men's incompetence in male-typed domains. This may also explain why men's versus women's proposals were funded despite more negative critiques" (ibid. S. 73/74).

such as third-party funding in competitive procedures, which enjoy a high reputation in most scientific systems. These applicants therefore have an advantage when it comes to application approval.[7]

In the online survey, the anonymization of applications was overwhelmingly welcomed – 73% were very satisfied and 8% satisfied (see Figure 3). The high level of approval for completely anonymized research proposals also underlines the fact that the associated waiver of the possibility of submitting further publications or expert reports is by no means seen as a disadvantage by the funding recipients. The anonymization of the applications – i.e., the double-blind procedure – had the additional important effect from the perspective of the grantees that it was primarily the project that had to be convincing and not the publication lists of the applicants.

> *"... one positive thing (is) ... that it is assessed independently of the CV, and so it has a little less bias towards the establishment and various established (...) structures within the academic system .... I think it's very good that this is done, because it's really only the idea that counts in the end." (Junior professor at a university, life sciences)*

It can therefore be assumed that the anonymization of applicants contributes to greater diversity. Specifically, another element of "Experiment!", the lottery procedure, shows that effects on diversity with regard to age, career stage and gender ratio can be recognized and can therefore counteract a possible bias in the review process. For this purpose, the cohort was compared whose projects in the first four years of the "Experiment!" funding line (2013 to 2016) were selected exclusively by the jury with the funding cohort from 2017 onwards, in which the partially randomized procedure was introduced.

A combined analysis of the age and gender of the grantees is revealing. The absolute number of eight women who were funded between 2013 and 2016 is very small (see Table 1). Nevertheless, the comparison shows that the proportion of women has increased since 2017, in particular the participation of established female researchers (over 50 years of age) and young female postdocs (under 35 years of age) has increased. Overall, the proportion of funded persons under the age of 39 has increased in the partially randomized procedure. The proportion of younger women has risen from around 13% to around 37% and the proportion of younger men has risen from around 33% to around 54%.

7 In a study on the evaluation of the Excellence Initiative, it became clear that the CVs of the Principal Investigators played a decisive role in the evaluation of the Cluster of Excellence applications (Möller et al., 2012).

FIGURE 3: AGE AND GENDER RATIO BEFORE AND AFTER THE INTRODUCTION OF THE PARTIALLY RANDOMIZED PROCEDURE

Cohorts 2013 - 2016					
	Over 50 Years	40 - 49 Years	35 - 39 Years	Under 35 Years	No Information
Female (n = 8)	0,0%	62,5%	12,5%	0,0%	25,0%
Male (n = 40)	27,5%	32,5%	27,5%	5,0%	7,5%
Cohorts 2017 - 2019					
	Over 50 Years	40 – 49 Years	35 - 39 Years	Under 35 Years	No Information
Female (n = 19)	21,1%	26,3%	10,5%	26,3%	15,8%
Male (n = 56)	17,9%	21,4%	39,3%	14,3%	7,1%

Relative frequencies within gender

Source: Online-Surveys„Experiment!“ (JR, 2018-2022)

Furthermore, the introduction of the partially randomized procedure has led to a slight change in the career stages of those funded. The proportion of researchers who hold a professorship has remained almost the same (reduction from around 40% to around 39%), while the proportion of researchers in early career phases (postdocs and junior professorships) has increased slightly (from around 42% to around 47%).

Even if the case numbers are relatively small, it can therefore be concluded that the partially randomized procedure does result in shifts in favour of younger female scientists. Randomized procedures therefore have a certain potential to counteract an age and gender bias.

7. MORE EXPERIMENTS!

The experiment “Experiment” has shown that partially randomized procedures are widely accepted, particularly with regard to risky research and that they also have a positive effect on those receiving funding: Postdocs and scientists are more strongly represented than in the peer review process. These results are significant in view of the high importance of third-party funding, which will not decrease significantly in the foreseeable future.

With “Experiment!”, the Volkswagen Foundation has made an important contribution to the introduction and trialing of new selection procedures, thus opening up a space for experimentation for other funding bodies as well. There is no doubt that the high recognition of the Volkswagen Foundation in the scientific community has also contributed to the willingness of the scientific community to deal more intensively with the limits of peer review and new selection formats. The funding organizations have become more courageous: For example, after a pilot phase, the SNF in Switzerland has now offered the option of a lottery procedure for all funding programs, and the British Academy is using partially randomized selection for smaller funding projects (up to £10,000) in the social sciences and humanities. The Danish Novo Nordisk Foundation is also experimenting with this and with the anonymization of applications in some funding lines.

In general, a certain openness towards new funding formats, which may also contain experimental elements, can be observed in European research funding systems. This is supported above all by the fact that a variety of formats and orientations of research funding is conducive to fairer participation opportunities for applicants, as this means that deficits of one funding format can be compensated for by others (such as problems of peer review through partially randomized procedures). In addition, there are increasing signs that, in research evaluations among other things, a concept of quality that understands excellent research primarily as research whose quality can be measured by the number of publications in international refereed journals is being relativized and that different dimensions of quality can come into play (cf. Watermeyer et al., 2018; Muhonen et al., 2020) This trend is also related to the greater consideration of the social impact of research as well as interdisciplinary and transdisciplinary research, which represent a further reason for more diversity and experimentation in research funding.

REFERENCES

Barlösius, E. & Philipps, A. (2021). Verlosung von Forschungsgeldern: Welche Losverfahren können sich Wissenschaftlerinnen und Wissenschaftler vorstellen? Qualität in der Wissenschaft, 15(3/4), p. 67–72.

Bourdreau, K. J., Guiman, E. C., Lakhani, K. M. & Riedel, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. Management Science, 62(10),p. 2765–2783.

Foster, J. G., Rzhetsky, A. & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. American Sociological Review, 80(5), p. 875–908.

Herbert, D. L., Coveney, J., Clarke, P. et al. (2014). The impact of funding deadlines on personal workloads, stress and family relationship: a qualitative study of Australian researchers. BMJ Open, 4:e004462. doi: 10.1136/bmjopen-2013-004462.

Kaatz, M., Magua, W., Zimmerman, Carnes, M. (2015). A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators a one institution. Academic Medicine, 90(1), p. 69–75.

Knorr Cetina, K. (1984). Die Fabrikation von Erkenntnis. Zur Anthropologie der Wissenschaft. Frankfurt/M.

Kuhn, T. S. (1976). Die Struktur wissenschaftlicher Revolutionen. Frankfurt/M.

Kuhn, T. S. (1977). Die Entstehung des Neuen – Studien zur Struktur der Wissenschaftsgeschichte.

Li, D. (2015). Expertise vs bias in evaluation: Evidence from the NIH. HBS Working Paper 16-053. Boston.

Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T. & Pomeroy, L. (2020). The acceptability of using a lottery to allocate research funding: a survey of applicants. Research Integrity and Peer Review, 5(3). https://doi.org/10.1186/s41073-019-0089-z.

Mayring, Philip (2015). Qualitative Inhaltsanalyse: Grundlagen und Techniken. (11. Aufl.), Weinheim und Basel 2015.

Merton, R. K. (1968). The Matthew Effect in Science. Science, 159(3810), p.56–63.

Möller, T., Philipp, A., Hinze, S. & Hornbostel, S. (2012). Exzellenz begutachtet. Befragung der Gutachter in der Exzellenzinitiative. iFQ-Working Paper No. 11.

Muhonen, R.; Benneworth, P. & Olmos-Peñuela, J. (2020). From productive interactions to impact pathways: Understanding the key dimensions in developing SSH research societal impact. Research Evaluation, 29(1), p. 34–47.

Neidhardt, F. (2016). Selbststeuerung der Wissenschaft: Peer Review. In D. Simon et al. (Hrsg.), Handbuch Wissenschaftspolitik, Wiesbaden, p. 261–277.

Osterloh, M., Frey, B.S. (2019). Dealing with Randomness. Management Revue, 30, p.331–334.

Philipps, A. (2022). Research funding randomly allocated? A survey of scientists' view on peer review and lottery. Science and Public Policy, 49, p.365–377.

Röbbecke, M. & Simon, D. (2023). Riskante Forschung und teilrandomisierte Begutachtungsverfahren: Neue Wege der Förderlinie „Experiment!“ der Volkswagen Stiftung. In: Beiträge zur Hochschulforschung 2/2023, p. 8–31.

Roumbanis, L. (2019). Peer Review or Lottery? A critical analysis of two different forms of Decision-making Mechanisms for Allocation of Research Grants. Science, Technology, Human Values, 44(6), p. 994–1019.

Schweiger, G., Barnett, A., van den Besselar, P. (2024): The costs of competition in distributing scarce research funds. In: PNAS Direct Submission, December 2, 2024, 121(50), e2407644121.

Simon, D. & Knie, A. (2021). Vom Libero zur Viererkette? Eine Neubewertung transdisziplinärer Forschung in der akademischen Wissenschaft. In J. Herberg, J. Staemmler & P. Nanz (Hrsg.), Wissenschaft im Strukturwandel. Die paradoxe Praxis engagierter Transformationsforschung, München, p. 63–82.

Veugelers, R., Wang, J., Stephan, P. (2022): Do funding agencies select and enable risky research: Evidence form ERC using novelty as a proxy of risk taking; National Bureau of Economic Research, Tech. Rep.

Watermeyer, R. & Chubb, J. (2018). Evaluating ‘impact’ in the UK’s Research Excellence Framework (REF): liminality, looseness and new modalities of scholarly distinction. Studies in Higher Education, 44(9), p. 1554–1566.

Wenneras, C. & Wold, A. (1997). Nepotism and Sexism in Peer-Review. Nature, 387, 341–343. http://dx.doi.org/10.1038/387341a0.

Wissenschaftsrat (2017). Begutachtungen im Wissenschaftssystem. Berlin. Drs. 6680-17.

AUTHOR

DAGMAR SIMON
Geschäftsführerin Evaconsult GbR
Emserstr. 22, 10719 Berlin
Gastwissenschafterin der Forschungsgruppe Digitale Mobilität und gesellschaftliche Differenzierung am Wissenschaftszentrum Berlin für Sozialforschung (WZB)
Email: dagmar.simon@wzb.eu

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e8, 1-20
DOI: 10.22163/fteval.2025.700

ENERGY AND RESOURCE EFFICIENCY IN THE ECONOMY: THE EVALUATION OF GERMANY'S LARGE INDUSTRIAL FUNDING PROGRAMME USING MIXED METHODS

LISA NEUSEL AND SIMON HIRZEL
DOI: 10.22163/FTEVAL.2025.700

ABSTRACT

In Germany, the Federal Funding Scheme for Energy and Resource Efficiency in the Economy (EEE) is a central financial support scheme that aims to promote energy and resource-efficient technologies and processes in companies. Due to reporting requirements both, with regard to spending public budgets, as well as progress reporting towards the energy and climate goals, annual evaluations of the EEE are required. These evaluations include a substantial set of quantitative indicators. The indicators are analysed using a common methodology drawing on administrative data as well as survey results among beneficiaries of the support scheme. The paper illustrates the quantitative evaluation approach of the EEE in a two-fold way: First, it outlines the mixed methods approach underlying the evaluation which follows a methodological framework of nine steps. Second, it emphasises three methodological issues arising from recent modifications of the EEE and its framework conditions, which have neither been discussed methodologically nor content-wise yet.

The experience from five consecutive years of evaluation of the EEE plus the evaluation of the EEE's predecessor shows that the overall methodology

ensures that the process is carried out uniformly. This helps to ensure a basis for comparability of broad multi-measure funding schemes, such as the EEE. Yet, it can be observed that there is a constant need for refinement and adaptation to changes e.g. due to changes in external conditions, shifting interests or new design elements. Therefore, the paper outlines three of the most recent methodological issues in more detail. These underline (#1) the need to be transparent about how the dynamic decarbonisation of the energy system is taken into account in the impact assessment, (#2) the need to make conscious decisions on how to consider resource efficiency in GHG accounting and (#3) that, when using funding efficiency as a criterion for the design of such instruments, the context of these values has to be sufficiently appreciated.

Keywords: energy efficiency, resource efficiency, industrial funding, energy policy, climate change, evaluation

BACKGROUND AND AIM

Energy use is a very substantial source of anthropogenic greenhouse gas (GHG) emissions. In consequence, improving the deployment of renewable energies and energy efficiency measures is crucial to limit global warming to the 1.5°C climate target of the Paris Agreement. As a response to this, energy policymaking uses a variety of regulatory, informational and financial measures to enhance the uptake of corresponding action. Publicly funded subsidy schemes aimed at promoting energy efficiency in companies serve as an important cornerstone in many countries.

In Germany, a central scheme is the Federal Funding Scheme for Energy and Resource Efficiency in the Economy (EEE). It aims to specifically promote energy-efficient technologies and processes, available on the market to support companies in improving energy efficiency. This multi-measure scheme is structured into six modules and offers grant-based, credit-based and competition-based subsidies (Figure 1). In its original setup at its initiation in 2019, it covered support for investments in four modules: 1) energy-efficient cross-cutting technologies, 2) process heat from renewable energies, 3) measurement and control equipment, sensors and energy management software and 4) energy optimizations of plants and processes. In 2021, the last module was extended to also cover resource efficiency. Furthermore, two new modules have been added in 2021 and 2023, focusing on 5) transformation plans and 6) electrification in micro and small enterprises.

According to its latest amendment, the EEE aims at facilitating the implementation of energy and resource efficiency measures in companies, thereby targeting saving of 35 TWh of final energy and 19 million tons of carbon dioxide emissions from 2022 until the end of 2028 (BMWK 2024b, 2024a). In 2023, the EEE's subsidies exceeded 1 billion Euros for the first time.

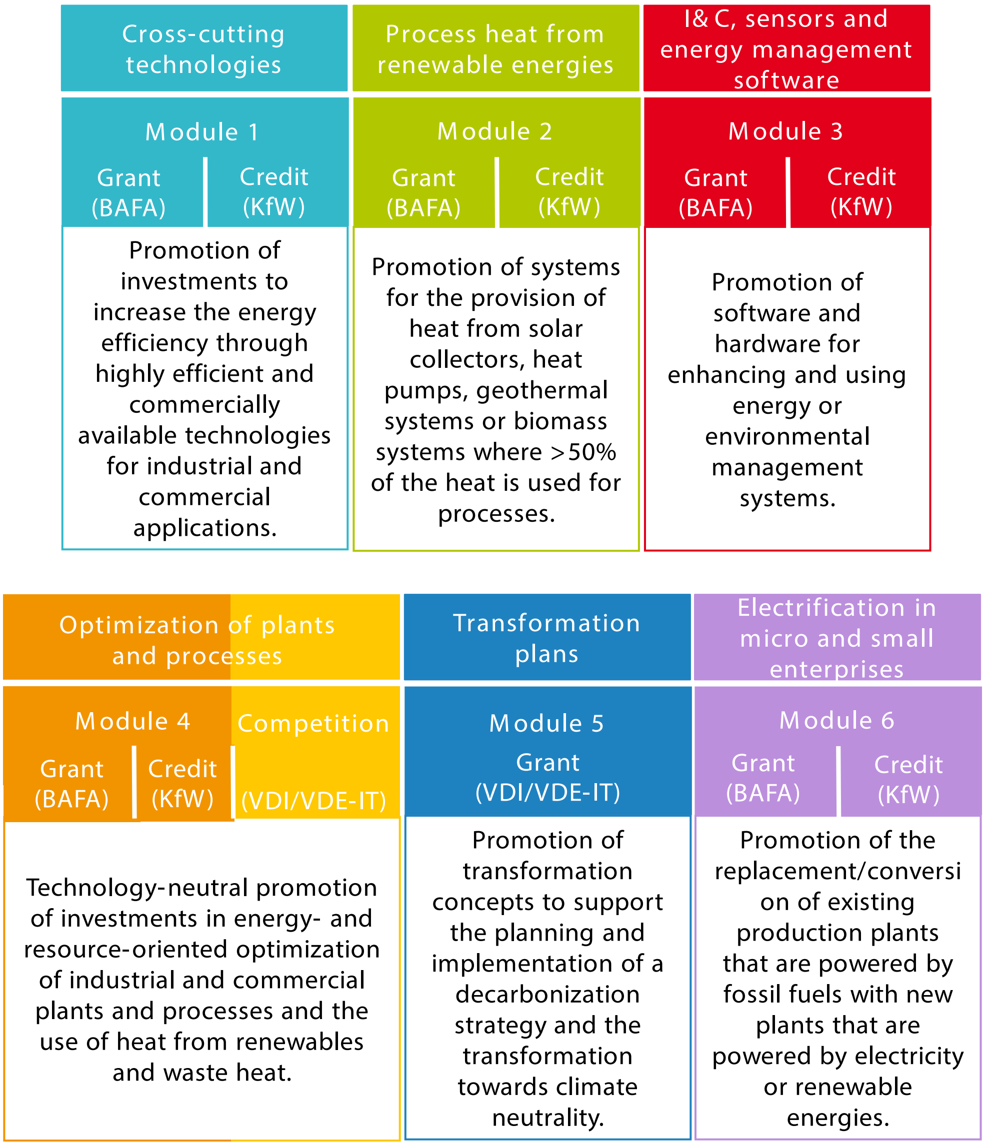


Figure 1: Overview of the architecture of the EEE (source: Neusel et al. 2024a).

Since such schemes as the EEE use public money, ex-post evaluations are regularly required to review their efficiency and effectiveness. Also, reporting requirements on measures addressing European and national energy efficiency and climate targets have increased considerably in recent years. On the European level, the most detailed ones are requirements for the communication of measures and methods for the implementation of Article 8 of the recast Energy Efficiency Directive (Directive 2023/1791/EU, Annex V).

On the level of Germany's national energy and climate targets, an overall goal of achieving GHG neutrality in 2045 is legally required by the revised Federal Climate Change Act (KSG) of 2021 and the Energy Efficiency Act (EnEfG) of 2023. Both also include quantitative reporting requirements in several places.

This contribution illustrates such a quantitative evaluation: First, it outlines the mixed methods approach underlying the evaluation, which follows a methodological framework of nine steps. Second, it emphasises three methodological issues which arise from recent modifications of the scheme and its framework conditions. These have neither been discussed methodologically nor content-wise yet.

EVALUATION METHODOLOGY

The evaluation of the EEE covered five annual rounds between 2019 and 2023. It is based on a methodological framework formalised by Schlomann et al. 2020 within the EEE's predecessor programme (Hirzel et al. 2019), as well as on previous expertise in energy policy evaluation. Table 1 provides an overview of prior publications related to the EEE and its underlying methodology. The purpose of this methodology is to:

- Monitor target achievement: To what extent were the objectives of the funding achieved?
- Assess impact: Is the funding the cause of the impact or suitable for triggering it?
- Control efficiency: Are both the funding provided (efficiency of implementation) and the objectives achieved (efficiency of measures) in an economical manner?

Table 1. Overview of prior publications related to the EEE and the underlying methodology

Publication	Topic
Schlomann et al. 2017	A first outline of the general methodology and its application
Voswinkel et al. 2018	An analysis of the German waste heat programme within the Fund
Voswinkel 2018	An account of eight ways for determining energy savings in evaluations
Voswinkel 2019	Catches in evaluations of multi-programme schemes like the Fund
Voswinkel 2020	Shares the experiences with an overview of an unified harmonisation methodology
Hirzel et al. 2022	Overall impact of the Fund and aggregation issues in multi-measure schemes
Hirzel et al. 2022	Comparison of the classical vs. the competitive funding line of the EEE
Brunzema et al. 2022	Ex-ante impact evaluations by the example of the EEE
Weinert et al. 2024	Resource efficiency as a new funding element in the EEE
Hirzel et al. 2024	Funding of sensors, measurement and control equipment as part of the EEE
Neusel et al. 2024	Evolution and impact of the EEE

The overall methodology consists of nine steps (Table 2). After its application for the year 2019, it was continuously refined to address methodological issues identified for the EEE. Using the EEE as an example, **selected steps** are detailed below to illustrate the methodology.

Table 2. Overview of the evaluation methodology (source: Hirzel and Schlomann 2022).

Step and purpose	Main tasks
1: Characterisation Description of the covered policy measures	▪ General **outline of the measure** covering its type, target group/sectors, budget, funding bodies/ implementation agencies, legal basis, related policy measures and the funding process ▪ Analysis of the **impact model** of the measure ▪ Consideration of potential **distortions** (e.g. overlaps, double counting, side-effects such as free-rider effects, spill-overs or follow-up-effects)
2: Framework data Definition of common data and assumptions	▪ Definition of **harmonised input data** (e.g. emissions factors, primary energy converters) ▪ Provision of **default choice lists** (e.g. lifetimes by type, energy prices)
3: Targets and requirements Identification of the targets of the overall programme, its policy measures and the specific requirements for the evaluation	▪ Description of **requirements and expectations** for the evaluation ▪ Analysis of **top-down targets** for energy efficiency improvements based on governmental documents, directives and laws ▪ Analysis of **bottom-up targets** of individual support schemes from ex-ante estimation, funding guidelines ▪ Definition of the **main areas of interest** for the evaluation
4: Indicators Setting up performance values to measure the achievement of targets	▪ Selection of **indicators that reflect progress** in the areas of interest ▪ **Operationalisation** of the indicators: choice between qualitative/quantitative type, description and delimitation, computational model, type of result, units (quantitative) or scales including interpretation rules (qualitative)
5: Data collection Identification and collection of data for establishing the indicators	▪ Establishing a **data collection concept** based on the selection and setup of the indicators ▪ Implementation of the **data collection process**
6: Data review Processing incomplete or missing information	▪ Review of data (e.g. error correction, missing parameters) ▪ Method selection and implementation of backcasting and proj**ections** of data where needed
7: Data analysis Processing of the data to measure the achievement of the target values	▪ Selection of appropriate **method of analysis** (descriptive/analytical) ▪ Computation of **gross values** for indicators
8: Net impact estimation Eliminations of distortions in the results	▪ **Identification of distortions** ▪ Computation of undistorted **net values** for indicators ▪ Conclusions for the **individual measures**
9: Overall assessment Merging individual results	▪ Determination of **areas for aggregation** and comparison ▪ **Correction for double counting** when aggregating quantitative values ▪ Computation of the **overall assessment** ▪ Formulating conclusions for the **entire scheme**

One of the initial steps, which is a part of the characterisation of the evaluated policy measures, is the development of an impact model (**step 1**). The impact model is a logical causal chain and deliberate simplification of the influences to make impact relationships manageable in the evaluation. The basic impact model follows an input-output-outcome-impact-logic: The input depicts the effort put into the program, and the output reflects the immediate result, the outcome of the content-wise changes and the impact the final result of the intervention on the level of the overall aim. For each of the six modules of the EEE, a specific impact model is used to investigate the individual impact. Figure 2 illustrates such an impact model for Module 3 in the EEE: Module 3 is a support program for measuring and controlling equipment, including support for software and training. As **input**, support is provided in three areas: hardware, software, and training. New sensors and control systems are implemented to enhance data collection, complemented by available software fully connected to this data. Comprehensive training ensures staff can effectively utilise the new systems. The **output** includes the deployment of the new sensors, established software, and trained personnel. In the **outcome**, automated data collection in daily operations is emphasised. Analysed data leads to optimised processes and identifies areas for corrective measures, initiating new strategies to enhance efficiency. Finally, the impact model depicts the logic of the overall intervention, highlighting energy savings as a key impact resulting from optimised processes and enhanced staff capabilities.

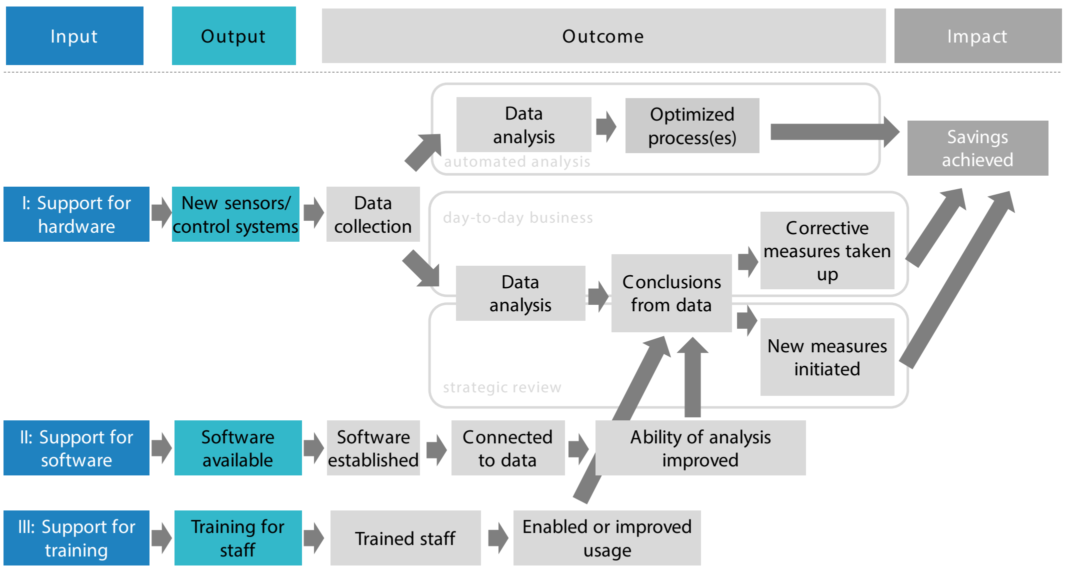


Figure 2: Illustration of an impact model at the example of Module 3 of the EEE (source: Neusel et al. 2023).

The impacts, such as energy savings but also various other aspects are mostly quantitatively assessed along a set of key performance indicators (KPI, **step 4**). For the EEE, these add up to about 100 individual indicators (including sub-indicators) across all categories. Table 3 provides an overview of the KPI chosen for the EEE's evaluation by core evaluation areas. The list includes a set of indicators of general knowledge interest (G), which serve to provide a general characterisation for each of the six funding modules. This is followed by the actual description of target achievement (A), effectiveness (B) and economic efficiency (C). Additional indicators seek to cover the quality of the procedural implementation (D), and the last category (E) contains indicators addressing module-specific and technology-specific issues and questions beyond the evaluation core objectives (e.g. the role of the module as a 'door-opener' for later participation in other modules). Most of the KPIs are quantitative values (e.g. GHG savings in tonnes of CO_2-eq.), yet some are provided qualitatively (e.g. quality of the funding process). For further details on the KPI shown in Table 3, the reader is referred to Neusel et al. 2024b.

Table 3: Overview of key performance indicators for the EEE (based on Neusel et al. 2024b).

(**G**) - **General knowledge interest:** Structural data on applications, approvals and funding
Availment by region, by type of company, by company size, by sector, by funding object, etc.
(**A**) - **Target achievement:** To what degree have the established targets been achieved?
Reduction in ▪ final and primary energy consumption ▪ GHG-emissions ▪ energy and resource costs
(**B**) - **Effectiveness:** To what degree is the measure causal to the achievements?
Total value of the effect adjustment: Gross impact - Free-rider and pull-forward effects + Spill-over and follow-on effects = Net effect
(**C**) - **Economic efficiency:** How efficient is the measure from the implementer's perspective with regard to achieving the targets and concerning the use of resources?
▪ Total costs (funding and administrative costs) ▪ Funding efficiency ▪ Total triggered investments ▪ Leverage effect (triggered investments to amount of funding)
(**D**) - **Procedural implementation:** How is the operational implementation perceived?
▪ Process from company and implementer's perspective (qualitative) ▪ Response time and complaints management
(**E**) - **Specific knowledge interest:** Module-specific questions that go beyond the specified evaluation objectives

The quantitative KPI for the EEE are determined using two sources of information (**step 5**): Data from the administration of the support scheme and data from the beneficiaries. Available administrative data contains information from the application on the beneficiary (e.g. name, location, company size) and financial and administrative information on the activity submitted for funding. For Modules 4 and 5, the competition line, information on greenhouse gas and resource savings is also partially available in the administrative data set. Complementary data and views on the funding process from the participants are collected via an annual online survey among beneficiaries. The survey consists of common questions for all modules and specific questions addressing aspects of individual modules or implementing agencies. Participation typically takes about 15 to 25 minutes. In the most recent evaluation of 2023, almost 11,000 beneficiaries were invited, with a response rate of around 20%, which is similar to previous years.

Using this data, gross KPI values are determined. For analysing the role of the EEE in triggering investments in energy and resource efficiency, an effect adjustment by calculating the category (B) indicators is carried out (**step 8**). For this, the online survey contains several control questions. These questions address the extent and role of the funding scheme for the investment by both, taking negative effects (e.g. free-riders: subtraction from gross values) and positive effects (e.g. spill-overs: addition to gross values) into account (Table 4) (Schlomann et al. 2017).

The evaluation accounts for both free-rider and spill-over effects, as shown in Table 4. Free-riders refer to investments or savings that would have occurred even without the funding scheme, including those that were already planned but brought forward due to the programme (pull-forward effect). Spill-over effects capture additional investments or savings triggered indirectly by the EEE, those that did not receive funding but were inspired by the programme, potentially leading to further energy efficiency actions.

A survey-based approach including a logic of several pre-defined questions is applied to quantify these effects. It is described in detail by Voswinkel 2018 for the EEE's predecessor programme. Net values are then calculated by subtracting free-rider effects from gross values and adding spill-over effects. For 2023, the effect adjustment reduces gross values by around 12 percentage points across all six funding modules (Figure 3). Violette and Rathbun 2017 give an account of other methods, including randomised control trials and quasi-experimental methods.

Table 4: Approach for net impact estimation of the EEE (Source: Voswinkel (2019)).

Impact / Effects	Description
Gross value	Impact before considering effects
- Free-rider and pull-forward effects	Saving that would have occurred without policy and early replacement
+ Spill-over effects	Effects trough spill-over (transfer) on third parties and other areas not directly credited to the programme
= **Net value**	Impact after adjusting for effects

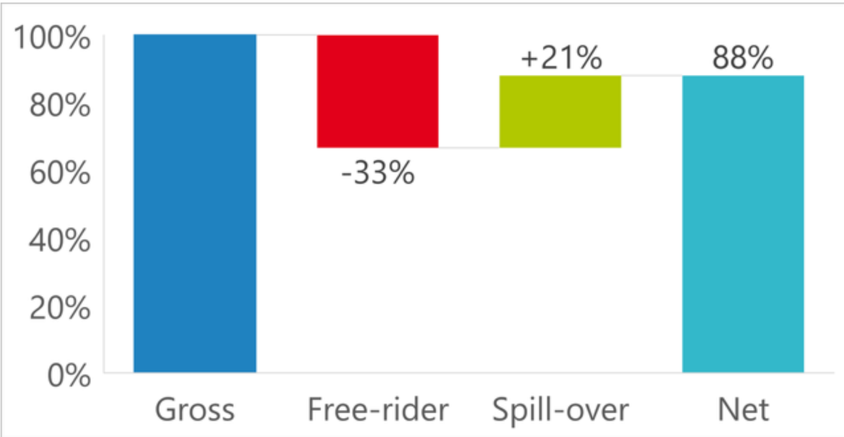


Figure 3: Effect adjustment from gross to net values as part of the EEE evaluation of 2023 (source: Neusel et al. 2024a).

DISCUSSION ON THREE METHODOLOGICAL ISSUES GAINING MOMENTUM

The foundational evaluation methodology is based on a nine-step approach, as outlined in Table 2. This methodology has been consistently applied over the years. However, the dynamic character of the EEE repeatedly poses new methodological and operational challenges to the annual evaluation. As a result, refinements or extensions of the methodology are necessary in certain areas. Summarised in Figure 4, previous publications have already covered some of these issues arising from the various changes in the EEE itself over the years, as for example described in Neusel et al. 2024b. However, three challenges of the most recent evaluation of the EEE for 2023 have not yet been addressed and are increasingly gaining momentum.

[1] Voswinkel 2020, [2] Hirzel and Schlomann 2022 [3] Neusel et al. 2024b, [4] Voswinkel 2018

Figure 4: Selection of methodological issues related to the evaluation of the EEE and its predecessor programme covered in prior publications (grey) and three novel issues (green) (own illustration).

#1 THE SUCCESSIVE DECARBONISATION OF THE ELECTRICITY SYSTEM

Some of the KPIs in Table 3 seek to project the impact of a measure on the energy and greenhouse gas (GHG) emissions over several years. This period typically spans the lifetime of the measure. Energy savings are usually expressed in terms of final energy and primary energy. Final energy refers to the energy used to operate an application, such as the amount of electricity used to operate an electric furnace. In contrast, primary energy reflects the amount of energy required to produce that final energy from the original energy source. For example, it includes the energy contained in the coal needed to generate electricity. To make this transition, a primary energy factor is used, which describes a ratio between primary and final energy. The GHG savings are determined in the same way, using an emission factor that reflects the amount of emissions per unit of final energy used. The selection of these factors can substantially influence the KPI.

While the emission factors for fuels such as coal and gas remain largely constant, the emission factor for electricity has been declining over the last couple of years. This decrease has gained momentum over the last couple of years in Germany (Umweltbundesamt 2024), and mid-term reference projections until 2030 have gained importance (Öko-Institut e.V. et al. 2023). These projections are based on consideration of the price for emission allowances in Europe, the implementation of measures in the energy industry sector including the deployment of renewable energies and hydrogen, as well as the coal phase-out. Consequently, GHG emissions from electricity generation in the German energy industry sector are expected to change significantly by 2030. According to the scenario-based projections of the

2023 projection report, it is expected that the GHG emissions from electricity generation in the German electricity mix will fall from 482 g $CO_{2\text{-eq}}$/kWh in 2023 to less than 92 g $CO_{2\text{-eq}}$/kWh in 2030 (Öko-Institut e.V. et al. 2023, with-measure scenario). Consequently, while the dynamics of the emission factors were less relevant in the first years of the EEE evaluation, the more solid projections of the emission factors become more relevant for the EEE's future impact over the lifetime.

Table 5 illlustrates the comparison of using a static vs. a dynamic emission factor on lifetime savings: This illustration assumes new measures each year with a lifetime of five years and annual energy savings of 1,000 kWh_{el} from 2021 until 2029. This adds up to lifetime energy savings of 25,000 kWh_{el}. With a static emission factor, this converts into lifetime-related GHG emission savings of 11.25 t $CO_{2\text{-eq}}$, whereas with dynamic emission factors, it converts into 8.49 t $CO_{2\text{-eq}}$. Due to the successive decarbonisation of the electricity system, lifetime emission savings are lower in the dynamic case.

To address the issue of successive decarbonisation of the electricity system, it is essential to maintain transparency in the framework data used. This is reflected in step 2 of the methodological framework shown in Table 2, where framework conditions and harmonised parameters (e.g. emission factors, primary energy factors, underlying lifetimes of the respective measures and energy prices) are defined.

Table 5: Illustrative comparison of the impact of using a static and a dynamic emission factor on lifetime savings in case of annual energy savings of 1.000 kWh_{el} with a lifetime of 5 years from 2021 to 2029.

Year	2021	2022	2023	2024	2025	2026	2027	2028	2029	Lifetime savings
Energy savings in the respective year [kWh_{el}]	1,000	2,000[1]	3,000	4,000	5,000	4,000	3,000	3,000	1,000	25,000
Static emission factor* [g CO_2/kWh_{el}]	Constant at 450									-
GHG reduction in the respective year [t CO_2]	0.450	0.900	1.350	1.800	2.250	1.800	1.350	0.900	0.450	11.25
Dynamic emission factor* [g CO_2/kWh_{el}]	450	466	482	410	362	282	214	165	116	-
GHG reduction in the respective year [t CO_2]	0.450	0.932[2]	1.446	1.640	1.810	1.128	0.642	0.330	0.116	8.49

* Öko-Institut e.V. et al. 2023

#2 ATTRIBUTING SAVINGS FROM RESOURCE EFFICIENCY MEASURES

The second issue revolves around attributing savings from resource efficiency measures, i.e. emission reductions from saving material. The EEE is the first energy efficiency funding scheme in Germany to incorporate resource efficiency using a life cycle assessment (LCA) approach. Since November 2021, investments in resource efficiency measures have been eligible for funding besides energy efficiency measures. EEE evaluation results from resource efficiency projects in 2022 indicate that material saving or substitution measures are particularly substantial for savings in sectors such as cement, metal, and plastics. These projects are characterised by high absolute GHG reductions (Weinert et al. 2024). This underlines that other resources besides energy may substantially affect the EEE's target impact.

The computation of GHG savings as an effect of resource efficiency – similar to the calculation of GHG savings from energy – is based on a conversion factor, more specifically, a material-specific CO_2 conversion factor (e.g. in kg CO_2/kg) which is multiplied with the amount of savings of that material (e.g. in kg).

1 Measures implemented in 2021 with a lifetime of 5 years will save 1.000 kWh in 2022, and measures implemented in 2022 (again with a lifetime of 5 years) will save a further 1.000 kWh in 2022.

2 Measures implemented in 2021 and 2022, saving a total of 2.000 kWh in 2022, are converted into GHG emission savings using the 2022 emission factor of 466 g CO2/kWhel.

Such savings from resource efficiency can be viewed from two perspectives: The first is a climate perspective, which considers the reduction in emissions released into the global atmosphere, regardless of their origin. The second is an accounting perspective, which seeks to identify the savings resulting from of a particular measure. Care must be taken when analysing the impact of a measure from an accounting perspective, as some resources are global commodities, i.e. traded between regions. To illustrate this, Table 6 shows a simplified case with two regions which exchange resources and different regional resource conversion factors (e.g. due to different regional industries).

Table 6: Illustration of the attribution and accuracy of savings from resource efficiency measures to a Region A where the resource efficiency measure is implemented.

		Resources originating from	
		Region A	**Region B**
Resource conversion factors reflect the actual situation in	**Region A**	Savings attributable to Region A and accurate	Savings not attributable to Region A and likely over-/ underestimated
	Region B	Savings in principle attributable to the Region A, but likely over-/ underestimated	Savings not attributable to Region A and accurate

From a climate perspective, attribution errors are irrelevant as long as the conversion factors accurately reflect the actual impact (upper left and lower right quadrants in Table 6). Yet attribution errors may have two consequences for accounting: First, the average resource conversion factors used to determine GHG emission savings from resource efficiency measures may not necessarily reflect the actual impact of the respective resources. This can occur if only regional proxy values for the resource conversion factors are available (lower left quadrant). Alternatively, the correct factor may be available, but the resources could originate from a different region (upper right quadrant).

The implication points at the issue that savings cannot not be simply attributed to one region. In extreme cases, this could result in GHG savings exceeding the actual GHG inventory of a region/country if the inventory does not account for the “grey emissions” associated with imported resources. Therefore, within evaluations such as the EEE, it is essential to make conscious decisions regarding how to handle emission savings derived from imported materials. This consideration may become relevant for national reporting. It is particularly important when these savings are related to national targets or national emissions.

#3 DECISIONS THAT CONSIDER FUNDING EFFICIENCY ALONE

The third issue concerns using funding efficiency as the lead criterion for programme design only. Funding efficiency in this context is defined as the Euros expended (funding including administrative costs) per saved tonne of CO_2 over the measure's lifetime. This indicator is especially relevant in the context of limited public budgets, as it provides a straightforward metric for policymakers when evaluating effectiveness. However, relying solely on funding efficiency can be overly simplistic and may not capture the full impact of the measures. Factors that must be considered include (Schlomann et al. 2020):

- **Activation of target groups and potential:** Some target groups, such as small and medium-sized enterprises (SMEs), are more difficult to reach, i.e. with greater effort and higher funding volumes. This means that the funding efficiency of measures aimed at these groups is often lower. It should also not be neglected that it becomes more difficult to activate potential over time. This means that realising the last potential in an area of application is more complex and expensive than at the beginning.
- **Low-hanging fruits:** In that line of thought, the so-called "low-hanging fruits" with high economic efficiency are more likely to lead to an attractive funding efficiency than in-depth investments, e.g. those with a high degree of innovation and/or lighthouse character. However, addressing "low-hanging fruits" alone (particularly in the area of cross-cutting technologies) appears far from sufficient to achieve the long-term climate targets.
- **Economies of scale:** Small measures are administratively easier to implement, but they are often associated with smaller and short-term savings. Large measures are generally more expensive and administratively complex, but are usually associated with long-term and far-reaching savings, even if they may be less efficient in terms of funding. Nevertheless, such measures are also necessary in order to utilise the entire savings potential in an area.

In the EEE, there is a large variation in the funding efficiency of the six funding modules. For example, over the entire 2019-2023 funding period, the funding efficiency of Module 4 is 42 Euros per tonne of CO_2 for an assumed lifetime of 8 years, while it is 113 Euros per tonne of CO_2 (8 years lifetime) for Module 1. Among other factors, this can be attributed to the fact that Module 4 finances significantly larger projects than Module 1 due to its technology-open focus. In addition, Module 1 is mostly dominated by SMEs and funds smaller, more cost-effective cross-cutting measures. While from a purely monetary view of funding efficiency, Module 4 appears much more effective, a sole focus on this value would neglect to take the particularities of the modules.

It can be concluded that funding efficiency should not be used alone. Instead, it needs to be seen in the context of the characteristics and objectives of the measure. This includes factors such as the type and size of the measure, the type of reduction potential addressed as well as the long-term nature and depth of the effect of the induced energy efficiency measures. Therefore, it is essential to interpret funding efficiency within the specific context of each measure. This helps to gain a comprehensive understanding of their effectiveness.

CONCLUSIONS

The aim of this paper was, on the one hand, to outline the mixed methods approach underlying the evaluation of the Federal Funding Scheme for Energy and Resource Efficiency in the Economy (EEE). On the other hand, the aim was, to discuss methodological issues which arise from recent modifications of the scheme and its framework conditions.

The evaluation of energy efficiency funding schemes tends to be a complicated matter and obtaining the results of such an evaluation depends on many methodological choices along the way of the evaluation. Particularly in case of complex funding schemes- such as the EEE - which include multiple funding modules, it is crucial to rely on a standardized methodology. This ensures comparability and supports the meaningful interpretation of evaluation results. The methodology presented in this paper promotes a consistent evaluation process, while allowing flexibility to account for scheme-specific design, data availability, and contextual factors.

Experience from five consecutive years of evaluating the EEE, along with insights from the predecessor programme, demonstrates that the overall methodology, based on a nine-step approach, provides a robust and reliable foundation. Yet, it can be observed that there is a constant need for refinement and adaptation to changes such as shifts in external conditions, evolving interests or the introduction of new design elements. In this paper, three of these issues have been outlined in more detail. They underline that it (#1) requires transparency how the dynamic decarbonisation of the energy system is taken into account, that (#2) conscious decisions on how to consider resource efficiency in GHG accounting are needed and that (#3) funding efficiency can only be applied for design of such instruments if the context of these values has been sufficiently appreciated.

ACKNOWLEDGEMENTS

This paper is based on results of a project carried out for the German Federal Ministry for Economic Affairs and Climate Action (BMWK). We express our gratitude to the representatives of the Ministry and the implementing agencies for their support, as well as to our colleagues from Prognos, IER and Öko-Institut who contributed to the study.

REFERENCES

BMWK (2024a): Richtlinie zur Bundesförderung für Energie- und Ressourceneffizienz in der Wirtschaft – Förderwettbewerb, Richtlinie Förderwettbewerb, revised 1/25/2024. Source: Bundesanzeiger. In : Bundesanzeiger.

BMWK (2024b): Richtlinie zur Bundesförderung für Energie- und Ressourceneffizienz in der Wirtschaft – Zuschuss und Kredit, Richtlinie Zuschuss und Kredit, revised 1/25/2024. Source: Bundesanzeiger. In : Bundesanzeiger.

Brunzema, Iska; Schlomann, Barbara; Kemmler, Andreas (2022): How to increase the transparency of ex-ante impact evaluations of energy efficiency and climate policies. Illustrated by the example of a funding program for energy and resource efficiency in industry in Germany. With assistance of Fraunhofer-Gesellschaft.

Hirzel, Simon; Antoni, Oliver; Heinrich, Stephan; Jessing, Dominik; Paar, Angelika; Pehnt, Martin et al. (2019): Evaluierung und Weiterentwicklung des Energieeffizienzfonds (Projekt Nr. 63/15). Abschlussbericht - Langfassung. Fraunhofer ISI; Prognos AG; ifeu; Stiftung Umweltenergierecht. Karlsruhe/

Heidelberg/Basel/Würzburg. Available online at https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/evaluierung-und-weiterentwicklung-des-energieeffizienzfonds.pdf?__blob=publicationFile&v=8, checked on 1/3/2022.

Hirzel, Simon; Berger, Carmen; Neusel, Lisa (2024): A measure of control: about sensors, measurement and control equipment in German companies. In: Proceedings of the eceee summer study 2024.

Hirzel, Simon; Neusel, Lisa; Heinrich, Stephan; Weinert, Karsten; Schlomann, Barbara (2022): Evaluation of the German Funding Scheme for Energy Efficiency in the Economy: Classical versus Competitive Financial Measures. In: Energy Evaluation Europe Conference 2022.

Hirzel, Simon; Schlomann, Barbara (2022): Evaluating Multi-measure Schemes for Enhancing Energy Efficiency: The German Energy Efficiency Fund. In: Proceedings of the eceee summer study 2022.

Neusel, Lisa; Hirzel, Simon; Berger, Carmen; Schlomann, Barbara; Heinrich, Stephan; Weinert, Karsten et al. (2023): Evaluation der „Bundesförderung für Energie- und Ressourceneffizienz in der Wirtschaft“ (Zuschuss und Kredit/Förderwettbewerb). Projekt für das Bundesamt für Wirtschaft und Ausfuhrkontrolle (BAFA) - Referat 123 / Projekt BfEE 08/2020. Jahresbericht 2022. Fraunhofer ISI, Prognos, Universität Stuttgart, Öko-Institut. Karlsruhe, Basel, Stuttgart, Berlin. Available online at https://www.bmwk.de/Redaktion/DE/Evaluationen/Foerdermassnahmen/231031-evaluation-eew-jahresbericht-2022.pdf?__blob=publicationFile&v=10, checked on 1/25/2024.

Neusel, Lisa; Hirzel, Simon; Berger, Carmen; Schlomann, Barbara; Heinrich, Stephan; Weinert, Karsten et al. (2024a): [Duplikat] Evaluation der „Bundesförderung für Energie- und Ressourceneffizienz in der Wirtschaft“ (Zuschuss und Kredit/Förderwettbewerb). Projekt für das Bundesamt für Wirtschaft und Ausfuhrkontrolle (BAFA) - Referat 123 / Projekt BfEE 08/2020. Jahresbericht 2023. Fraunhofer ISI, Prognos, Universität Stuttgart, Öko-Institut. Karlsruhe, Basel, Stuttgart, Berlin. Available online at https://www.bmwk.de/Redaktion/DE/Evaluationen/Foerdermassnahmen/241217-evaluation-eew-jahresbericht-2023.pdf?__blob=publicationFile&v=6, checked on 1/9/2025.

Neusel, Lisa; Hirzel, Simon; Weinert, Karsten; Heinrich, Stephan; Grodeke, Anna-Maria (2024b): Efficiency Unleashed: Evolution and Impact of Germany’s Funding Scheme for Energy and Resource Efficiency in the Economy. In: Proceedings of the eceee summer study 2024.

Öko-Institut e.V.; Fraunhofer ISI; IREES; Thünen-Institut (2023): Kernindikatoren des Projektionsberichtes 2023 (Datentabelle). Available online at https://www.umweltbundesamt.de/dokument/datenanhang-kernindikatoren-projektionsbericht-2023, checked on 8/7/2024.

Schlomann, Barbara; Hirzel, Simon; Nabitz, Lisa; Heinrich, Stephan; Jessing, Dominik; Paar, Angelika et al. (2017): From Targets to Impacts: Eight Steps for Evaluating Energy Efficiency Policies. In : International Energy Program Evaluation Conference - IEPEC 2017. Making Ambitious Reductions Real: Accurate and Actionable Evaluation. Baltimore, 8 - 10 August.

Schlomann, Barbara; Voswinkel, Fabian; Hirzel, Simon; Paar, Angelika; Jessing, Dominik; Heinrich, Stephan et al. (2020): Methodikleitfaden für Evaluationen von Energieeffizienzmaßnahmen des BMWi (Projekt Nr. 63/15 - Aufstockung). Fraunhofer ISI; Prognos AG; ifeu; Stiftung Umweltenergierecht. Karlsruhe/Heidelberg/Basel/Würzburg. Available online at https://www.bmwi.de/Redaktion/DE/Downloads/M-O/methodik-leitfaden-fuer-evaluationen-von-energieeffizienzmassnamen.pdf?__blob=publicationFile&v=6, checked on 1/3/2022.

Umweltbundesamt (Ed.) (2024): Entwicklung der spezifischen Kohlendioxid-Emissionen des deutschen Strommix in den Jahren 1990 - 2023. Available online at https://www.umweltbundesamt.de/publikationen/entwicklung-der-spezifischen-treibhausgas-10, checked on 1/15/2025.

Violette, Daniel M.; Rathbun, Pamela (2017): Estimating Net Savings - Common Practices. In NREL (Ed.): The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Available online at https://www.nrel.gov/docs/fy17osti/68578.pdf.

Voswinkel, Fabian (2018): Accounting of energy savings in policy evaluation. How to get a least 8 different (correct!) results from the same data. In: IEPPEC International Energy Policy and Programme Evaluation Conference. Vienna, Austria, 25–27 June 2018.

Voswinkel, Fabian (2019): Gotta catch 'em all - catches to evaluating heterogenous energy efficiency programmes. In: eceee 2019 Summer Study on Energy Efficiency: Is Efficient Sufficient? Eceee Summer Study proceedings. European Council for an Energy Efficient Economy. Stockholm: eceee, pp. 767–776.

Voswinkel, Fabian (2020): One size fits all? Emerging challenges to harmonising energy efficiency evaluations in a changing policy arena. In: Energy Evaluation Europe. Accelerating the energy transition for all:

Evaluation's role in effective policy making. Here East, Olympic Park, London, 29 June - 1 July.

Voswinkel, Fabian; Grahl, Andrea; Rohde, Clemens (2018): Less hot air for a less hot climate: evaluating the German waste heat reduction programme. In: Proceedings of the eceee Industrial Summer Study 2018. Kalkscheune, Berlin, Germany, 11 - 13 June, pp. 65–74.

Weinert, Karsten; Neusel, Lisa; Hirzel, Simon;: unav (2024): Resource efficiency: a new element in Germany's funding program for companies. In: Proceedings of the eceee summer study 2024.

AUTHORS

LISA NEUSEL
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Strasse 48
76139 Karlsruhe, Germany
E-mail: lisa.neusel@isi.fraunhofer.de
ORCID: 0009-0009-5537-1224

SIMON HIRZEL
Fraunhofer Institute for Systems and Innovation Research ISI
Breslauer Strasse 48
76139 Karlsruhe, Germany
E-mail: simon.hirzel@isi.fraunhofer.de
ORCID: 0000-0002-6323-917X

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e9, 1-18
DOI: 10.22163/fteval.2025.701

THE CONTRIBUTION OF R&I PROGRAMMES TO TRANSITION:

EVIDENCE ON TRANSFORMATIVE OUTCOMES IN THE AREAS OF THE GREEN TRANSITION AND THE ENERGY TRANSITION

MICHAEL DINGES, CHRISTIANE KERLEN, SURYA KNÖBEL AND KATHLEEN TOEPEL
DOI: 10.22163/FTEVAL.2025.701

ABSTRACT

This paper examines the role of Research and Innovation (R&I) programmes in driving systemic transitions, with a focus on the Green and Energy Transitions in the EU and Germany. It analyses two major programmes—the 7th Energy Research Programme (Germany) and Horizon Europe Clusters 5 & 6—through a transformative outcomes framework grounded in sustainability transition theory and the multi-level perspective (MLP). Using a mixed-method evaluation approach, the study assesses how these R&I initiatives foster innovation, support niche development, influence regime change, and drive systemic transformation. Findings highlight meaningful contributions to building and expanding niches, yet limited impact on regime destabilisation and institutionalisation. The paper underscores both the potential and limitations of R&I policies in catalysing sustainability transitions, offering insights for programme design, policy alignment, and evaluation practice.

BACKGROUND & RESEARCH QUESTION

The pressing need for a paradigm shift in response to escalating human-induced environmental change has fuelled the quest for a Green Transition in policy, economy, and society.

In the European Union, the 2019 European Green Deal marked a turning point in the political landscape, adopting key concepts such as net-zero and climate targets alongside sectoral policies. A key element of the Green Deal is its emphasis on digitalisation as a strategic enabler of the green transition, recognising that digital technologies can facilitate energy efficiency, smart infrastructure, and more sustainable resource use. However, critiques about the aspired transformation process within the Green Deal, the Green-Growth paradigm, and limits to growth remain unresolved, raising questions about the desired nature of the Green Transition. Growth-critical concepts highlight the constraints on human impact, adding depth to the understanding of the Green Transition.

In light of these challenges, the role of Research and Innovation (R&I) becomes crucial. Transition-oriented R&I programmes can serve as catalysts for innovative solutions, promoting the development of sustainable technologies and innovative practices to navigate the complexities of the Green Transition. Ultimately, they could play a pivotal role in shaping a more resilient and sustainable future.

To effectively assess and guide these processes, it is essential to employ analytical instruments that can accurately portray the transformation dynamics at play. For analysing transformation processes, tools such as systems mapping, scenario modelling, and policy simulations have been developed to provide insights into how transformation unfolds over time, highlighting interdependencies, potential trade-offs, and emergent properties of change. In the area of R&I policy evaluations, the necessity for a nuanced understanding of innovation pathways becomes evident, as it helps policymakers assess the feasibility and effectiveness of various strategies within the transition process. Systematic, transformation-oriented analytical instruments may inform decision-makers about the most effective leverage points for intervention.

Against this background, this paper examines the role of R&I programmes in facilitating transition processes, focusing on their contributions to both the Green Transition and the Energy Transition. Adopting a transition theory perspective, the paper provides empirical evidence on how two major R&I initiatives drive systemic change by fostering innovation, supporting technological development, and enabling institutional shifts.

From a policymaker's perspective, the paper offers insights into how R&I programmes can serve as strategic tools for steering transitions, identifying key leverage points for intervention. Additionally, for researchers in R&I policy evaluation and Science, Technology, and Innovation studies, it demonstrates how the contributions of these programmes can be systematically measured, offering a framework for assessing their effectiveness in driving sustainable transformations.

PROGRAMMES UNDER CONSIDERATION AND THEIR EVALUATIONS

This research draws upon two independent evaluation studies that have been conducted for the following two major R&I programmes related to the Green and Energy Transitions:

1. The 7th Energy Research Programme 'Innovations for the Energy Transition' (2018-2023) is pursuing a strategic approach for energy research, focusing on the transfer of technology and innovation. The Federal Ministry for Economic Affairs and Climate Action (BMWK) in Germany funds Collaborative Projects (TRL 3-7), Real-World Labs (TRL 7-9), Micro Projects, and accompanying measures. The programme also strengthens research on cross-system issues and strives for close networking at the international and European levels. Open calls for participation offer continuous opportunities for application. In total, 6,499 sub-projects from companies, universities, research institutions, and other organizations were funded from 2018 until the end of 2023. These sub-projects were based on proposals independently developed by the grant recipients. The total funding volume (federal funds) amounted to €3.317 billion.
2. Horizon Europe Cluster 5 (Climate, Energy, and Mobility) and Cluster 6 (Food, Bioeconomy, Natural Resources, Agriculture, and Environment) are key pillars of the Green Transition, supporting research and innovation to tackle climate change, enhance sustainability, and ensure the responsible use of natural resources. The European Commission funds Research and Innovation Actions (TRL 2-6), Innovation Actions (TRL 6-8), as well as Coordination and Support Actions and Partnerships,

> facilitating technological advancements and the implementation of policies. Both clusters operate through specific thematic calls in designated work programmes, fostering collaboration across sectors and disciplines. By June 2023, a total of 1,016 projects have been funded under these clusters, driving innovative solutions for Europe's environmental and energy challenges. With a combined budget of €24.075 billion for 2021-2027, Horizon Europe Clusters 5 and 6 support international cooperation, multi-disciplinary research, and systemic transformations needed for a sustainable future.

Both the 7th Energy Research Programme (ERP) and Horizon Europe Clusters 5 and 6 aim to drive sustainable innovation and systemic transformation through research and development funding. They share a commitment to supporting decarbonisation, fostering technological advancements, and hence facilitating policy implementation for the Green and Energy Transitions. A key difference lies in their scope and scale: the 7th ERP is a national programme focused on Germany's energy transition, emphasising applied research and real-world demonstration projects, whereas Horizon Europe operates at the EU level, covering a broader range of environmental and energy challenges with a stronger emphasis on interdisciplinary research and international collaboration.

The evaluation study "Horizon Europe and the Green Transition: Interim evaluation support study" (European Commission 2024) was part of a back-to-back evaluation for the ex-post evaluation of Horizon 2020 and the interim evaluation of Horizon Europe conducted on behalf of the European Commission, with a thematic focus on Green Transition. The interim evaluation support study was conducted between February 2023 and January 2024. During the inception phase of the study, a specific methodological approach was designed in agreement with the Steering Committee, utilising a combination of various data collection and analysis tools, including bibliometrics, case studies, surveys, and benchmarking.

The evaluation of the 7th ERP commenced in 2021 as a five-year accompanying evaluation. The evaluation employs a mixed-methods approach to develop ongoing analyses, reflections and recommendations as a basis for steering and continuous improvement of the programme („programme learning"), while also contributing to an assessment of its effectiveness and impact. It informed the design of the 8th Energy Research Programme, which was launched as a mission-oriented research programme in 2023, with funding commencing in 2024.

ANALYTICAL FRAMEWORK

Both evaluation studies underlying the present research paper employed an evaluation framework that aimed to assess the contributions of these programmes to ongoing systemic transformations. The evaluations employed an enhanced programme theory approach, based on the Transformative Outcomes Framework (Ghosh et al., 2021), which is embedded in the Multi-Level Perspective (MLP) on System Innovation and helps to grasp systemic transformations better (see Dinges et al., 2022). A programme theory is often built on a conceptual model that explains how change is supposed to happen and what leads to impact. These models can be either explicitly stated or assumed without being clearly outlined. They might be based on theories, real-world experiences, or specific perspectives. In addition to traditional input-output-outcome-impact (I-O-O-I) frameworks, this paper adopts a theoretical approach to analyse the theory of change behind these programmes, drawing on insights from sustainability transition research.

The field of sustainability transition research defines transformative change as a fundamental shift in how socio-technical systems operate (Markard et al., 2012). This perspective is based on the MLP theory, which explains change as an interaction between three levels (Geels, 2011):

1. Stable Regimes – existing systems and structures that dominate the status quo.
2. Niche Innovations – new ideas, technologies, or practices that challenge the existing system.
3. Landscape Pressures – external forces (e.g., climate change, economic shifts, policy changes) that push for transformation.

The way these three levels interact determines how transitions unfold, leading to different transition pathways (Geels & Schot, 2007). Managing sustainability transitions is challenging because these changes are complex and evolutionary—no single group of actors can fully control them (Kivimaa et al., 2019). Instead, successful transformation requires carefully designed processes that 1) protect and support the growth and expansion of new, sustainable alternatives (niche innovations), and 2) break down or phase out existing, unsustainable systems (incumbent regimes).

These transformative processes comprise 12 key Transformative Outcomes (Ghosh et al., 2021), which provide a structured approach to understanding and guiding change through policies. In evaluation studies, outcomes refer to the measurable effects, results, or changes that occur as a result of a policy, programme, or intervention. Outcomes (intended and unintended) are typically assessed in relation to predefined goals and objectives, using indicators to track progress. Unlike conventional outcomes, the concept of Transformative Outcomes focuses on systemic and structural change rather than just measuring success against predetermined objectives. The key distinction is that Transformative Outcomes are not static results but ongoing, process-oriented mechanisms that contribute to transformative change over time. They play a crucial role in driving transformation, which is why it is important to focus on understanding how specific programmes contribute to these processes.

The twelve Transformative Outcomes are grouped into the following three overarching processes of transformative change: Processes 1 and 2 are related to supporting the growth and expansion of new, sustainable alternatives 1) building and nurturing niches; 2) expanding and mainstreaming niches; Process 3) is related to break down or phase out existing, unsustainable regimes:

Building and nurturing niches: This process focuses on developing and supporting emerging innovations through 1) Shielding - active protection of the niche (e.g. through R&I subsidies for development), 2) Learning - encouraging experimentation and knowledge exchange among niche actors that challenges beliefs and assumptions of (incumbent) actors, 3) Networking – strengthening connections between (niche) actors to support innovation, 4) Navigating Expectations – developing and sharing visions of change to inspire action.

Scaling and mainstreaming niches: This process focuses on mechanisms for scaling and broadening the reach of successful innovations and experiments beyond their niche, thereby increasing their scale and scope. It consists of 1) Upscaling – increasing the adoption of new practices/technologies by involving more users, 2) Replicating all or parts of the innovations in new contexts, 3) Circulating ideas and resources through learning, and transferring ideas from one niche to another through ensuring funding, skills and infrastructure support and 4) Institutionalising – modifying policies, norms, and regulations through shared narratives, definitions, standards and interpretations.

Opening up and unlocking regimes: This process focuses on embracing new perspectives, deconstructing rules that characterise a dominant regime, and creating space for alternatives to emerge and grow. It consists of 1) De-aligning and destabilising mechanisms that facilitate the decline or transformation of existing dominant socio-technical regimes, those entrenched systems of practices, institutions, rules, and technologies that maintain the status quo, 2) Unlearning and deep learning – mechanisms helping regime actors to questioning existing mindsets and values in comparison to new rules and routines, associated with solving sustainability challenges, 3) strengthening regime-niche interactions – creating stronger links between niche actors and regime actors, and 4) changing perceptions of landscape pressures – challenging collective perceptions about wider socio-political developments.

To understand whether real, lasting change is happening, we developed project beneficiary questionnaires in both evaluation studies that help track progress in two important ways:

First, to assess the current state of transformation processes, respondents were asked to evaluate the current state of transformation processes within their area of expertise as experts in the field. This section of the questionnaire, conducted only as part of the Energy Research Programme evaluation, aimed to capture the broader picture:

- To what extent are transformative processes already underway?
- Is the country actively driving sustainable change, or are existing systems remaining unchanged?
- Do researchers perceive tangible shifts toward systemic transformation, or does the status quo persist?

By gathering insights on the Transformative Outcomes as seen by experts in the field, this evaluation helped to assess whether and how fundamental changes are taking place, providing a clearer understanding of the progress and challenges in sustainability transitions.

Second, are the funded projects truly driving meaningful transformation? While many projects succeed in meeting their immediate objectives, we aimed to assess whether they also contribute to more profound, lasting change. Specifically, are they helping to shift the underlying systems and structures necessary for a sustainability transformation, as outlined by the 12 Transformative Outcomes? From the researchers' perspective, to what extent are these projects influencing the broader change processes, and how well do their impacts align with the outcomes that are known to catalyse systemic transformation?

By applying this survey concept, we provided a structured framework for how R&I programmes and their projects contributed to actively shaping systemic transformation. Table 1 shows how the Transformative Outcome survey operationalised each Transformative Outcome through multiple survey items.

Transformative Outcome	Survey operationalisation
Building and nurturing niches	
Shielding	*Establishing and promoting new fields of innovation* 1. Development of new, ground-breaking solutions 2. Establishing new fields of knowledge 3. Supporting pioneers 4. Protecting new fields of innovation from dominant interest 5. Protecting new fields of innovation from market influence
Learning	*Learning and exchange of experiences* 6. Learning about subject specific problems 7. Exchange of experience on innovative solutions 8. Reflection on new solutions and their application 9. Promotion of professional competences 10. Open communication of failures *Promoting awareness of problems and new ways of solving them* 11. Awareness of new ways of solving problems 12. Questioning conventional ways of solving problems 13. Critical questioning of established basic assumptions 14. Breaking down established ways of working
Networking	*Networking between and within young innovation fields* 15. Networking between new innovation actors 16. Networking between young innovation fields 17. Synergies between young innovation fields 18. Cooperation between pioneers
Navigating expectations	*Managing expectations and promoting shared visions* 19. Strengthening innovative solutions as legitimate alternatives for the future 20. Establishing and promoting new fields of innovation contributing to a green transition 21. Common understanding of the future direction of innovation fields 22. Anticipation of future trends and shocks 23. Reduced uncertainty about context conditions for innovations
Expanding and mainstreaming niches	
Upscaling	*Expansion of new fields of innovation* 24. Broad acceptance of novel approaches by various stakeholder groups 25. Large-scale use of innovations 26. Accelerated implementation of innovations 27. Recognition of new “rules of the game” associated with innovations
Replicating	*Replication of innovative solutions in new contexts* 28. Application of innovations in other places or regions 29. Transfer of innovations into other application areas 30. Re-interpretation and adaptation of solutions in other contexts 31. Transfer of innovation into other contexts

Circulating	*Dissemination and diffusion of innovative solutions* 32. Widespread dissemination of new, innovative ideas 33. Open communication of novel solutions 34. Transfer of knowledge beyond the boundaries of one's field of knowledge 35. Intensive discussion of innovations from other contexts
Institutionalising	*Institutionalisation of new strategies and norms* 36. Institutionalisation / mainstreaming of new solutions 37. Establishment of new, common definitions or norms 38. Establishment of new legal and regulatory foundations 39. Establishment of new rules of conduct
Opening up and unlocking regimes	
De-aligning and destabilising	*Breaking up outdated structures and strategies* 40. Open-mindedness of established actors for new ideas 41. Opening the system to new strategies 42. Breaking up outdated processes 43. Shaking up the established system through radical innovations
Unlearning and deep learning	*Abandoning outdated habits and rules* 44. Willingness of established actors to engage in new ways of solving problems 45. Questioning the usefulness of prevailing solutions 46. Acceptance of risks that innovations entail 47. Unlearning outdated rules and habits
Strengthening regime-niche interactions	*Exchange between "old" and "new" areas of knowledge* 48. Networking between pioneers and established players 49. Exchange between "old" and "new" areas of knowledge 50. Opportunities for pioneers to enter the dominant system 51. Opening up traditional patterns of cooperation to new actors
Changing perceptions of landscape pressures	*Systemic changes in terms of flexible response to changing framework conditions* 52. Recognition of the need for action due to new developments 53. Critical (re)interpretation of framework conditions 54. Rapid reaction to changing framework conditions 55. Flexible reactions to trends and shocks

In the Green Transition Evaluation, each question was tailored as follows: "To what extent do you expect that your HEU project will contribute to the Green Transition in terms…". In the Energy Research Evaluation, researchers were asked to "assess from your personal perspective the extent to which the following developments for the energy transition in Germany are taking place in your area of expertise." In addition, researchers were asked to what extent their project contributes to processes related to the Energy Transition, except from the process of de-aligning and destabilising, as this process was not part of the theory of change. For questions related to the program's contribution to transition processes, only the headlines (e.g., Establishing and promoting new fields of innovation) have been asked.

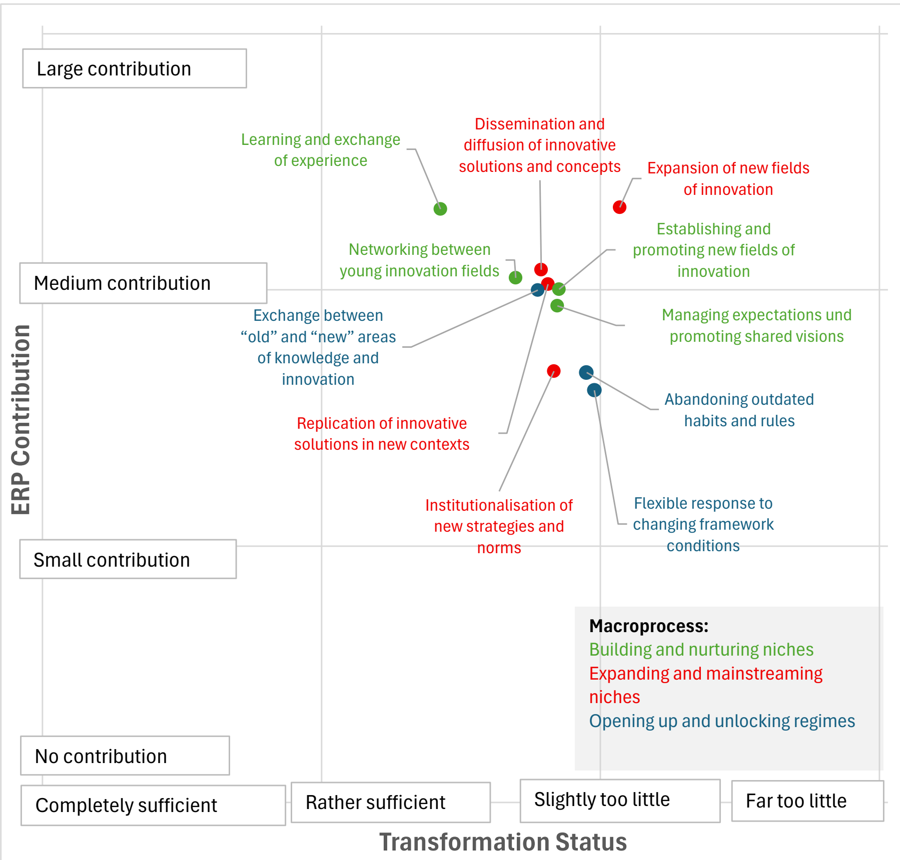
Source: Own compilation based on the surveys in the Green Transition evaluation and the Energy Research Programme evaluation.

SELECTED RESULTS

In the German case (7th ERP), the survey aimed to investigate how transformation processes are perceived by actors in the energy system and to what extend the programme contributes to these developments.

Figure 1 illustrates the relationship between the State of Transformation of the energy system (x-axis) and the Contributions of the 7th ERP to the Transformation (y-axis). The x-axis represents the degree of progress in the transformation process, moving from left (completely sufficient) to right (far too little). The y-axis measures the contribution of the 7th ERP to each Transformative Outcome operationalised in the survey.

Figure 1: Contributions of the 7th ERP in relation to the perceived transformation status



Note: Answers to the question: 'Before you answer questions about the energy research programme itself and your project, please assess from your personal perspective the extent to which the following developments for the energy transition in Germany are taking place in your area of expertise.' N= 5,235, average of 3,450 responses. As well as answers to the questions: To what extent does your project in the 7th Energy Research Programme contribute to the following developments in the energy system? To what extent do real-world labs of the 7th Energy Research Programme contribute to the following developments in the energy system? To what extent do accompanying measures of the 7th Energy Research Programme contribute to the following developments in the energy system?' N = 364 projects, average of 2,403 responses.

Source: Online survey 7th ERP (2024)

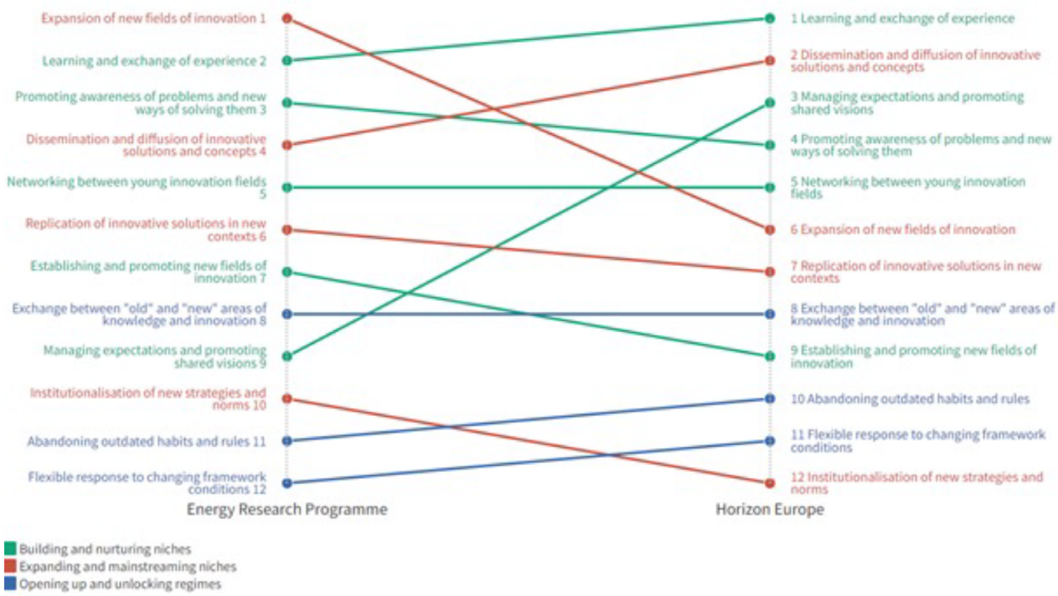
Several outcomes are located in the upper right quadrant, where transformation is perceived as insufficient, but the 7th ERP is making a significant contribution. Transformative outcomes in this area include ‘expansion of new fields of innovation’ and ‘establishment and promotion of new fields of innovation’. These cases demonstrate that the 7th ERP is actively involved in areas where change is still required. The alignment between need and support indicates effective targeting of resources in areas where further transformation is both possible and supported by research activities.

The transformative outcomes ‘learning and exchange of experience’, as well as ‘networking among young innovation fields’, are viewed as comparatively well progressed, though not yet fully sufficient, and are attributed high contributions from the 7th ERP. These may be interpreted as areas of successful research engagement, where sustained 7th ERP support has contributed to the maturation of knowledge systems and actor networks.

A third group of transformative outcomes is located in the lower right quadrant: ‘institutionalisation of new strategies and norms’, ‘abandoning outdated habits and rules’, and ‘flexible response to changing framework conditions’. These are areas where transformation is perceived as significantly lacking, yet the 7th ERP is considered to make a comparatively limited contribution. The transformative outcomes in this area comprise matters of regime-level or structural change, such as shifts in governance, regulations, or established routines. The 7th ERP’s lower contribution here may reflect the inherent limitations of a research program, which may lack direct instruments to influence institutional or political frameworks. Nonetheless, the perceived gap highlights a critical tension: while the 7th ERP may be structurally constrained in these domains, these areas are perceived to be central to advancing the energy transition, suggesting a need to strengthen the interface between research and institutional change, potentially through collaboration with policy actors or the integration of research insights into decision-making processes.

When comparing the 7th ERP with Horizon Europe in terms of its contribution to transformative outcomes, the similarities outweigh the differences between the two R&I programmes. However, there are some striking differences (Figure 2).

Figure 2: Rank Order of Programmes' contribution to transformative outcomes



Source: ERP and Horizon Europe Survey data, own analysis.

Most respondents in the online surveys indicated that their projects contribute particularly well to the macro-processes of 'building and nurturing niches' and 'expanding and mainstreaming niches'. Compared to the 7th ERP, Horizon Europe makes a greater contribution to managing expectations and promoting shared visions. In the European Programmes, no significant differences were found across the different Societal Challenges or Clusters (e.g., energy or mobility), and the anticipated results from Horizon Europe exceed those from H2020.

In the 7th ERP, it is noticeable that the programme's contribution is more focused on the macro-process 'expanding and mainstreaming niches' than on 'building and nurturing niches', compared to Horizon Europe. The activities funded by the 7th ERP have made significant contributions to technology development, the promotion of innovation, and the demonstration and application of new solutions in new contexts. The contributions vary according to the instruments used in the 7th ERP. The scheme R&I projects focused on individual technologies contribute in particular to the macro-process of 'building and nurturing niches'. Real-world labs support the expansion and mainstreaming of niche areas. The accompanying measures (e.g., energy research networks) effectively facilitate learning and the exchange of experiences at the project level, while also raising awareness of new and innovative solutions.

Within the macro-process 'expanding and mainstreaming niches', the results for 'institutionalisation of new strategies and norms', are significantly lower for both considered programmes (Horizon Europe as well as 7th ERP), calling into question the boundaries of an R&I programme, and the links between R&I, policy making, and deep learning of system actors.

Overall, contributions to the transformative outcomes within the macro-process ‘opening up and unlocking regimes’ are distinctly lower than those to the other transformative outcomes. There are only low contributions to ‘abandoning outdated habits and rules’, and limited contributions to a ‘flexible response to changing framework conditions’.

Both evaluations also analysed the involvement of stakeholders in programme planning and the funded projects. The analyses for the Energy Research Programme show that stakeholder groups outside the direct target groups (research organisations and industry) are reached to a lesser extent (Dinges et al. 2023). For the European FPs, it becomes evident that although the involvement of regulatory authorities and standardisation bodies has improved, stakeholder involvement is still not sufficient in some areas. Relevant needs for the energy or green transition that go beyond the traditional focus of a research programme receive comparatively little support.

BENEFITS OF ANALYSING TRANSFORMATION PROCESSES IN EVALUATIONS AND REMAINING CHALLENGES

Analysing transformation processes in the evaluation of instruments and programmes provides various benefits, including the ability to assess their contributions to intended outcomes better. Empirically verifying impact mechanisms is critical to understanding whether interventions are achieving their objectives and how these mechanisms function in practice. This evidence-based approach enables evaluators to identify gaps in implementation or unintended consequences, allowing for adjustments to improve overall effectiveness.

Additionally, making changes in prioritisation visibly ensures transparency and adaptability, allowing stakeholders to realign efforts in response to shifting conditions or emerging priorities. By enabling systematic comparisons between instruments, evaluations contribute to the identification of best practices and foster the development of more targeted, effective, and scalable solutions.

Tracking the status of transformative outcomes over time offers critical insights into the dynamics of ongoing transformation processes and helps to navigate the complexities of systemic change more effectively. A key benefit is the ability to record and track changes in perceptions of the energy transition among innovation actors in the field, such as researchers, project managers, and industry stakeholders. Their perceptions serve as crucial, real-time indicators of how deeply change is being internalised and implemented within socio-technical systems. Identifying ongoing transformation processes and needs in different sectors is essential to ensure tailored interventions that address sector-specific challenges and opportunities. External influences, such as political, economic, or social factors, as well as internal dynamics within organisations, should also be systematically considered.

Another benefit includes the development of an empirical survey design, which plays a vital role in capturing transformation processes as well as internal and external influences, providing robust data on how perceptions of energy transition and programme impacts evolve. This allows for an assessment of project-specific characteristics, including the progress achieved, the nature of the projects themselves, and the attributes of the organisations involved.

Despite these benefits, challenges remain. One critical challenge is finding a comprehensive explanation of the concept of transformation. Despite its frequent application in research and policy discussions, the term often lacks a clear, unified definition, which can hinder its practical implementation and evaluation. Relatedly, the design of survey items to capture the nuances of transformation processes presents difficulties, as highlighted by Knöbel et al. (2023). Developing effective survey instruments requires careful balance—while extensive survey sections with numerous items may provide in-depth insights, they also impose significant demands on respondents, potentially leading to lower participation rates and reduced data quality.

A further challenge lies in avoiding oversimplified interpretations of transformation processes. An inadmissible shortcut, for example, is equating R&I programmes fostering transformation solely with niche development processes. Technical change itself drives transformation by persistently generating new niches (Schot and Geels, 2007). These niches serve as critical sources of path-breaking innovations but require temporary protective spaces to develop, as emphasised in foundational works by Schot et al. (1994) and Kemp et al. (1998). The need to balance niche protection with broader systemic change presents a significant challenge for fostering innovation.

Regulatory learning represents another underexplored yet highly important aspect of transformation, particularly in the energy sector. There is limited understanding of the role that research, technology, and innovation can play in advancing regulatory frameworks that facilitate systemic transformation. The interaction between regulatory development and innovation systems demands greater attention to ensure that policies are both enabling and adaptive.

Finally, the importance of interfaces between old and new actors, as well as their influence on behavioural change, remains a complex area of study. Transformation processes often require bridging gaps between established systems and emerging actors, fostering collaboration while addressing resistance to change. Understanding how these interfaces influence behavioural dynamics and long-term systemic transitions is essential for effective transformation but remains insufficiently addressed in current research.

CONCLUSION

Embedding research and innovation programmes within the multi-level perspective of science and technology studies provides a robust framework for developing programme theory. By situating R&I initiatives within the broader context of socio-technical systems, the multi-level perspective allows for a deeper understanding of how transitions unfold across niche, regime, and landscape levels. This approach also helps delineate the boundaries of an R&I programme, offering insights into its limitations and identifying areas where coordination with other policy or sectoral efforts is necessary. Such coordination is crucial to ensuring that programmes effectively contribute to systemic transformation.

The empirical reviews of R&I programmes further underscore the need for continuous refinement of programme theory. These reviews provided valuable indications of areas requiring further development, whether in conceptual design or operational implementation. Moreover, they facilitate the identification of particularly transformation-relevant instruments and topics, enabling policymakers and researchers to focus on interventions with the greatest potential for driving systemic change.

The contributions of individual projects within R&I programmes to broader transformation processes also offer critical insights into transformation needs that warrant targeted policy attention. These contributions can reveal gaps in existing interventions, highlight emerging challenges, and suggest priorities for future action. However, addressing these needs requires sharpening measurement concepts and developing more precise empirical tools. Advancing methodologies for assessing transformation processes will enhance the ability to evaluate the effectiveness of R&I programmes, ensuring they remain relevant and impactful in addressing complex societal challenges.

REFERENCES

Almeida, D. V., Kolinjivadi, V., Ferrando, T., Roy, B., Herrera, H., Gonçalves, M. V., & Van Hecken, G. (2023). The "Greening" of Empire: The European Green Deal as the EU first agenda. Political Geography, 105, 102925.

Dinges, M., Coatanroch, G., Kerlen, C., Soloviy, V., Toepel, K., Montalvo Rojo, A. (2024). Navigating the Green Transition. Unveiling the Impact of European Framework Programmes for R&I through a Multi-Level-Perspective in Socio-Technical Transitions Research (Forthcoming).

Dinges, M., Kerlen, C., et al. (2023). Begleitevaluation der BMWK-Förderung im 7. Energieforschungsprogramm. Erster Zwischenbericht, Stand Juli 2023. https://www.energieforschung.de/lw_resource/datapool/systemfiles/elements/files/d8450fc8-77f0-11ee-ac3d-a0369fe1b534/current/document/Erster_Zwischenbericht_Unabh%C3%A4ngige_Begleitevaluation_7.EFP.pdf

EEA (2018). Perspectives on transitions to sustainability. European Environmental Agency, Report 25/2017, Luxembourg.

European Commission, Directorate-General for Research and Innovation, Dinges, M., Coatanroch, G., Evaluation study on the European framework programmes for research and innovation for addressing global challenges and industrial competitiveness – Focus on activities related to the green transition – Final report phase 1 – Horizon 2020, Dinges, M. (editor), Coatanroch, G. (editor), Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2777/422725.

European Commission: AIT, Directorate-General for Research and Innovation, Fraunhofer ISI, Kerlen Evaluation, Science Metrix, Technopolis Group, ZSI, Dinges, M., Coatanroch, G., Horizon Europe and the green transition – Interim evaluation support study – Final report ("Phase 2" study), Dinges, M.(editor) and Coatanroch, G. (editor), Publications Office of the European Union, 2024, https://data.europa.eu/doi/10.2777/67934.

European Environment Agency (2021), Growth without economic growth, Briefing no. 28/2020, ISBN 978-92-9480-320-7.

Geels, F.W., 2006. Multi-Level Perspective on System Innovation: Relevance for Industrial Transformation, in: Olsthoorn, X., Wieczorek, A.J. (Eds.), Understanding Industrial Transformation: Views from Different Disciplines, Environment & Policy. Springer Netherlands, Dordrecht, pp. 163–186. https://doi.org/10.1007/1-4020-4418-6_9.

Geels, F.W., Sovacool, B.K., Schwanen, T., Sorrell, S., 2017. The Socio-Technical Dynamics of Low Carbon Transitions. Joule 1, 463–479. https://doi.org/10.1016/j.joule.2017.09.018.

Ghosh, B., Kivimaa, P., Ramirez, M., Schot, J., Torrens, J., 2021. Transformative outcomes: assessing and reorienting experimentation with transformative innovation policy. Science and Public Policy 48, 739–756. https://doi.org/10.1093/scipol/scab045.

Hermoso, V., Carvalho, S. B., Giakoumi, S., Goldsborough, D., Katsanevakis, S., Leontiou, S., & Yates, K. L. (2022). The EU Biodiversity Strategy for 2030: Opportunities and challenges on the path towards biodiversity recovery. Environmental Science & Policy, 127, 263-271.

Kemp, R., Schot, J., & Hoogma, R. (1998). Regime shifts to sustainability through processes of niche formation: the approach of strategic niche management. Technology analysis & strategic management, 10(2), 175-198.

Kivimaa, P., Boon, W., Hyysalo, S., & Klerkx, L. (2019). Towards a typology of intermediaries in sustainability transitions: A systematic review and a research agenda. Research policy, 48(4), 1062-1075.

Leonard, M., Pisani-Ferry, J., Shapiro, J., Tagliapietra, S., & Wolff, G. B. (2021). The Geopolitics of the European Green Deal (No. 04/2021). Bruegel policy contribution.

Markard, J., Raven, R., & Truffer, B. (2012). Sustainability transitions: An emerging field of research and its prospects. Research policy, 41(6), 955-967.

Mastini, R., Kallis, G., & Hickel, J. (2021). A green new deal without growth?. Ecological Economics, 179, 106832.

Ossewaarde, M., & Ossewaarde-Lowtoo, R. (2020). The EU's green deal: a third alternative to green growth and degrowth?. Sustainability, 12(23), 9825.

Rip, A., & Kemp, R. (1998). Technological change. Human choice and climate change, 2(2), 327-399.

Schot, J. (1998). The usefulness of evolutionary models for explaining innovation. The case of the Netherlands in the nineteenth century. History and Technology, an International Journal, 14(3), 173-200.

AUTHORS

MICHAEL DINGES
Head of Competence Unit Transformation Governance
Center for Innovation Systems & Policy
AIT Austrian Institute of Technology GmbH
Giefinggasse 4, 1210 Vienna, Austria
Email: michael.dinges@ait.ac.at
ORCID: 0000-0003-0433-4318

CHRISTIANE KERLEN
Kerlen Evaluation ltd.
16 West Savile Gardens
Edinburgh, Scotland, EH9 3AB
Email: christiane.kerlen@kerlen-evaluation.de

SURYA KNÖBEL
Austrian Institute of Technology
Giefinggasse 4, 1210 Vienna, Austria
Email: surya.knoebel@ait.ac.at
ORCID: 0009-0000-9865-9273

KATHLEEN TOEPEL
Kerlen Evaluation Ltd
16 West Savile Gardens
Edinburgh EH9 3AB
Email: kathleen.toepel@kerlen-evaluation.de
ORCID: 0009-0002-2322-6777

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e10, 1-27
DOI: 10.22163/fteval.2025.702

ENHANCING EU POLICY THROUGH COMPLEXITY METRICS: A NEW LENS FOR RESEARCH AND INNOVATION

FLORENCE BENOIT, VALENTINA DI GIROLAMO, DARIO DIODATO, ERIK CANTON AND JULIEN RAVET
DOI: 10.22163/FTEVAL.2025.702

ABSTRACT

In a knowledge-based economy, understanding local capabilities is essential for identifying regional specialisations and technological trajectories. Recognizing where valuable knowledge resides and how innovation systems evolve is vital for enhancing the European Union's competitiveness and overcoming some of the multifaceted challenges the world is facing today. From a policy perspective, this understanding is a crucial input for research and innovation (R&I) policies. Effective R&I policies require data-driven decisions based on comprehensive analysis. However, traditional indicators tend to miss the necessary nuances of technological progress by focusing on the quantity instead of quality of knowledge output. This has prompted a growing interest in complementary quality-based metrics, including the concepts of complexity and relatedness. Complexity captures the diversity and interdependencies of economic activities, while relatedness measures the connections between different economic activities. This paper explains how incorporating these metrics can enhance the ability of the EU to foster economic growth and address societal challenges through the design of more impactful policies. In particular, the paper focuses on how these metrics can inform three EU policy priorities: safeguarding access to critical technologies, fostering the green transition and promoting greater territorial cohesion.

Keywords: Complexity, relatedness, R&I policy

1. INTRODUCTION

In today's economy, knowledge is a critical resource for long-term economic growth (Romer, 1990). It tends to concentrate in densely populated areas, where geographical proximity facilitates spillovers, rapid idea diffusion, and the recombination of capabilities. Much of this knowledge, however, remains tacit and context-dependent, thereby reinforcing the role of localised learning environments and institutional ecosystems (Polanyi, 1966; Lundvall, 1992; Cooke, 2001). This localised concentration can further be enriched by global knowledge flows through collaborations and networks. Through these processes, economies can obtain a set of capabilities that form the basis for the development of unique technological assets (Storper & Venables, 2004). These unique assets, which are difficult to replicate, become the cornerstone of a sustainable competitive advantage and contribute significantly to long-term economic development.

Consequently, it is important to understand the depth and the breadth of knowledge capabilities within an economy. Traditional innovation indicators often fall short due to their dominant focus on output quantity, which fails to capture the more qualitative aspects underlying knowledge development and application (Balland & Rigby, 2017). Newer quality-based indicators, such as citation-weighted impact and novelty metrics, offer important improvements, but do not assess how knowledge is structured, connected or embedded within the economy. Yet, understanding these patterns is important as they shape an economy's capacity to absorb new ideas, diversify into emerging sectors, and sustain long-term innovation-driven growth. In response to these limitations, the concepts of knowledge complexity and relatedness have been gaining more prominence. Complexity measures an economy's ability to produce a diverse range of sophisticated technologies, thereby emphasizing both, the variety of technologies it produces and their global rarity (Balland et al., 2022). Relatedness measures the degree of similarity between different economic activities based on the shared knowledge and competencies required for their production (Boschma, 2017). Together, these concepts provide a complementary and comprehensive framework for understanding the unique strengths of an economy's existing knowledge base.

Such insights offer valuable guidance for addressing challenges facing the European Union (EU), e.g., increasing sustainability (Santoalha et al., 2021; Sbardella et al., 2022; Mealy and Teytelboym, 2022; European Commission, 2024) and reskilling (Stephany and Teutloff, 2024). The complexity framework can help policymakers assess an economy's current capabilities and growth potential, highlight opportunities for diversification and design more targeted interventions. It also enables continuous monitoring to ensure that policies remain adaptive to changing economic and technological landscapes. This approach is particularly valuable for research and innovation (R&I) policies that aim to move beyond traditional models, which often focus solely on technological innovation and economic growth, to create more holistic policies capable of addressing grand societal challenges (Cavicchi et al., 2023).

This paper is structured as follows: Section 2 provides an overview of the theory and measurement of complexity, while Section 3 discusses how complexity can serve as a tool for R&I policy. Section 4 presents case studies demonstrating the potential of complexity in guiding current EU policy discussions, including the safeguarding of critical technologies, the acceleration of the green transition, and the promotion of territorial cohesion. Through these examples, this paper highlights the value of complexity for innovation policy practitioners.

2. COMPLEXITY: THEORETICAL BACKGROUND

Knowledge plays a key role in shaping economic systems and driving long-term economic growth (Romer, 1990). It accumulates through the exchange of ideas and the combination of diverse expertise, facilitated by interactions between individuals, firms and institutions located in close geographical proximity (Storper & Venables, 2004). This implies that knowledge is not created in isolation but through a systemic and interactive process embedded within networks of economic and institutional actors, where learning occurs through mutual engagement and contextual collaboration (Lundvall, 1992). This learning is rooted in specific places, giving rise to regional innovation systems in which innovation is fostered through local knowledge interactions (Cooke, 2001). Since much of this knowledge is tacit, it remains closely tied to its social and geographical context and cannot be easily codified or transferred (Polanyi, 1958; Polanyi, 1966). However, local knowledge bases can be further enriched by engaging within global knowledge flows through licensing agreements, collaborations or personal networks, which complement and build upon local capabilities (Archibugi & Michie, 1995; Doel & Hubbard, 2002; Bathelt et al., 2004).

As knowledge accumulates and diversifies, it creates a unique set of capabilities (Storper, 1997) essential for technological development. However, not all technologies are equally valuable and have the same growth potential. Advanced technologies, such as the Internet of Things, demand expertise across multiple domains like cloud computing, wireless communication and embedded systems. As a result, their development requires a deeper understanding and seamless integration of these diverse knowledge areas, making them inherently complex and challenging to replicate. These high-value, non-ubiquitous technologies (Nelson & Winter, 1982) offer significant competitive advantages due to their rarity and sophistication (Hidalgo & Hausmann, 2009). Understanding these capabilities is crucial for uncovering a territory's technological trajectory and revealing the geographical patterns of economic growth and development (Schumpeter, 1942; Romer, 1990; Pugliese et al., 2018; Pintar & Sherngell, 2022; Hidalgo & Hausmann, 2009; Hausmann et al., 2014; Tacchella et al., 2018).

The concept of knowledge complexity is central to these ideas. Knowledge complexity relates to economic complexity, which explains an economy's ability to produce and export a wide range of goods (Hidalgo & Hausmann, 2009). Knowledge complexity, however, focuses on an economy's capacity to produce a diverse range of sophisticated technologies. More precisely, it considers both, the variety and ubiquity of knowledge capabilities required for their production (Balland & Rigby, 2017). Since these capabilities are not directly observable, knowledge complexity is inferred using outcome-based approaches, such as analysing patent data (Antonelli et al., 2017; Ivanova et al., 2017).

Among the measures of knowledge complexity[1], the Knowledge Complexity Index (KCI) adapted from the Economic Complexity Index (ECI) (Hidalgo & Hausmann, 2009; Balland & Rigby, 2017), examines patent applications to measure technological diversity (the number of technologies in which an economy specialises) and ubiquity (the number of economies specialising in this technology). Higher values of KCI signify that an economy produces diverse technologies that are less commonly produced globally, thereby revealing a deeper knowledge base. Similarly, the Technology Complexity Index (TCI)[2] captures how complex a specific technology is by assessing how difficult it

1 Like with economic complexity, this literature has proposed different measures of knowledge complexity; the most prominent are Sbardella et al. (2018)'s Technological Fitness and Balland et al. (2019)'s Knowledge Complexity Index (KCI).

2 From a technical standpoint, KCI measures the complexity of a region/country's knowledge base considering how unique and diversified its innovation output is; whereas, TCI measures technological complexity by evaluating how specialised and widely distributed different technologies are across regions/countries (Balland & Rigby, 2017). More in general, the KCI of a location is defined as the average of the TCI of the technological activities (typically proxied by patent activities) that are located in it. Similarly, the TCI of a technology is defined as the average KCI of the locations where that technology is observed (Hidalgo, 2021). For more information see also The Observatory of Economic Complexity.

is to patent in that area. Hence, TCI zooms in on the complexity of individual technologies. Together, KCI and TCI indices can provide an indication of proximity to the technological frontier (Schetter, 2022).

Another closely related concept is technological relatedness, which measures the similarity between technologies based on the knowledge and competencies required to produce them (Boschma, 2017). Two technologies are considered related when they share similar knowledge or require overlapping skills (Hidalgo et al., 2018; Balland et al., 2019; Pugliese et al., 2019). Relatedness is often inferred from the geographic co-location of innovation activities, i.e., technologies that are often produced in the same location are considered similar as their co-location suggests that they leverage the same local capabilities. For example, economies with a strong robotics industry may also excel in the automotive industry due to overlapping expertise. In other words, relatedness provides insights into an economy's proximity to a specific technology, complementing traditional specialisation metrics like the Balassa Index[3] by identifying untapped opportunities for diversification.

Knowledge complexity and relatedness are often used together to provide a comprehensive understanding of an economy's knowledge base. While knowledge complexity reflects an economy's ability to produce sophisticated technologies, relatedness indicates how well it can expand into new, related areas. These concepts reinforce each other dynamically: advancing into related technologies increases knowledge complexity, which in turn facilitates further diversification. This cycle boosts the ability to innovate and adapt, contributing to long-term economic growth and resilience by enabling economies to engage in technological advancements and respond to changing global demand (e.g., Liao, 2015).

3 The Balassa Index, also known as the Revealed Comparative Advantage (RCA) index, is a measure that quantifies a country's comparative advantage in the production and export of specific goods or services. It helps identify sectors in which a country is relatively more competitive in international trade compared to others.

3. USING COMPLEXITY IN RESEARCH AND INNOVATION POLICY

The development of effective policies relies on evidence-based decision-making and clear data analysis. Traditional indicators based on patent counts, scientific publications and similar metrics offer valuable insights by measuring the quantity of innovation outputs making them useful for quick policy assessments, the monitoring of (relative) innovation performance and easy communication to the wider public (Hollanders et al., 2009). However, using the number of patents or publications implicitly assumes that they all have the same value, failing to fully capture all nuances of technological progress (Balland & Rigby, 2017), including potential heterogeneity in terms of underlying know-how and links to pre-existing specialisation patterns. Indeed, the dynamics of technological progress depend not only on output, but also on the sophistication and relationships within an economy's innovation ecosystem.

To address these gaps, policy frameworks are increasingly adopting a layered approach to indicators, combining output metrics with more quality-driven citation-weighted impact measures, novelty detection tools and composite indicators. Citation-weighted indicators (Aksnes et al., 2019) capture the influence of knowledge over time, showing how innovations contribute to future developments. Novelty metrics (Verhoeven et al., 2016), based on textual and structural analysis of patents, identify whether innovations represent incremental improvements or radical shifts. Composite indicators (Nardo et al., 2008), like the European innovation Scoreboard or the Global Innovation Index, go further by offering a broader view of innovation by integrating data on R&D inputs, collaboration and commercialisation.

Complexity metrics offer a distinct yet complementary perspective to the aforementioned indicators by focusing not just on innovation outputs, but also on how they are embedded and connected within the system. These indicators provide structural insight into technology linkages, assess the absorptive capacity, and offer a forward-looking perspective by identifying potential innovation pathways based on existing strengths. Complexity metrics are compiled based on matrix factorization, which help preserve these relational patterns, thereby enabling more nuanced analysis of innovation and growth potential (Hidalgo, 2021; Hidalgo & Hausmann, 2009; Hausmann et al., 2014). This makes the complexity framework particularly well-suited for guiding innovation policies by examining both the current state and future potential of an economy's innovation ecosystem.

The concept of relatedness further enhances this understanding. Relatedness assesses the feasibility of developing new technologies based on existing capabilities and network connections, highlighting that technologies closely linked to current strengths offer more feasible, less risky, and less costly opportunities for diversification (e.g., Frenken et al., 2007; Boschma & Frenken, 2011). These insights allow economies to strengthen their innovation performance by leveraging existing assets and helping to build on current strengths to explore new areas of innovation.

When combined with other factors, such as the social returns to research and innovation investments, these insights enable policymakers to craft tailored interventions that align with regional strengths and development trajectories. Indeed, although complexity indices have been mainly used to identify diversification paths (e.g., Hausmann et al., 2014; Deegan et al., 2021), it can also provide a forward-looking perspective to help create environments that support transitions into new, complex technological areas. This approach encourages policymakers to plan and organise long-term strategies for developing emerging technologies (e.g., Alshamsi et al., 2018; Waniek et al., 2020). By understanding how technologies are interconnected, policymakers can make more informed decisions about resource allocation, investments in skills and infrastructure, and potential collaborations to support the development or adoption of these technologies. Additionally, the complexity toolbox can enable continuous monitoring, allowing strategies to be adapted to evolving technological landscapes and economic conditions.

All these characteristics make complexity a powerful tool in the pursuit of ambitious EU policy objectives. By obtaining a clear understanding of the dynamic capabilities, more tailored interventions can be designed that allow Member States to build on their strengths with the purpose of enhancing overall innovation capacity, while mitigating the risks associated with technological and economic shifts. In addition, tailored advice can also help economies catch up with more advanced ones by identifying opportunities for investment that promote convergence. This approach enhances Europe's competitiveness, improves living standards (European Commission, 2024), reduces regional disparities, and increases cohesion through innovation.
As such, complexity thinking aligns well with the principles of National and Regional Innovation Systems, offering a way to map interdependencies and guide place-based strategies, but may require more adaptive and iterative policy tools than currently standard.

At the same time, the ability to structure and organise for long-term strategic development is particularly relevant for the EU's mission-oriented R&I policies aimed at addressing societal challenges, such as climate change and sustainable development. These challenges require a coordinated approach across various economic areas. Complexity can offer valuable insights to guide the development and coordination of the assets needed to address these challenges. This approach enhances the capacity to deliver innovations that are both technically feasible and socially meaningful, contributing to broader systemic change. However, they may also challenge conventional top-down governance approaches that favour linear planning.

Although promising, complexity's use in R&I policy is relatively recent. Initially, it was applied to national-level phenomena, such as economic growth (e.g., Hidalgo & Hausmann, 2009; Pugliese & Tacchella, 2021), income inequality (e.g., Chu & Hoang, 2020) and sustainability (e.g., Mealy & Teytelboym, 2022; Sbardella et al., 2022). These applications allowed policymakers to capture economic interdependencies and formalise principles of development, such as complementarity (e.g., Rosenstein-Rodan, 1943; Hirschman, 1977). As the limitations of linear innovation models became evident in addressing the complexity of technological systems, complexity theory began to shape R&I policies. Notably, it became a key component of the EU's Smart Specialisation Strategies (S3) under Cohesion Policy, moving away from one-size-fits-all approaches and instead fostering regional innovation by leveraging unique regional strengths.

4. CASE STUDIES: COMPLEXITY FOR A COMPETITIVE EUROPE

This section explores how complexity can be leveraged to address three critical transformations which the EU must navigate to secure its future competitiveness (Draghi, 2024): increasing technological sovereignty and reducing dependencies, advancing the green transition and closing the innovation divide. Case 1 highlights the usefulness of complexity metrics to identify key technologies for future growth and investment decisions while Case 2 explains how the European Green Deal can benefit from the same metrics by identifying regions with the potential for green technology development. Case 3 explores how complexity analysis can inform policy practitioners on how to reduce regional disparities in technological capabilities by fostering cross-border collaboration.[4]

CASE 1: COMPLEXITY AND TECHNOLOGICAL SOVEREIGNTY

In the last decade, political and economic shocks have challenged the standard globalisation growth model and its division of labour (European Commission, 2024). Protectionist policies and a revival of industrial policy are reshaping Global Value Chains (GVCs), aiming to reduce reliance on imports, while boosting national innovation, investments and growth (Aghion et al., 2023). At the same time, the growing securitisation and weaponisation of science and technology policies have intensified debates on how the EU can safeguard access to critical technologies and reduce foreign interferences. While ensuring the availability of critical technologies has always been a priority for policymakers, the approach to securing these technologies is evolving as free international technological cooperation and trade are undergoing significant restructuring (European Commission, 2024).

To address these challenges, complexity metrics provide policymakers with a data-driven lens towards technological sovereignty. As highlighted by Edler et al., (2023), an effective strategy for technological sovereignty begins with identifying which technologies are critical to the functioning of an economic system, followed by an assessment of the system's ability to access and develop them. Yet, this identification and assessment is complicated by the rapid pace of technological change, the expanding set of policy objectives as well as the general lack of data to inform development decisions. To this purpose, complexity measures can offer valuable guidance by identifying technologies with the highest growth potential, thereby providing data-driven insight on where to redirect public resources. For example, as shown in Figure 1, complexity analysis indicates that digital technologies, such as the Internet of Things (IoT), artificial intelligence (AI) and cybersecurity stand out as complex and hard to replicate fields. These are precisely the fields where strategic investment could offer significant growth and competitiveness advantages (Balland & Rigby, 2017; European Commission, 2024).

4 All cases are syntheses of existing studies, performed by the authors, curated to support the policy argument.

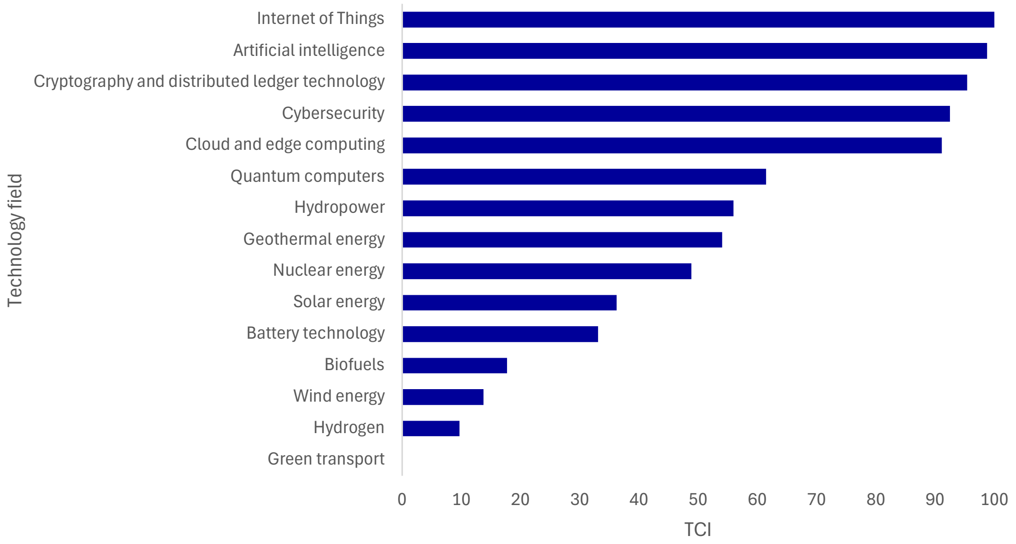


Figure 1. The complexity of key strategic technologies

Note: On the y-axis, technologies are ranked by Technology Complexity Index (TCI), which measures complexity at the technology level, normalised between 0 and 100. Source: The Science Research and Innovation Performance Report (SRIP) 2024.

Once critical technologies are identified, it is key to monitor their development over time and assess both current capacity and future potential. Indicators using patent counts or specialisation indices clearly highlight that the EU's technological performance in digital fields has been weakening thereby broadening the gap with competitors like the US and China (European Commission, 2024). However, these metrics are backward-looking and offer limited insights into the EU's future technological potential. In this context, relatedness indicators offer a more forward-looking lens, identify technological trajectories that a country is more likely to pursue based on current strengths.

As shown in Figure 2, the EU faces a significant technological gap compared to other key players in technologies such as Internet of Things, AI, blockchains, cybersecurity and quantum computers - not only in terms of current specialisation, but also when looking at the diversification potential in these fields (captured by the relatedness density indicator). This implies that the EU's current ability to build up capacity in such technologies is limited (European Commission, 2024). On the contrary, the EU's current specialisation is stronger in technologies such as wind energy, hydrogen, and green transports, whereas a higher potential for technological development is observed in other important green fields (e.g., hydropower, geothermal energy). This is a crucial aspect, as indicators such as relatedness density can be used as a tool to identify the types of strategic technologies in which the EU could better leverage its existing capabilities for further specialisation, thereby helping to optimise investment priorities and improve the efficiency of public support.

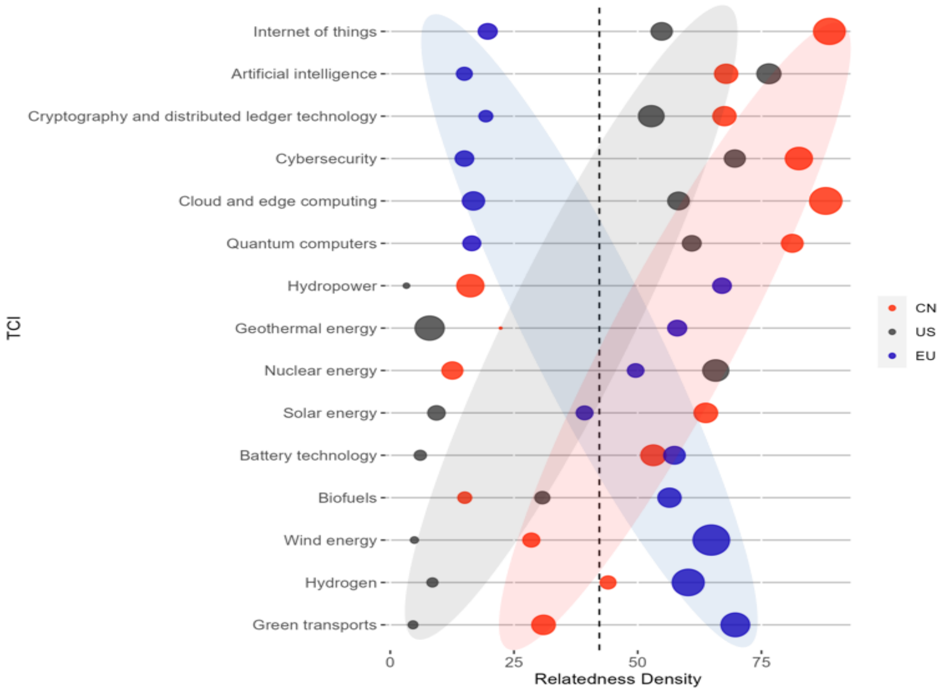


Figure 2. The EU's position in complex technologies versus the US and China, 2019-2022

Note: The x-axis indicates the relatedness density in any of the technology fields considered. On the y-axis, technologies are ranked by complexity levels, normalised between 0 and 100. The size of the bubble captures the degree of specialisation that each country reports in a given technology field, as measured by the revealed comparative advantage (RCA). Source: The Science Research and Innovation Performance Report (SRIP) 2024.

To further build up technological capacities and to guarantee access to critical technologies, the EU can either deepen its domestic innovation capabilities or rely on international sources of knowledge to acquire new capabilities (e.g., via collaborations) (Boschma, 2005; Edler et al., 2023). While the EU has long prioritised openness and international collaboration in science and technology, there exists a natural tension between the priority of safeguarding the EU's technological sovereignty and fostering international R&I cooperation (European Commission, 2024). This calls for empirical approaches able to support policymakers in identifying potential international partners from which the EU can gain in terms of technological complementarities, helping diversify the EU's partners pool and reduce the risk of exacerbating one-sided dependencies. As showcased by Figure 3, relatedness metrics can be used also to this purpose[5], as they can provide insights on the extent to which non-EU countries can complement the EU's technological deficiencies in different technology fields, especially more complex ones.

5 In this context, the concept of 'relatedness added' can be used to capture technological capabilities around a given technology that are missing in a country and that are available in other countries. For more information, please refer to Balland & Boschma (2021).

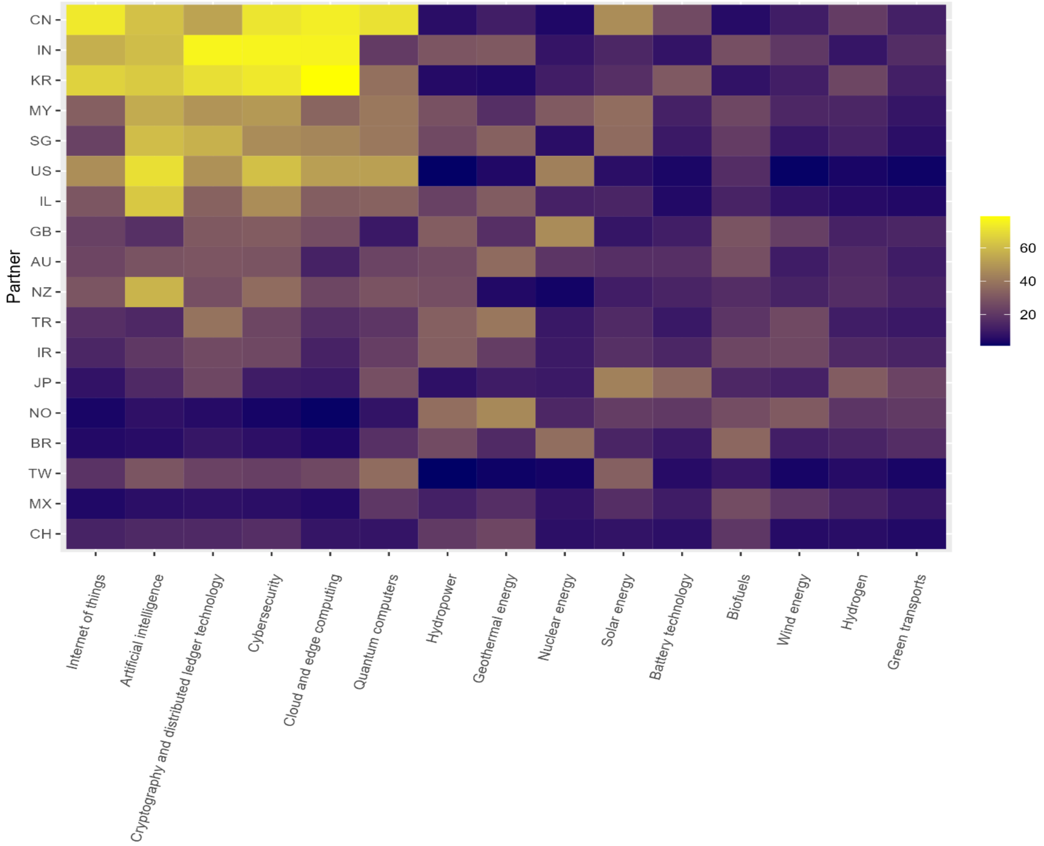


Figure 3. The EU's technological complementarities

Note: On the x-axis, technologies are ordered according to the degree of technology complexity (TCI index). On the y-axis, countries are ranked according to the average relatedness density added. Example of interpretation: South Korea, India and China show strong specialisation in technologies which are closely related to cloud and edge computing, and in which the EU shows weaker specialisation.
Source: The Science Research and Innovation Performance Report (SRIP) 2024.

CASE 2: COMPLEXITY AND FOSTERING THE GREEN TRANSITION

The European Green Deal aims to offset greenhouse gas emissions by 2050 while enhancing economic growth. To meet carbon neutrality goals, the EU will have to accelerate the development of climate-related technologies as climate targets cannot be met by only relying on existing technologies. In this case, complexity and relatedness metrics can provide guidance regarding the direction of policy intervention by evaluating which green technologies have the potential to be developed in the EU and which areas are better placed to do so based on their existing capabilities. This type of analysis can thus provide insights into identifying investment opportunities to develop a particular green technology and on which green technologies the EU should be focusing on.

The literature on economic geography has long argued that regions, often functioning as clusters of knowledge exchange, are the most apt unit of analysis when thinking about capabilities, accumulation of know-how, specialisation, innovation and diversification (Glaeser et al., 1992). It is therefore likely that regional policies will increasingly embed European Green Deal objectives. Vice versa, it is also likely that European Green Deal policies will look towards regions for sources of innovation.

In this context, the framework of Economic Complexity proves particularly relevant. There are already many studies that apply the paradigm of economic complexity to understand the technological evolution towards the green transition. In the academic literature, for instance, Sbardella et al. (2018) calculate the complexity of green technologies. Mealy and Teytelboym (2022), instead, propose indices of green complexity and green potential in traded commodities. Caldarola et al. (2024) review these and other contributions to document the emergence of the economic complexity approach to analyse the sustainable transition. This rise is reflected also in the policy discourse, especially at the regional level. In a JRC policy brief, Sbardella et al. (2022) analyse the green potential of European regions. This report builds a mapping of relatedness between non-green and green technologies to assess the green potential of EU regions, based on their non-green technologies (as illustrated in Figure 4).

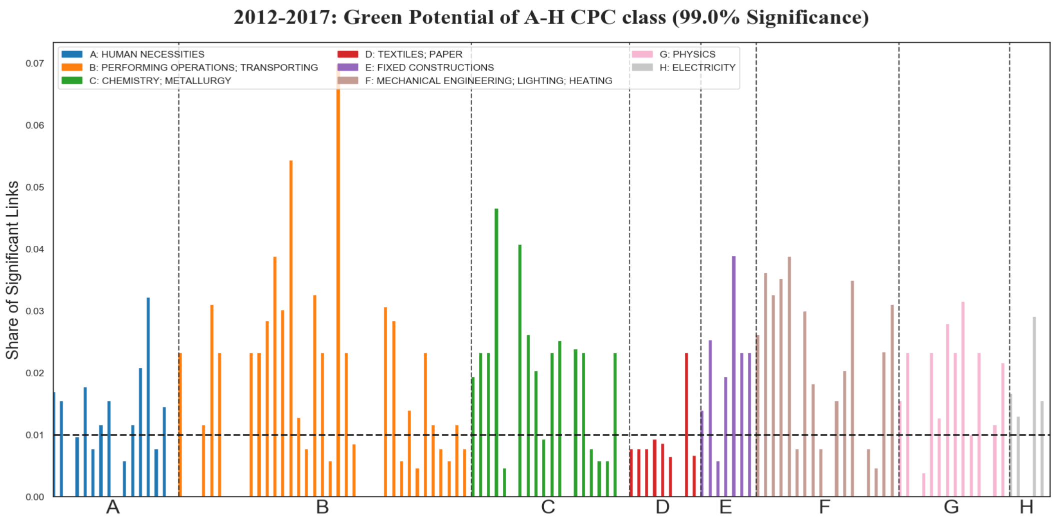


Figure 4. Relatedness between non-green and green technologies

Note: Share of 99% statistically significant links in the non-green—green technology space of each A-H CPC non green technology at 4-digit aggregation to all Y02 green technologies at 8-digit aggregation level. Source: JRC policy brief (Sbardella et al., 2022).

More recently, the Science, Research and Innovation Performance Report (European Commission, 2024; Chapter 9) focuses on specific green technologies. The analysis finds that, for instance, the EU is lagging in climate-change mitigation technologies related to aeronautics. These are technologies such as efficient propulsion systems or drag reduction techniques. The analysis then assesses the capacity of European regions in this domain as well as their potential, again based on the relatedness between the target green technology and the non-green technologies that do exist in EU regions. Both pieces of information are depicted in Figure 5. The regions in light and dark blue already have high patenting activity in green aeronautics. The regions in yellow and orange, instead, currently have no capacity in this technology, but medium to high potential, given their current specialisation in related technologies.

This analysis is useful for national or super-national policies that target technologies of strategic importance. It can, in fact, help identify regions with the greatest potential for development and impact. The analysis is also useful for regional and cohesion policies as it can identify new pathways – or untapped opportunities, which is how these are often refer to in the Smart Specialisation literature – for development in lagging regions.

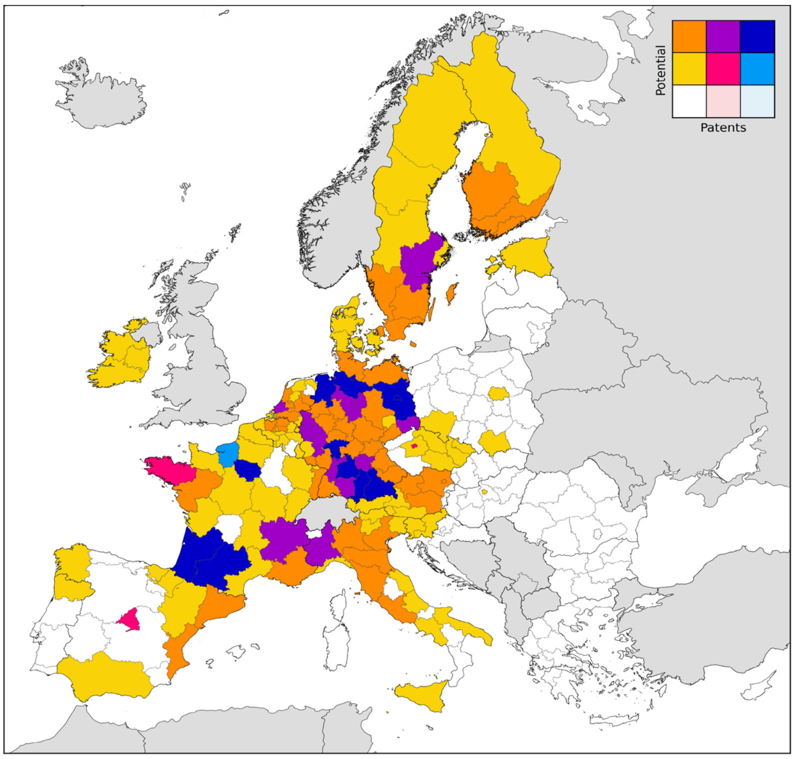


Figure 5. Map for green technology "aeronautics"

Source: The Science Research and Innovation Performance Report (SRIP) 2024.

As a matter of fact, the ideas of Economic Complexity resonate well with Smart Specialisation. Both approaches are place-based, aiming at exploiting local capabilities to foster new technological trajectories; both are vertical, meaning that investments should be targeted to a limited number of economic or technological areas and both are based on the idea that unknown, untapped opportunities exist (Diodato et al, 2023). As Foray (2015) argues, Smart Specialisation is "the capacity of an economic system (a region for example) to generate new specialties through the discovery of new domains of opportunity and the local concentration and agglomeration of resources and competences in these domains."

CASE 3: COMPLEXITY AND R&I CONNECTIVITY NETWORKS

Technological and innovative capabilities vary significantly across the EU, resulting in a concentration of advanced technologies in certain regions. This is especially evident in the development of complex technologies, which require multidisciplinary expertise and cross-border collaborations. Indeed, as technologies become more complex, they increasingly rely on diverse knowledge inputs from multiple regions and sectors, creating a greater need for more interconnected R&I ecosystems (e.g., Balland & Rigby, 2017).

Consequently, regions that can effectively integrate into international collaborative networks gain a clear advantage in developing and scaling complex technologies (Fleming & Sorenson, 2001; Balland & Rigby, 2017). In fact, there is a correlation between the ranking by complexity index of a specific technology category and its level of European inter-country collaborations, with a stronger correlation observed for more complex technologies (see Figure 6). However, cross-border cooperation in the EU remains limited, where the regional co-patenting network is fragmented along national lines (European Commission, 2024), hindering the sharing of knowledge and resources necessary for advancing complex technologies such as IoT, blockchain and cybersecurity. This fragmentation also exacerbates regional disparities, as innovation remains concentrated in leading regions that already possess the necessary expertise and infrastructure, while others struggle to catch up.

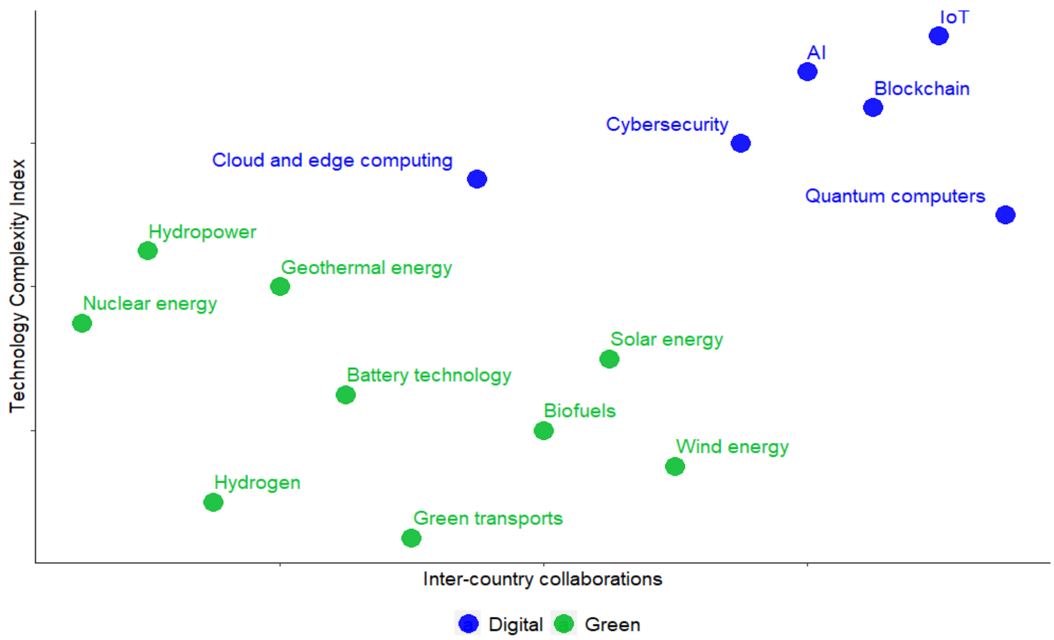


Figure 6. European inter-country collaborations by technology ranked according to complexity index, 2014-2020

Source: The Science Research and Innovation Performance Report (SRIP) 2024.

Bridging these regional gaps and fostering more integrated, cross-border innovation ecosystems is crucial for ensuring that Europe remains competitive in the global technological landscape. R&I policies play an important role in promoting international collaborations and knowledge diffusion across borders (European Commission, 2022), helping to overcome traditional barriers to knowledge exchange and enabling regions to benefit from collective expertise. One example is the initiatives under the Framework Programme for R&I, which aims to align regional strengths with broader European objectives, enabling less-developed regions to contribute to complex technological advancements.

Nevertheless, in practice, the increasing complexity of knowledge that is being produced and the speed at which new technologies are evolving can have a profound impact on achieving truly inclusive collaborations. Indeed, despite the EU's inclusive objectives, Balland et al., (2019b) show that pre-2004 member states are more frequently positioned as central players in high-complexity projects while post-2004 member states tend to participate in lower-complexity projects. This division risks reinforcing a spatial concentration of complex knowledge within pre-2004 member states. Without targeted strategies to support capacity-building in post-2004 member states, this cycle may continue, limiting these regions' ability to engage meaningfully in the EU's broader technological goals.

Complexity analysis can support this capacity-building by identifying regional strengths and weaknesses, revealing synergies between regions with complementary expertise. This can help guide targeted EU interventions to bridge gaps, strengthen knowledge networks and create more inclusive innovation ecosystems. By highlighting regions' relative positions in knowledge networks, complexity metrics can help policymakers create partnerships that ensure all regions can meaningfully engage in high-complexity technological development. This approach enables the EU to strategically allocate resources, promote cross-regional collaborations and support a more resilient and balanced technological landscape.

LIMITATIONS OF COMPLEXITY

To ensure the correct application and interpretation of complexity for policy purposes, it is important to recognise several limitations. First, although the complexity metrics presented in this paper go beyond traditional well-established metrics, they are still largely based on patent data, which may underestimate local capabilities, as not all knowledge is captured in patents. This is a well-known limitation of the use of patent data for analyses related to the measurement of knowledge and innovation. However, standard patent analysis perceives patents simply as an output value, while complexity uses patents as a proxy to identify technological specialisation irrespective of the overall absolute patent production. Complexity should therefore be less affected by this limitation (Diodato et al., 2023). Second, complexity is more effectively analysed when historical data on technologies are abundant and may be less accurate for emerging technologies, where it may struggle to capture rapid changes. Third, the use of patents to assess technological opportunities assumes that the region or country can always enter in the development of a technology if it possesses the necessary know-how. However, even with the required capabilities, territories can choose not to be active in a field for various reasons. Hence, while complexity can be used as an instrument to assess capabilities or technological opportunities, it does not provide direct solutions for the most appropriate strategy based on specific territorial or sector characteristics.

CONCLUSION

In the rapidly evolving global economy, the EU faces multifaceted challenges in ensuring its future competitiveness while encouraging inclusivity and sustainability. Navigating these challenges demands innovative approaches that can complement more traditional innovation metrics. Knowledge complexity and relatedness can be considered pivotal frameworks in this context, offering powerful tools to understand and help improve economic and innovation policies. Both concepts emphasise the dynamic and interconnected nature of knowledge accumulation and technological specialisation. Complexity metrics reveal existing strengths and new opportunities for diversification and growth, while relatedness can identify synergies between current capabilities and new technologies. Therefore, these concepts can contribute to the formation of more tailored, data-driven interventions that align with regional strengths and promote diversification.

REFERENCES

Aghion, P., Ahuja, K., Bown, C.P., Cantner, U., Criscuolo, C., Dechezlepretre, A., Dewatripont, M., Hausmann, R., Lalanne, G., McWilliams, B. et al., (2023). Sparking europe's new industrial revolution-a policy for net zero growth and resilience, Technical report, Bruegel.

Aksnes, D., Langfeldt, L. Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. SAGE Open, 9(1).

Alshamsi, A., Pinheiro, F. L., Hidalgo, C. A., (2018). Optimal diversification strategies in the networks of related products and of related research areas. Nat. Commun, 9, pp. 1328.

Antonelli, C., Crespi, F., Mongeau Ospina, C., Scellato, G., (2017). Knowledge complexity, Jacobs externalities and innovation performance in European regions. Reg. Stud., 51(11), pp. 1708-1720.

Archibugi, D., Michie, J., (1995). The globalisation of technology: a new taxonomy. Cambridge journal of Economics, 19(1), pp. 121-140.

Balland, P. A., Rigby, D., (2017). The geography of complex knowledge, Economic geography, 93(1), pp. 1–23.

Balland, P.A., Boschma, R., Crespo, J., Rigby, D.L., (2019). Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. Reg. Stud, 53 (9), pp. 1252–1268.

Balland, P. A., Boschma, R., Ravet, J. (2019b). Network dynamics in collaborative research in the EU, 2003–2017. European Planning Studies, 27(9), 1811–1837.

Balland, P. A., Broekel, T., Diodato, D., Giuliani, E., Hausmann, R., O'Clery, N., Rigby, D., (2022). The new paradigm of economic complexity, Research Policy, 51(3).

Boschma, R., (2005). Proximity and innovation: a critical assessment. Reg. Stud, 39, pp. 61–74.

Boschma, R., Frenken, K., (2011). The emerging empirics of evolutionary economic geography. J Econ Geogr, 11, pp. 295–307.

Boschma, R., (2017). Relatedness as driver of regional diversification: A research agenda, Regional Studies, 51(3), pp. 351-364.

Bathelt, H., Malmberg, A., Maskell, P., (2004). Clusters and knowledge: Local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography, 28(1), pp. 31-56.

Caldarola, B., Mazzilli, D., Napolitano, L., Patelli, A., Sbardella, A., (2024). Economic complexity and the sustainability transition: A review of data, methods and literature. Journal of Physics: Complexity, 5.

Cavicchi, B., Peiffer-Smadja, O., Ravet, J., Hobza, A., (2023). The transformative nature of the European framework programme for research and innovation: analysis of its evolution between 2002-2023, European Commission, Directorate-General for Research and Innovation, Publications Office of the European Union.

Chu, L.K., Hoang, D.P., (2020). How does economic complexity influence income inequality? New evidence from international data. Economic Analysis and Policy, 68, pp. 44–57.

Cooke, P., (2001). Regional innovation systems, clusters, and the knowledge economy. Industrial and corporate change, 10(4), pp. 945-974.

Deegan, J., Broekel, T., Fitjar, R.D., (2021). Searching through the Haystack: The Relatedness and Complexity of Priorities in Smart Specialization Strategies. Economic Geography, 97, pp. 497–520.

Diodato, D., Napolitano, L., Pugliese, E., Tacchella, A., (2023). Economic Complexity for Regional Industrial Strategy. JRC Science for Policy Brief – Industrial Innovation & Dynamics Series. No. JRC136443. European Commission, Joint Research Centre – Directorate for Growth and Innovation, Seville (Spain), December 2023.

Doel, M., Hubbard, P., (2002). Taking World Cities Literally: Marketing the City in a Global Space of Flows. City. 6, pp. 351-368.

Edler, J., Blind, K., Henning, K., Schubert, T., (2023). Technology sovereignty as an emerging frame for innovation policy: Defining rationales, end and means. Research Policy, 52(6).

European Commission (2024). Science Research and Innovation Performance Report. European Commission, Directorate-General for Research and Innovation, Publications Office of the European Union.

Foray, D., (2015). Smart specialisation: opportunities and challenges for regional innovation policy. Routledge: Abingdon.

Frenken, K., Oort, F. V., Verburg, T., (2007). Related variety, unrelated variety and regional economic growth. Reg. Stud., 41, pp. 685–697.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., Shleifer, A., (1992). Growth in cities. Journal of political economy, 100(6), pp. 1126-1152.

Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M., Simoes, A., (2014). The Atlas of Economic Complexity: Mapping Paths to Prosperity. MIT Press.

Hidalgo, C. A., Hausmann, R., (2009). The building blocks of economic complexity, Proc. Natl. Acad. Sci. U.S.A., 106 (12), pp. 10570-10575.

Hidalgo, C.A., (2021). Economic complexity theory and applications. Nature Reviews Physics, pp. 1– 22.

Hirschman, A., (1977). The Passions and the Interests: Political Arguments for Capitalism before its Triumph. Princeton: Princeton University Press, pp. 153.

Ivanova, I., Strand, Ø., Kushnir, D., Leydesdorff, L., (2017). Economic and technological complexity: a model study of indicators of knowledge-based innovation systems. Technol. Forecast. Soc. Change, 120, pp. 77–89.

Liao, W., Gu, J., Li, K., (2025). Roles of related and unrelated external technologies in shaping regional breakthrough technological advantages. Technological Forecasting and Social Change, 210.

Lundvall, B. A., (1992). National systems of innovation: Towards a theory of innovation and interactive learning. Francis Printer.

Mealy, P., Teytelboym, A., (2022). Economic complexity and the green economy, Research Policy, 51(8).

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffmann, A., Giovannini, E. (2008). Handbook on constructing composite indicators: methodology and user guide. OECD publishing, Paris.

Nelson, R., Winter, S., (1982). An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA.

Pintar, N., Scherngell, T., (2021). The complex nature of regional knowledge production: Evidence on European regions, Research Policy, 51(8).

Polanyi, M., (1958). Personal Knowledge. Towards a Post-Critical Philosophy, London: Routledge & Kegan Paul.

Polanyi, M., (1966). The Tacit Dimension, London: Routledge & Kegan Paul.

Pugliese, E., Chiarotti, G. L., Zaccaria, A., Pietronero, L., (2018). Complex economies have a lateral escape from the poverty trap. PLoS One, 12(1).

Pugliese, E., Cimini, G., Patelli, A., Zaccaria, A., Pietronero, L., Gabrielli, A., (2019). Unfolding the innovation system for the development of countries: coevolution of Science, Technology and Production. Scientific reports, 9(1), pp. 16440.

Pugliese, E., Tacchella, A., (2021). Economic complexity analytics: Country factsheets. Joint Research Centre (Seville site).

Romer, PM., (1990). Endogenous Technological Change, Journal of Political Economy, 98(5), pp. 71-102.

Rosenstein-Rodan, P.N., (1943). Problems of industrialisation of eastern and south-eastern Europe. The economic journal, 53, pp. 202–211

Santoalha, A., Consoli, D., Castellacci, F., (2021). Digital skills, relatedness and green diversification: A study of European regions, Research Policy, 50(9).

Sbardella, A., Perruchas, F., Napolitano, L., Barbieri, N., Consoli, D., (2018). Green technology fitness. Entropy, 20(10).

Sbardella, A., Barbieri, N., Consoli, D., Napolitano, L., Perruchas, F. Pugliese, E., (2022). The regional green potential of the European innovation system (No. JRC124696). Joint Research Centre.

Stephany, F., Teutloff, O., (2024). What is the price of a skill? The value of complementarity, Research Policy, 53(1).

Storper, M., (1997). The regional world: Territorial development in a global economy. New York: Guilford.

Storper, M., Venables, A., (2004). Buzz: face-to-face contact and the urban economy, Journal of Economic Geography, 4(1), pp. 351-370.

Schetter, U., (2022). A Measure of Countries’ Distance to Frontier Based on Comparative Advantage, CID Research Fellow and Graduate Student Working Paper (135).

Schumpeter, J.A., (1942). Capitalism, Socialism and Democracy. Vol. 36, Harper & Row, New York, pp. 132-145.

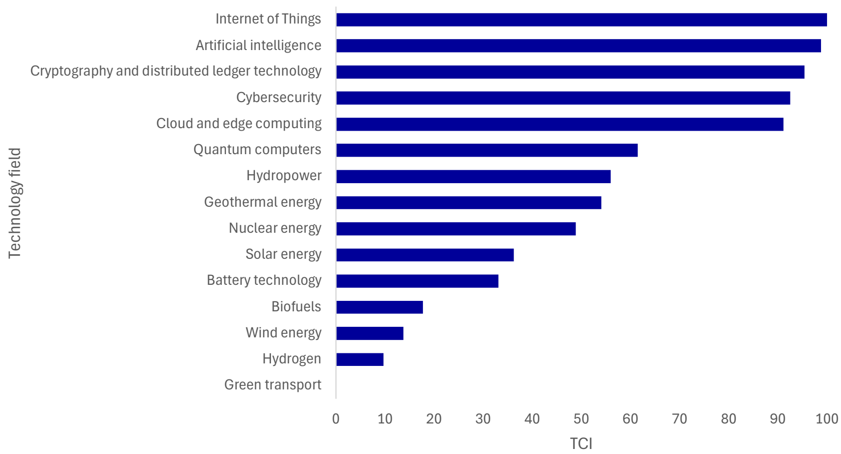
Tacchella, A., Mazzilli, D., Pietronero, L., (2018). A dynamical systems approach to gross domestic product forecasting. Nat. Phys. 14, pp. 861–865.

Verhoeven, D., Bakker, J., Veugelers, R., (2016). Measuring technological novelty with patent-based indicators. Research Policy, 45(3), pp. 707-723.

Waniek, M., Elbassioni, K., Pinheiro, F. L., Hidalgo, C. A., Alshamsi, A., (2020). Computational aspects of optimal strategic network diffusion. Theor. Comput. Sci, 814, pp. 153–168.

FIGURES

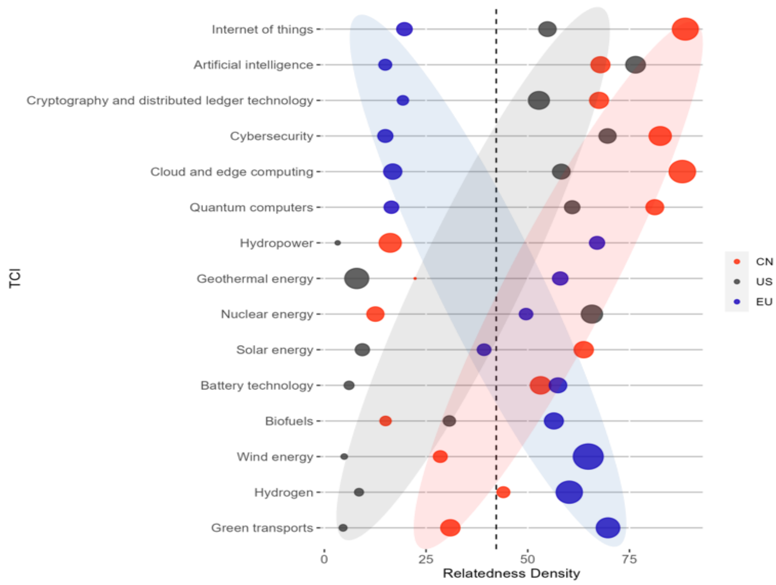
Figure 1. The complexity of key strategic technologies



Note: On the y-axis, technologies are ranked by Technology Complexity Index (TCI), which measures complexity at the technology level, normalised between 0 and 100.

Source: The Science Research and Innovation Performance Report (SRIP) 2024.

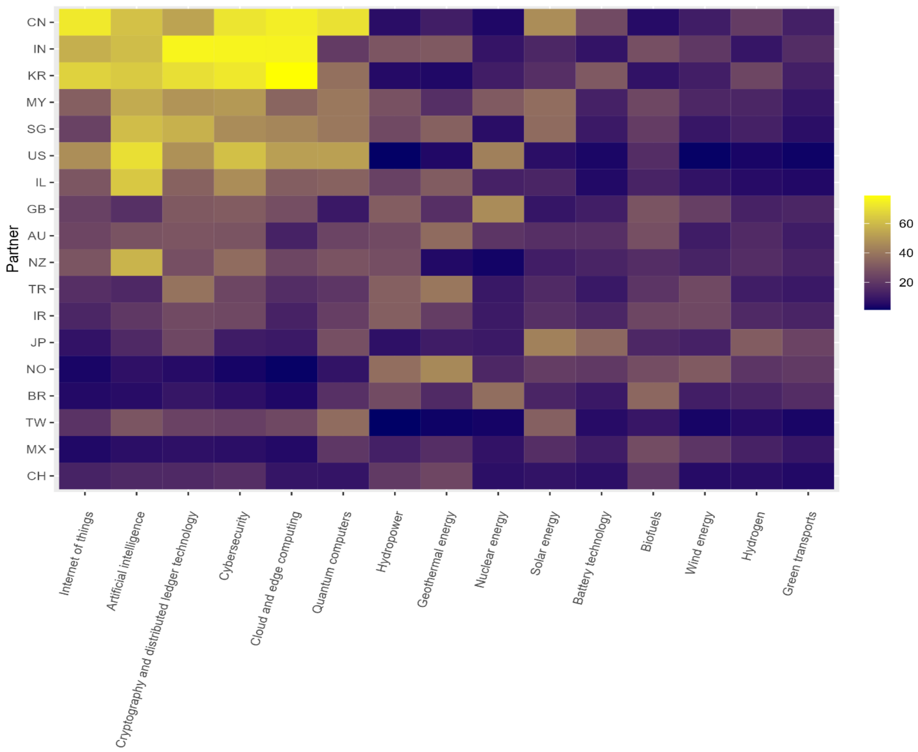
Figure 2. The EU's position in complex technologies versus the US and China, 2019-2022



Note: The x-axis indicates the relatedness density in any of the technology fields considered. On the y-axis, technologies are ranked by complexity levels, normalised between 0 and 100. The size of the bubble captures the degree of specialisation that each country reports in a given technology field, as measured by the revealed comparative advantage (RCA).

Source: The Science Research and Innovation Performance Report (SRIP) 2024.

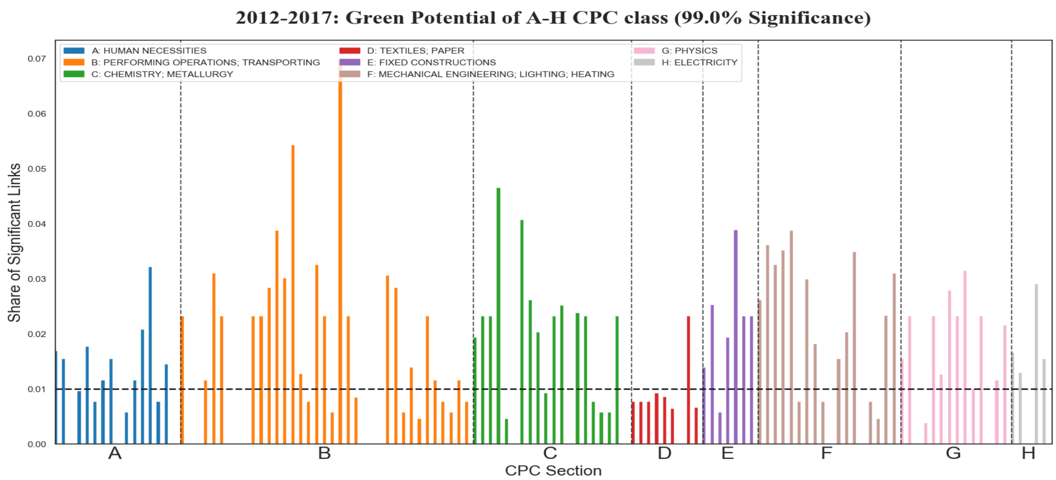
Figure 3. The EU's technological complementarities



Note: On the x-axis, technologies are ordered according to the degree of technology complexity (TCI index). On the y-axis, countries are ranked according to the average relatedness density added. Example of interpretation: South Korea, India and China show strong specialisation in technologies which are closely related to cloud and edge computing, and in which the EU shows weaker specialisation.

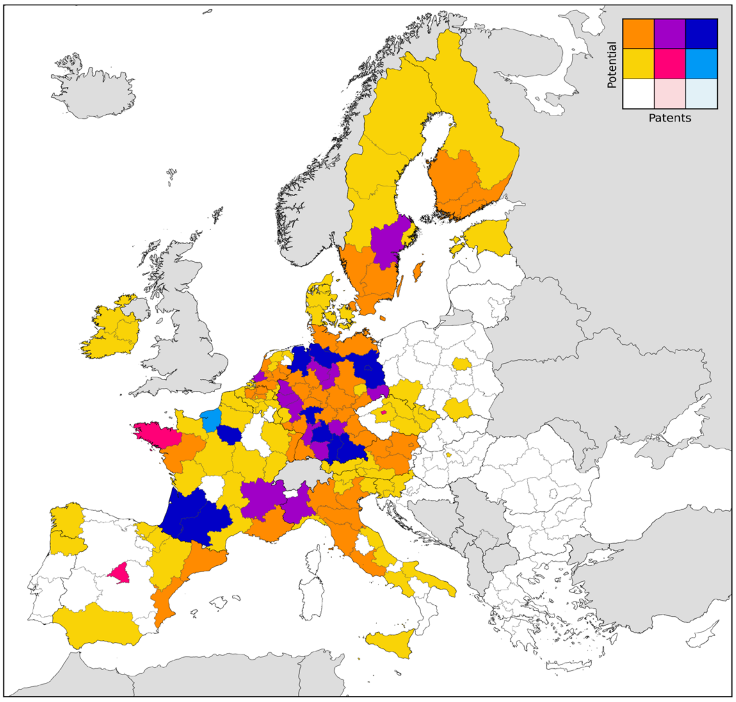
Source: The Science Research and Innovation Performance Report (SRIP) 2024.

Figure 4. Relatedness between non-green and green technologies



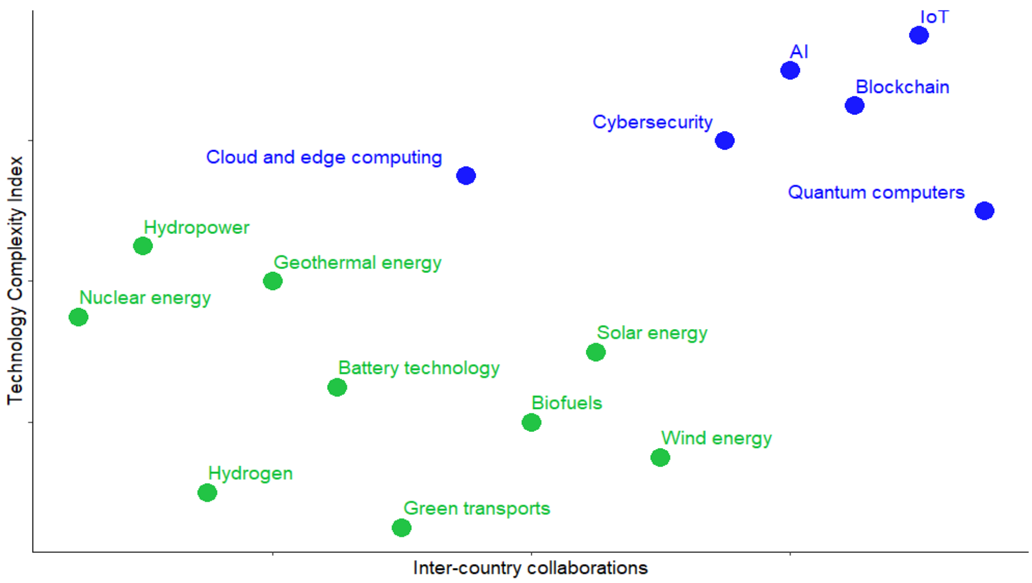
Note: Share of 99% statistically significant links in the non-green—green technology space of each A-H CPC non green technology at 4-digit aggregation to all Y02 green technologies at 8-digit aggregation level. Source: JRC policy brief (Sbardella et al., 2022).

Figure 5. Map for green technology "aeronautics"



Source: The Science Research and Innovation Performance Report (SRIP) 2024.

Figure 6. European inter-country collaborations by technology ranked according to complexity index, 2014-2020



Source: The Science Research and Innovation Performance Report (SRIP) 2024.

AUTHORS

FLORENCE BENOIT
Directorate-General for Research and Innovation (DG RTD)
European Commission, Brussels, Frère Orbansquare 8
Email: Florence.BENOIT@ec.europa.eu

VALENTINA DI GIROLAMO
Directorate-General for Research and Innovation (DG RTD)
European Commission, Brussels, Frère Orbansquare 8
Email: Valentina.DI-GIROLAMO@ec.europa.eu

DARIO DIODATO
Joint Research Centre (DG JRC)
European Commission, Seville, Calle Inca Garcilaso 3
Email: Dario.DIODATO@ec.europa.eu
ORCID: 0000-0001-6902-2468

ERIK CANTON
Directorate-General for Research and Innovation (DG RTD)
European Commission, Brussels, Frère Orbansquare 8
Email: Erik.CANTON@ec.europa.eu
ORCID: 0000-0001-6101-8333

JULIEN RAVET
Directorate-General for Research and Innovation (DG RTD)
European Commission, Brussels, Frère Orbansquare 8
Email: Julien.RAVET@ec.europa.eu

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e11, 1-12
DOI: 10.22163/fteval.2025.703

EVALUATING "LEARNING AND EXPERIMENTAL SPACES"

WHEN A TRADITIONAL APPROACH REACHES ITS LIMITS

CHRISTINA SCHUH, DANIEL SCHWERTFEGER AND SONJA FRINGES
DOI: 10.22163/FTEVAL.2025.703

ABSTRACT

The paper reflects on the evaluation of a case study that seeks innovative solutions for digital transformation. In particular, it discusses the tension between traditional evaluation approaches and new perspectives on the funding process in general, the used methods and the changing role of the evaluator. The subject of the evaluation is the funding guideline of the German Federal Ministry of Labour and Social Affairs: "*Sustainable companies and administrations in digital change*". The Learning and experimental spaces (LES) funding instrument supports small and medium-sized enterprises in developing innovative, tailor-made and consensual solutions for employees and companies in the digital transformation. These solutions must also be supported by social partnerships. The several LES should be of a fundamentally exemplary nature and transfer to further innovative solutions. Funding was provided for 17 LES in the first funding round (starting 2018) and a further 11 LES with focus on the use of artificial intelligence (AI) in the second funding round (starting 2020). The duration of each funded project was about three years. Each individual project was evaluated externally and additionally subjected to an overall evaluation by the Federal University of Applied Administrative Sciences (HS Bund), which included all projects of the respective funding round. The paper first provides a theoretical framework for the background of the funding guideline. The second part presents the LES funding guideline and its evaluation, including some project examples. Then we describe and discuss the evaluation process, using five factors to compare traditional evaluation approaches and new perspectives on the funding process and the changing role of the evaluator before we end with our conclusion.

Keywords: Learning and experimental spaces, small and medium-sized enterprises, linking digital transformation and social innovation, consensual solutions, developmental evaluation

1. INTRODUCTION

For several decades, the need to adapt innovation policy as well as the regarding evaluation paradigms and practices to the changed societal and environmental problems has been in the focus of attention. Since the mid-20th century, at least two paradigm shifts took place in innovation policy (IP) and along with corresponding changes in methods of assessment and evaluation (Schot & Steinmueller, 2018; Rohracher, Coenen and Kordas, 2022). Early IPs focused on innovation mainly for economic growth, prosperity and mass production, whereas after the first shift, IPs started dealing with international competition and the link between discovery and application. After another important paradigm shift, the focus is now on transformative change, that is, addressing major global and societal challenges such as the Sustainable Development Goals. Thus, for some time now, we have increasingly been confronted with wicked problems (Reale, 2021) which do not allow for a single solution (if there is even one) or predefined solutions (Rittel & Webber, 1973). As Schot and Steinmueller (2018) conclude, the model of innovation must be experimental in this paradigm. Despite this insight, a gap often remains between the claim of transformative IPs and their actual implementation, including the practices of their evaluation (Rohracher, Coenen & Kordas, 2022).

In this practice report, we describe our experiences that the implementation of interventions and funding as well as the expectations of (some of the) stakeholders can make it difficult to apply flexible evaluation methods. The subject of our report is a funding guideline from the German Federal Ministry of Labour and Social Affairs entitled “Sustainable companies and administrations in digital change”. A total of 28 projects were funded in this program, which were designated *Learning and experimental spaces* (LES), as they focus on learning experiences rather than products. The application of flexible evaluation methods is essential to address a key characteristic of wicked problems, namely, that each is a one-shot problem (Rittel & Webber, 1973). To gain knowledge of promising innovation pathways, one needs to gather the experiences of different actors with different perspectives (Schot & Steinmueller, 2018). This is far from trivial in the case (presented below), where an overarching evaluation was conducted across several distinct and highly diverse projects.

An example of an approach, that could meet the requirements of the transformative paradigm is, developmental evaluation (Patton, 2010), which allows flexible application of methods and an active role of the evaluator. It takes a systems- and innovation-oriented approach and focuses on adapting interventions to changing contexts, target groups, or emerging needs. Therefore, developmental evaluation involves flexible designs, flexible relationships, flexible budgeting and flexible reporting. Accountability in this approach is extended to accountability for learning, development and adaption, and evaluators are supposed to be part of the evaluation team (Patton, 2015).

2. OVERALL EVALUATION OF THE LES FUNDING PROGRAM

As introduced above, in 2017 the German Federal Ministry of Labour and Social Affairs announced a funding guideline entitled "Sustainable companies and administrations in digital change"[1], for which consortia of small and medium-sized companies and application-oriented research institutions, including universities, applied for project funding. A central concern of the funding is, to implement the connection between technological and economic change processes and social innovation within the framework of strong employee and company participation. The institutional framework for this is provided by the BMAS's *New Quality of Work Initiative* (INQA), which is based on social partnership and designed to promote innovative solutions. The aim was to promote innovative, tailored and consensual solutions for employees and companies in the digital transformation, supported by the social partners, e. g. work councils or trade unions. Company learning and experimental spaces should be fundamentally, exemplary in character and contribute to further innovative solutions. The objects of funding were so-called *Learning and experimental spaces (LES)* – interventions for which failure was allowed, and measures could experimentally be tested out. The 17 plus – in a second AI-oriented funding round[2] – 11 projects were all evaluated individually by

1 See BAnz AT 14.08.2017 B2 at https://www.bundesanzeiger.de (last visited 23.05.2025).

2 See BAnz AT 20.01.2021 B2 at https://www.bundesanzeiger.de (last visited 23.05.2025).

separate independent institutions. Our team of the Federal University of Applied Administrative Sciences was commissioned to conduct an overall evaluation of the guideline.

The funding instrument of the learning and experimental space is considered particularly suitable for finding such strategies due to its open-ended approach, which enables operational learning and experimentation with new forms of work as iterative processes. An essential feature of this funding logic is that 'failure' is permitted, as setbacks are seen as part of the learning process. This funding logic is intended to free projects from the restrictive (implicit or explicit) expectation that they must present 'working' innovations or products at the end of the funding period. Project results can therefore also be learning outcomes that arise from trying out new technologies, methods or forms of work.

All funded project interventions aimed to promote digitalization in and medium-sized enterprises. For example, bus drivers in Leipzig were equipped with tablets to enable better communication with each other. In the care sector, speech recognition software was tested to facilitate documentation. In another project, exoskeletons were used by companies specializing in the renovation of bathrooms to support heavy work.

In order to better understand the overarching objectives of the funding guideline, a few more detailed insights into selected projects are provided here: One exemplary project from the 17 projects in the first funding round was entitled *AgilKom*. This project brings together stakeholders from the local administration (the city of Essen and the administrative district of Soest), the United Services Union (ver.di) and the German County Association. The specific intention of the *AgilKom* project was to implement innovative and agile processes in the administration on a technological and organizational level. And contrary to the usual top-down-logic, such solutions were developed in so called *Innovation-Labs*, in which employees work together across hierarchies and disciplines. The overarching goal in the end was to improve flexibility, efficiency as well as closeness to citizens of the administration. Regarding the impact level, the project aspires to gain insights into the transferability of central principles of agile organization to the public sector. Another project was *Handwerksgeselle 4.0*, which was about the development of technological assistance systems in the sanitary, heating and air conditioning industry. Accordingly, stakeholders from both, the development sector and the application sector were working together in this LES with the aim of solving industry-typical problems (e.g. shortage of recruits, demographic change and competitive pressure). An exemplary project from the 11 projects in the second, AI-oriented funding round is the *KIDD* project. While the project primarily involved small and medium-sized enterprises, major corporations also contributed by collaborating on various topics (e.g. sales, evaluation of services, personnel management). With a focus on the application of digital systems in the work context, the KIDD project aimed to promote diversity among employees. This is achieved by developing standardized processes and criteria for these systems – especially in the application of AI.

As overall evaluators, our mandate included, on the one hand, compiling a synopsis of the findings from the individual evaluations across the 17 and 11 projects. During the clarification of the mandate, however, we also realized that the ministry wanted to know how well the new funding logic and the LES as funding instrument were working. This led us to expand the mandate to include this point on the other hand.

In concrete terms, the daily work of the overall evaluation consisted largely of networking and maintaining contact with the individual project evaluations. Data and findings from the individual experimental spaces, usually gathered at the beginning, throughout and at the end of the funding period, were incorporated into a survey instrument, developed specifically for the overall evaluation.

We drew on the approach of developmental evaluation (Patton, 2010), without claiming fidelity to its pure form or essential principles (Patton, 2016). Developmental Evaluation focuses on providing real-time feedback to support decision-making in complex, evolving environments. It prioritizes learning over accountability, helping stakeholders adapt strategies as programs unfold. Developmental Evaluation is highly collaborative, with evaluators working closely with program staff and stakeholders to co-create solutions. Unlike traditional evaluations this approach is flexible and emergent, adapting its methods and goals as the program evolves. The approach is designed for innovative or change-driven initiatives, emphasizing systems thinking and the continuous refinement of processes. We see a connection between the learning and experimental spaces and Patton´s (2015) principles of utilization-focus, developmental evaluation and co-creation principles that both, the project evaluators and our team aimed to uphold. This involved flexibility of the methods as well as dynamic designs (Patton, 2015) that adapted to the changing timelines and contents of the different projects. The reference to Patton became apparent during the evaluation process. When planning an overall evaluation, the benefits were not yet clear to us, which illustrates our adaptive approach.

Thus, based on the insights gained during the evaluation process, we applied an adaptive multi-method approach (see Figure 1). Besides a document analysis (see first column), the evaluation included several components. First, on the qualitative side (second column), we conducted guided, individual face-to-face interviews with representatives from all stakeholder groups, involved in the funding instrument LES. We also held focus groups with volunteer evaluators and responsible representatives from the ministry. In addition, we asked for written reflections from the project managers on the course of the project and on the funding instrument. Furthermore, in the third column: (longitudinal synopsis), we used an instrument to collect what we called descriptive fields. These were based on a framework and included recommended indicator categories such as learning and application, improved working conditions, sustainability and future viability, communication, participation and organization. We also conducted a standardized online survey with project managers (see last column).

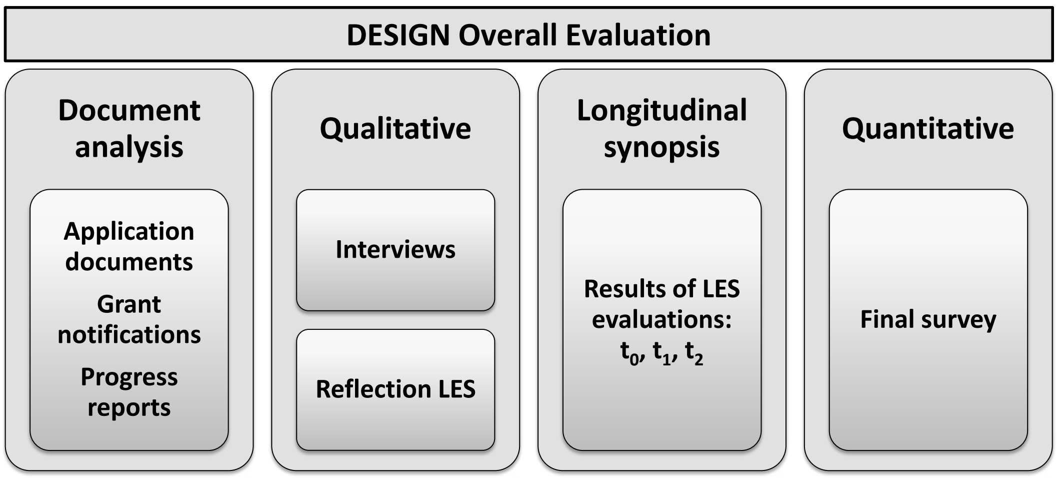


Figure 1: Design Overall Evaluation

3. REFLECTION ON THE EVALUATION PROCESS

In the following, the evaluation process is examined from the five perspectives: the financing logic (asking: How well did the linear path function?), the context in which the measures of the individual projects were implemented, the specific interventions, the methods applied and the role of the evaluators.

FINANCING LOGIC

The idea is that experimental spaces, in contrast to linear project funding, should make it possible to make mistakes and to allow failure during the research project. We found this idea has been well received by the project participants, as it was mentioned in both interviews and focus groups as a special feature of the funding instrument. However, the more interesting question is, to what extent these possibilities have been implemented during the project. We found indications that there is still room for improvement in this area. On the one hand, only a few projects documented processes of failure, on the other hand, there were also only few instances where projects deviated from their original goals or budget planning and made corresponding adjustments. One possible reason for this might be that the additional administrative effort involved in the projects (iterative cycles) was not considered. The challenge of evaluating non-linear funding logics, aimed to addressing wicked problems is described in the literature as lying primarily in several areas of tension: the attribution of outcomes to specific interventions, limited funding periods versus impact measurement, processes versus outcomes and reflexive learning versus external control (Rittel & Webber, 1973; Rohracher, Coenen and Kordas, 2022). In our overall evaluation, two-thirds of the project managers, surveyed in the final round, judged the three-year funding period as too short to effectively measure impact. This despite the fact that the funding guideline, as the name suggests, explicitly focuses learning. From an academic point of view, as represented by the technical support provided by the departmental research institution BAuA (Federal Institute for Occupational Safety and Health), the question of attributing intervention success and identifying causal effects remains largely unresolved. Instead, it is primarily anecdotal knowledge that has been generated, which could potentially be applied successfully to other areas. Short-term solutions to long-term problems cannot be implemented and therefore not measured. The consequence for the evaluation process is that it adopts a process-oriented, formative view instead of an ex-post, summative one. From an epistemological term, it shifts away from a predominantly positivist approach toward a more interpretive and constructivist one.

CONTEXT OF INTERVENTION

Solving complex problems with the help of project funding, requires accurate understanding of wicked problems. This includes an understanding of the possibilities of funding logics and the acceptance that compromises, made in the attribution of success and in the transfer of results. The 28 LES projects shared the characteristic of operating in highly dynamic and complex environments, especially when viewed in the context of the funding guideline. As mentioned above, the heterogeneity of the projects made it difficult to summarize their results, an issue stemming from the diversity of funding guidelines and subsequent results. On the other hand, this heterogeneity enables transferability to other domains at a sufficiently high level of abstraction, which is reflected in some of our recommendations. For example, in several projects we were able to show that expectations of the intervention were often too high. Accordingly, managing expectations realistically helps to more accurately reflect the potential impact of the measures. This could be achieved by specifying the technological focus in the funding guideline from the outset. Another important factor for all LES projects conducted over the past four years is COVID-19. The pandemic has led to a new way of dealing with uncertainty and unforeseeable situations. In response to this, we mainly received the message that there was a kind of *COVID-Boost,* in terms of the acceptance of digitalization measures. In summary, it is not only difficult to summarize the results of heterogeneous projects, but also to compare them meaningfully. This problem is exacerbated when these projects are implemented in complex and uncertain environments.

SPECIFIC INTERVENTIONS

The solutions within the individual projects were co-developed with the social partners and had not been predetermined at the project's outset. However, it remains unclear to what extent these solutions were further developed, or even discarded, over the course of the project. After all, this rather gradual development of interventions and solutions would have been entirely in line with the concept of the funding guideline. We gained the impression that the available opportunities for experimentation had not been fully utilized. However, it does not matter that the opportunities, created by the new intervention logic, will be utilized. Even if there is a will to make funding options more flexible to meet the need of transformational IPs, there must also be the courage to implement them (Rohracher, Coenen and Kordas, 2022).

APPLIED METHODS

From a methodological perspective, a challenge arose in conducting the overall evaluation regarding the fields of indicators recommended by the technical support from BAuA. These indicators were also recommended to the individual projects at the start of the funding period, without our ability to influence which ones would be used. In other words, there was no fixed set of indicators available at the outset that we could offer to all individual projects for reporting purposes. Rather, it was clear to us from the outset that we would develop such an instrument together with the evaluators of the individual projects. In our view, reflection on applicable indicators is generally linked not only to the overarching funding objectives, but also to how these can be addressed in cooperation with the individual projects. However, bundling the partially qualitative information proved difficult. The needed effort increased further over the course of the funding period, especially with regard to the AI funding round. In conceptual terms, we were able to conduct a formative evaluation because we were part of the funding guideline from the beginning of the projects. The synoptic presentation of the evaluation results over three measurement points was developed using a coding system for the data from the first measurement (t_0). Three coders employed a bottom-up approach to develop and compare responses from three randomly selected documents. For questions with largely identical content across later survey dates, supplementary codes were assigned where necessary and subsequently discussed to reach a consensus. The resulting coding system served the basis for the content and synoptic analysis of the three survey dates. The extent to which our overall evaluation and the synopsis of individual results have contributed to addressing the major problems must be assessed by others. In any case, we were able to obtain valid results to improve a more recent call for proposals utilizing the instrument of *Learning and experimental spaces*.

ROLE OF THE EVALUATORS

A high degree of flexibility in applying and adapting the survey instruments also demands considerable flexibility in determining the required resources. Our understanding of our role as evaluators – and the BAuA similar understanding – was characterized by adherence to scientific standards, both in our methods and in the assessment of the individual projects. In the developmental evaluation approach, the evaluator is seen as part of the innovation team. This was the case neither in the individual projects nor in our overall evaluation. In the individual projects instead, the role of individual evaluators was sometimes described as *critical friend* (Balthasar, 2012). The Ministry was ultimately responsible for the success of the funding guideline. However, our results also show that the project managers strongly identified with the interventions. On a subjective level, this represents an important additional factor for success. If the role of an evaluator is understood as a *critical friend* or even as *part of the innovation team*, like the developmental evaluation suggests, it can be stated that the role of the administrative actors is changing. In this way, the evaluators take on more of an advisory role and less of a hierarchical, top-down position.

4. CONCLUSION

Learning and experimental spaces are currently a widely used project funding instrument at municipal, state and federal level in Germany. Their evaluation requires a new perspective on the funding process, the role of the stakeholders and applied methods. Finally, based on our experiences, we would like to offer some conclusions for future overall evaluations. In any case, the primary challenge is to transition from anecdotal knowledge in very specific areas to identifying effective principles and patterns. A well-implemented intervention should only be extended to other environments if it has proven effective in a setting with a control group, something that is difficult to achieve in the real-world funding context. Focusing too much on transfer from the outset can also dilute the impact assessment. Furthermore, the *hunger for learning* – as opposed to the *fear of failure* – should be communicated and practiced from the outset. Although flexibility is required from both, the innovation team and its evaluators, evaluations of a certain scope cannot be conducted without some degree of top-down-control. It is not possible to collect valid and summable data in such a heterogeneous, agile and large-scale environment. The demands on evaluators' qualifications are once again increasing, in addition to qualitative and quantitative methodological knowledge, creativity and tolerance for ambiguity (skills we also put to test during in the pandemic) and strong social competencies are becoming increasingly important.

REFERENCES

Balthasar, A. (2012). Fremd- und Selbstevaluation kombinieren: der ‚Critical Friend Approach' als Option. In: Zeitschrift für Evaluation, 11(2), pp. 173-198.

Patton, M. Q. (2010). Developmental evaluation: Applying complexity concepts to enhance innovation and use. Guilford Press.

Patton, M. Q. (2015). State of the art and practice of developmental evaluation: Answers to common and recurring questions. In: Patton, M. Q., McKegg, K., & Wehipeihana, N. (Edts). Developmental evaluation exemplars: Principles in practice (Reprint). Guilford Press, pp. 1-24.

Patton, M. Q. (2016). What is essential in developmental evaluation? On integrity, fidelity, adultery, abstinence, impotence, long-term commitment, integrity, and sensitivity in implementing evaluation models. In: American Journal of Evaluation, 37(2), pp. 250-265.

Reale, F. (2021). Mission-oriented innovation policy and the challenge of urgency: lessons from Covid-19 and beyond. Technovation, 107, 102306.

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy sciences, 4(2), pp. 155-169.

Rohracher, H., Coenen, L., & Kordas, O. (2023). Mission incomplete: Layered practices of monitoring and evaluation in Swedish transformative innovation policy. In: Science and Public Policy, 50(2), pp. 336-349.

Schot, J., & Steinmueller, W. E. (2018). Three frames for innovation policy: R&D, systems of innovation and transformative change. In: Research policy, 47(9), pp. 1554-1567.

AUTHORS

CHRISTINA SCHUH
Federal University of Applied Administrative Sciences
Willy-Brandt-Straße 1
50321 Brühl, Germany
Email: christina.schuh@hsbund.de

DANIEL SCHWERTFEGER
Federal University of Applied Administrative Sciences
Willy-Brandt-Straße 1
50321 Brühl, Germany
Email: gesamtevaluation-experimentierraeume@hsbund.de

SONJA FRINGES
Federal University of Applied Administrative Sciences
Willy-Brandt-Straße 1
50321 Brühl, Germany
Email: gesamtevaluation-experimentierraeume@hsbund.de

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e12, 1-16
DOI: 10.22163/fteval.2025.704

UNVEILING INNOVATION: USING INNOVATION BIOGRAPHIES IN EVALUATION PRACTICE - A REFLECTION

KATHLEEN TOEPEL
DOI: 10.22163/FTEVAL.2025.704

ABSTRACT

A wide range of RTI funding measures are implemented at regional level in European Structural and Investment Funds programmes. During the 2014-2020 funding period, the Berlin European Regional Development Fund (ERDF) programme included direct funding for R&D projects by companies conducted in collaboration with research institutions, support for innovative start-ups through venture capital funds as well as the establishment of application labs and validation centres at research institutes and universities. A multi-year, accompanying evaluation (2016-2024) was conducted on behalf of the Berlin Managing Authority for the ERDF programme.

In this evaluation, the use of innovation biographies represented a novel approach. This new research approach allowed the reproduction and analysis of the entire process of knowledge generation and application in specific innovation processes of companies, universities, and non-university research institutions. A total of 23 innovation biographies were created and analysed in a joint manner in accordance with the impact pathways of the theory of change.

From the perspective of a policy maker, the results of this analysis demonstrate the interplay between various funding measures and the influence of external factors on innovation processes. The innovation biographies present an authentic portrayal of the extended periods of time required for the outcomes of funding to be realised.

The contribution presents insights gained through the utilisation of this qualitative approach in evaluation with illustrative examples. This method is particularly suited to map the complexity of innovation processes at the micro level. Consequently, the paper contributes to the advancement of R&I policy evaluation frameworks and methods.

Keywords: innovation biographies, ERDF, qualitative evaluation method, innovation process, accompanying evaluation

ZUSAMMENFASSUNG:

In Programmen der Europäischen Strukturfonds werden vielfältige FTI-Fördermaßnahmen umgesetzt. Das Berliner EFRE-Programm enthielt in der Förderperiode 2014-2020 u.a. die direkte Förderung von FuE-Vorhaben von Unternehmen im Verbund mit Forschungseinrichtungen, die Unterstützung von innovativen Gründungen durch Risikokapitalfonds und den Aufbau von Applikationslaboren und Anwendungszentren an Forschungseinrichtungen und Hochschulen. Für das EFRE-Programm wurde (2016-2024) eine mehrjährige begleitende Evaluierung im Auftrag der Berliner EFRE-Verwaltungsbehörde durchgeführt.

Zur Evaluierung der Prioritätsachse 1 „Innovation" des EFRE-Programms in Berlin wurden erstmalig Innovationsbiografien als zentrale Evaluationsmethode angewendet, mit dem der gesamte Prozess der Wissensgenerierung in konkreten Innovationsverläufen der untersuchten Unternehmen, Hochschulen und außeruniversitären Forschungseinrichtungen nachgebildet werden konnte. Erstellt wurden insgesamt 23 Innovationsbiografien, die für die Evaluierung entlang des Wirkungsmodells ausgewertet wurden.

Für die Stakeholder aus Politik und Verwaltung wird anhand der Innovationsbiografien insbesondere das Zusammenspiel verschiedener Fördermaßnahmen und der Einfluss externer Faktoren auf die Innovationsprozesse deutlich. Die Innovationsbiografien zeigen den Zeitaufwand, der bis zur Realisierung von Forschungsergebnissen benötigt wird.

Der Beitrag stellt die Erfahrungen mit dieser qualitativ ausgerichteten Methode anhand von Beispielen vor. Mit dieser Methode lässt sich insbesondere die Komplexität von Innovationsprozessen auf der Mikroebene abbilden. Dadurch wird ein Beitrag zur Weiterentwicklung der Evaluationsmethodik geleistet.

Schlagworte: Innovationsbiografien, EFRE, qualitative Evaluationsmethode, Innovationsprozess, begleitende Evaluierung

1. INTRODUCTION

A wide range of RTI funding measures are implemented at regional level in European Structural and Investment Funds programmes. For the 2014-2020 funding period, Berlin's ERDF (European Regional Development Fund) programme included among other actions direct funding for R&D projects by companies conducted in collaboration with research institutes, support for innovative start-ups through venture capital funds and the establishment of application labs and validation centres at research institutions and universities. The specific objective of this programme is to intensify and expand the innovation activities of the business sector. The entire innovation process is to be strengthened.

A total of almost 600 million euros in eligible expenditure was available for this purpose, this is 45% of ERDF funding for the entire funding period in Berlin, thereby making this axis the largest component of the Berlin programme. The actions under this priority axis covered all phases of the innovation process (applied industrial research, experimental development and production set-up / market launch) as well as finding co-operation partners through network funding (in the clusters of the Regional Innovation Strategy and within the cultural industries).

A multi-year, accompanying evaluation for the ERDF programme was conducted on behalf of the Berlin Managing Authority in the Senate Department for Economic Affairs, Energy and Public Enterprises. In accordance with Article 54 of Regulation (EU) No. 1303/2013 and Berlin's evaluation plan, the effectiveness of the funding, its efficiency and its impact of each priority axis were assessed. The accompanying evaluation of the 'Innovation' priority axis started in June 2016. An interim report was produced in 2018 (IfS 2018). The final report was published in 2023 (IfS 2022).

2. METHODOLOGY

The study design for the evaluation of the priority axis was programme theory based (Funnel and Rogers 2011, Rogers 2014). The aim of the study was to reconstruct impact pathways and show how and under what circumstances the interventions work or do not work. In accordance with the evaluation plan, the study design was based on the understanding of the theory of change of ERDF funding as described in the 'Guidance Document on Monitoring and Evaluation', according to which external factors in addition to the intervention also have an impact on the results (European Commission 2014: 6). In this priority axis, the interplay between the various measures and external factors was examined during the programme period. The design thus corresponds to the complex objective of the priority axis ('strengthening the entire innovation process') and the diverse measures.

As a novelty in evaluations, innovation biographies (adapted from Butzin et al. 2012) were chosen as the central method. The final report is largely based on the cross-evaluation of the innovation biographies. It was accompanied by an analysis of monitoring data provided by the programme owners.

Innovation biographies as such are not new, but their use in evaluation is. The main methodological source was the work of the Institute for Labour and Technology Gelsenkirchen (Helmstädter / Widmaier 2001), which compiled the first innovation biographies around 2000. As part of the European project EURODITE - Regional trajectories to the knowledge economy[1] (6th EU Framework Programme, 2005-2010), 60 innovation processes were analysed. In addition, there were further applications at national level, e.g. in the construction industry, nanotechnologies and renewable energies in Germany.

Innovation biographies are a research approach that can be used to empirically capture knowledge dynamics in innovation processes from a spatial and sectoral perspective. Innovation biographies make it possible to model the process of knowledge generation in concrete innovation processes, from the initial idea to the concrete form of a new product or service, production set-up and market launch. This approach considers changes in the theoretical and empirical debate on innovation (Rammert 2000, Crevoisier/ Jeannerat 2009). At the same time, an exploratory approach enables the mapping of the influence of external factors in the same manner as that of the various support measures, thus facilitating a more comprehensive understanding of the mode of action.

The evaluation design, in the form of such innovation biographies, allows for a thorough examination of the individual case. This is particularly evident when considering the specifics of the project, historical coincidences, and external

1 https://www.uni-marburg.de/de/fb19/forschung/forschungsprojekte/geographie-der-dienstleistungen-kommunikation-und-innovation/docs/eurodite_abstract_engl_long.pdf

influences. The development history of the projects in their interaction with the environment in which they are operating can thus be analysed.

A ‘panel’ of 23 innovation projects was utilized to map the respective innovation development. The cases were selected from different measures (R&D projects of companies, venture capital investments and application labs) in three consecutive funding years (2015-2017). Based on monitoring data, the selection of cases was informed by an appropriate mix of sectors, fields of technology and company sizes, as well as types of organisations (enterprises, universities, non-university research institutions). Investment stages (seed, startup, growth) as well as cooperation patterns (individual or collaborative project) have also been considered. This resulted in an equal representation of the project types and corresponded to the approval patterns of the individual actions.

The accompanying evaluation of the ERDF allowed for a long-term study design over the entire funding period. The observation period ran from July 2016 to May 2021. Each selected project was interviewed once a year. The observation period for each individual case is at least four and up to six years in duration. In some cases, the history of each innovation is documented over a period of up to 10 years. This allowed a long-term perspective beyond the (limited) duration of the funded project. In addition to the annual interviews with the management of the company or project leaders, exploratory interviews were conducted with key cooperation partners and investors.

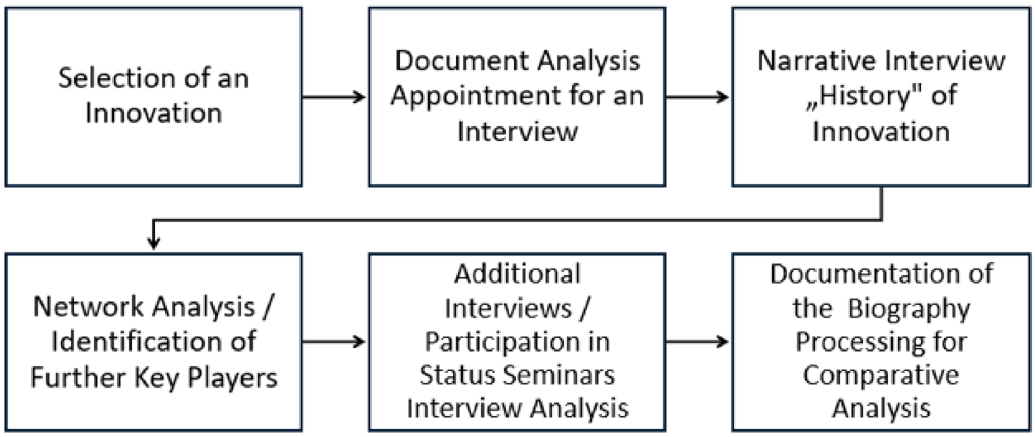


Figure 1: Approach in the Innovation Biographies
Source: Own elaboration based on IfS 2022

An innovation biography was created for each case study, in which all significant aspects and factors influencing the development of the innovation were presented and integrated into their context. This included an explanation of the impetus that led to the initial idea, the obstacles and difficulties encountered, a trajectory of knowledge development and the associated network of stakeholders. Furthermore, the chronological sequence of predecessor and successor projects was also considered. The innovation biographies were updated annually. Finally, the analysis of the innovation biographies was conducted in a joint manner, in accordance with the impact pathways of the theory of change.

The basis for the synthesis was laid by the interviews in the form of transcripts or notes plus internet research and documents. The company or research organisation received its own updated innovation biography for information purposes. In a first section, it contains a description of the innovation that largely abstracts from technical details. The core is the understanding of its Unique Selling Proposition, which problem it solves or need it fulfils, its price model, target customers, etc. The second section describes the chronological processes and events that influence or relate to the innovation. Additionally, these have been visualised in a detailed timeline. The documents of the innovation biographies comprised up to 30 pages.

In an additional working paper per case, the evaluation team reflected on the case based on the impact pathways. This internal working paper documented the activities of the evaluation team, recorded additional information as well as special features and open questions. At the same time, the contents of the innovation biographies were prepared for the cross-evaluation (e.g. success factors and obstacles) of the cases. The content of the interviews was immediately processed in the form of these two documents (innovation biography and working paper). Both together provided the basis for the cross-evaluation and synthesis. In the synthesis, the findings from the individual innovation biographies were collated using selected questions and parameters based on the impact pathways and assumptions.

3. EXAMPLES OF INNOVATION BIOGRAPHIES

Two examples from different funding measures are presented here to illustrate the method. It is not possible within the scope of this article to present all the details of each innovation biography, but these case studies show that the funded project is part of a long-term chain of activities and projects. It is preceded by other funded or non-funded projects, or is followed by other projects and other activities, each with their own objectives. A variety of factors influence the success or failure of innovations.

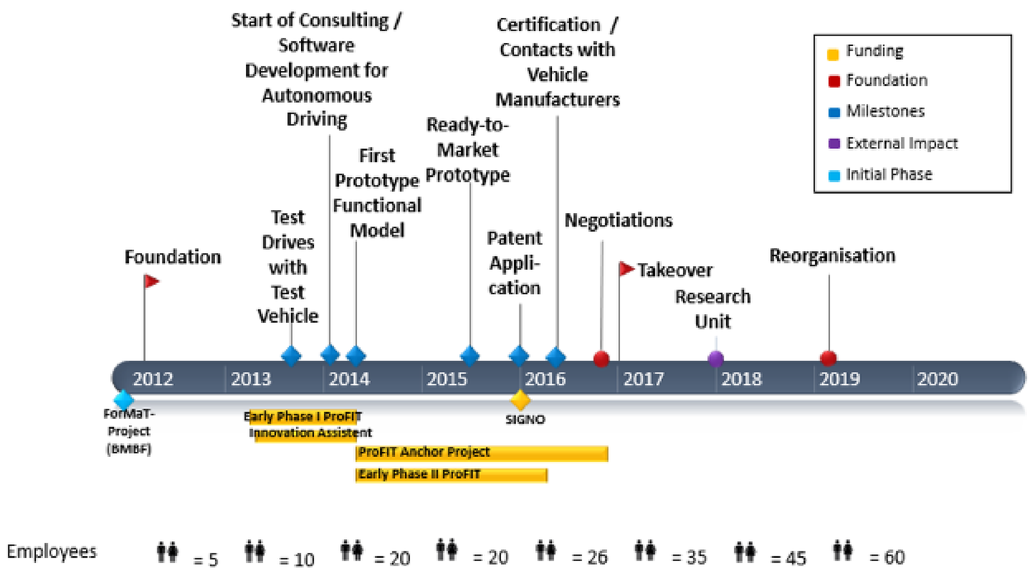


Figure 2: Example 1 - Reversing Assistance System for Refuse Collection Vehicles
Source: IfS 2022

Refuse collection vehicles have frequent accidents when reversing. The idea of this R&D project was to develop a camera-based system with image recognition instead of sensors to avoid collisions.

The timeline clearly shows that various support measures are interlinked here over time. The initial idea was developed in a ForMat project[2] funded by the Federal Ministry of Education and Research (BMBF). Here, engineers and business economists at the TU Berlin developed use cases for image recognition components that emerged from research work on autonomous driving. The company was a university spin-off. The Berlin ProFIT funding in the early phase made it possible to establish and develop the company's organisational structures. An innovation assistant was hired to support innovation management. The only measure that was financed by the ERDF is the ProFIT anchor R&D project. In a later phase, the patent application was subsidised at federal level (SIGNO Programme).

This example also illustrates the influence of the regulatory framework. During the development of the product, a ban on reversing for refuse collection vehicles was discussed. This uncertainty about the chances of use led to months of delays. The certification requirements regarding the robustness

2 The ForMaT Programme (Forschung für den Markt im Team) placed a particular focus on designing research results in such a way that they have practical applications on the market and can be successfully placed there. The programme ran from 2005 to 2017.

of the product for heavy goods vehicles also made the camera system considerably more expensive.

Alongside this product development, the company developed software related to image recognition for autonomous driving. At the end of the project a demonstrator of the reversing assistance system was available. But the project ended early due to a company takeover. There was no market launch of the assistance system, but R&D jobs in new enterprise were retained in Berlin. Regarding the achievement of the objectives of the ERDF programme it can be said in this case, that there is no new product in the market, but local R&D capacities have been strengthened.

This innovation biography contributed to the following selected findings:

- The subsidised project is only a small part of a longer innovation process (32 months preliminary phase from the first innovation idea and 33 months project duration).
- Product development would not have been able to take place without the simultaneous funding of the company set-up.
- Innovation processes can be stopped by taking over the companies, but the acquired knowledge lives on and is utilised elsewhere.
- What matters is not the specific new product on the market, but that knowledge-intensive R&D jobs are retained in the region.
- Pending regulatory procedures and high regulatory requirements slow down and jeopardise product development and make the product more expensive.

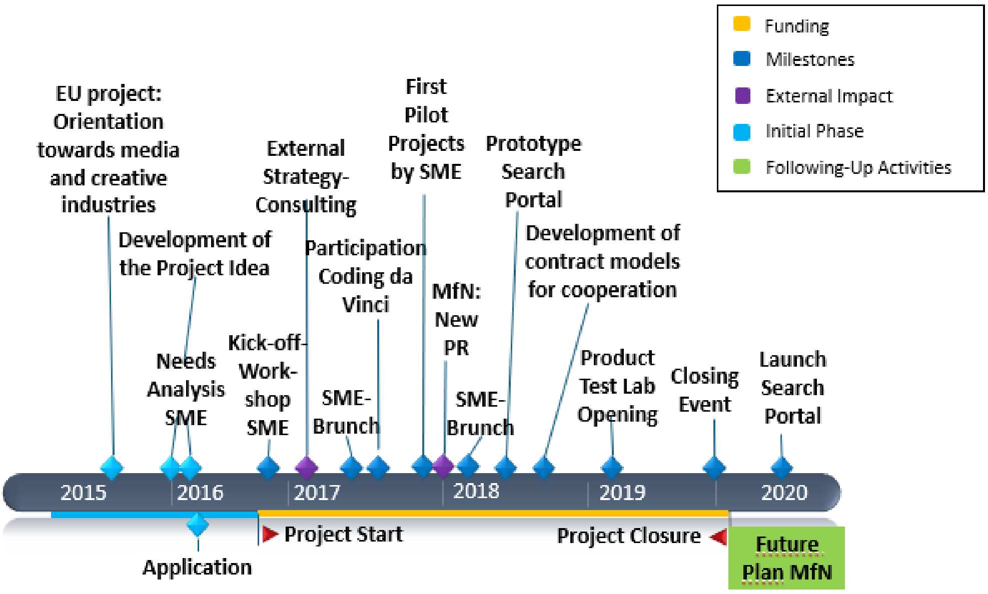


Figure 3: Example 2 - Mediasphere for Nature at Natural History Museum
Source: IfS 2022

The Mediasphere for Nature is a multimedia application laboratory at the Natural History Museum in Berlin. It is a research museum, so it is part of the Leibniz Institute for Evolution and Biodiversity Research. The objective was to make the rich collections (over 30 million collection items) available to users (especially SMEs) from the cultural and creative industries, whereby neither a strategy nor contacts of the museum to this sector existed beforehand. The idea of opening the collections to the creative industries came about as part of an earlier EU project[3].

Numerous networking activities were carried out, contacts have been established with interested SMEs and collaborations (28 cooperation agreements) have been realised. Demand among SMEs was high and has risen steadily. The museum developed internal structures and expertise for cooperation with such companies (e.g. granting of rights, contract models, work processes). A wide range of applications (VR, AR, videos, games, educational materials, touchable exhibition objects for the blind) were developed by the SMEs. They used animal voice recordings, documentation of research trips or expeditions, plant textures, digital animal models, specific research results like: “How does a frog perceive its surroundings?”

3 The Europeana Creative project (2013-2015) enabled and promoted greater re-use of cultural heritage resources by creative industries. The Museum was involved in a pilot project in the field of natural history. See https://pro.europeana.eu/project/europeana-creative-project

(for an interactive VR experience[4]). It was also possible to test products, apps or games with visitors to the museum.

After the end of the funded project, cooperations are being continued in new projects with SMEs. As part of the Future Plan for the Museum, considerable funds have been made available for the further digitisation of the collections. The Mediasphere will be a sub-project for improved access, innovation and networking. The media repository and research portal, the experimental field and the SME network will be continued within this framework.

This innovation biography contributed to the following learnings:

- For a research institute to cooperate successfully with SMEs, certain requirements must be met. Internal organisational structures, work processes and knowledge of the needs and requirements of SMEs were lacking. A transfer strategy was necessary to reach the target group.
- In communication between scientists and SMEs, there are often major differences in expectations and time management that need to be overcome.
- Delays in implementation resulted primarily from difficulties in precisely defining the required task profiles for newly created positions and finding the appropriate personnel.
- The lessons learnt from the project are valuable for other research museums at an international level.

Both innovation biographies illustrate the range of possible findings, only a fraction of which could be presented here. Finally, a total of 23 innovation biographies were available for the summarised analysis and synthesis. The table shows the parameters conceptualising the results in the two reports. The focus shifts from the first phases of the innovation processes in the interim report to the results and impacts in the final report.

4 https://inside-tumucumaque.com

Table 1: Main parameters of the synthesis

	Parameters conceptualising the synthesis results
Interim Report 2018 (IfS 2018)	▪ Role of funding in the innovation process: expansion of R&D capacities in the company, effects in addition to cost reduction and risk mitigation ▪ Expansion of R&D capacities in application laboratories ▪ Access to financing for innovative start-ups ▪ Innovative and creative impulses through co-operation ▪ Development of the initial innovation idea: problem seekers, problem solvers and customer understanders ▪ Other factors influencing innovation: economic situation, choice of location in innovative hotspots, shortage of skilled labour, technical and commercial expertise in the founders' team, legal framework conditions ▪ Interplay of funding measures in the innovation process ▪ Experiences with the application process and implementation of funding
Final Report 2022 (IfS 2022)	▪ Expansion of R&D capacities in companies: Implementation of R&D projects and direct job effects, development of R&D intensity ▪ Market launch of new products and services, status and duration of the development process from idea to market launch ▪ Effects on R&D behaviour, innovative capacity and technological skills ▪ Effects on cooperation behaviour in R&D and in other networks and clusters ▪ Strengthening the transfer activities of research institutions and universities: utilisation of funded infrastructures, strengthening of application orientation, cooperation with SMEs ▪ Growth of companies, creation and safeguarding of jobs, stability and future of companies ▪ Effects of the COVID-19 pandemic

4. EXPERIENCE, BENEFITS AND APPLICABILITY AS AN EVALUATION METHOD

The consolidation and synthesis of the innovation biographies for the overall evaluation reports was challenging and time-consuming because a lot of material (innovation biography, working document, timeline) was available.

The main risks in applying this method were whether the companies would participate for so long and allow these insights into innovation activities. Consistent cooperation and trust were built up during the interviews. This worked well in Berlin but required staff continuity in the evaluation team throughout the entire observation process.

The advantages of the method lie in the longer-term perspective, which extends far beyond the funding period and after project end. The explorative approach can discover impact factors that were not anticipated by the evaluation team based on field knowledge and literature (e.g. impact on innovation management). The process view shows all interactions and loops during the innovation process. It allows for a better understanding of mechanisms and time sequences. The method made it possible to map the influence of both, external factors and ERDF funding alike. In contrast to conventional case studies, the focus of innovation biographies is on the innovation, while other case studies often only focus on the funded project and remain within the logic of a single programme.

The method is based on storytelling.[5] The evaluation team comes as listeners. Storytelling is a natural way to share information and experience. Lively stories beyond bare facts and figures have been received. The depth of stories allows to show complexity in the innovation process as well as in its context. It is not always possible to assume that practitioners from the administration (as clients of the evaluation) or political actors and other stakeholders (as addressees of the evaluation results) have knowledge of operational practice in companies or research institutions. Here, the method proved to be particularly fruitful in achieving a greater and more realistic common understanding of how the funding works for the beneficiaries.

5 Several approaches use storytelling to evaluate impact, especially the Most Significant Change, see Snow et al. (2021).

It was also possible to show which time periods are realistically required until an outcome (e.g. the utilisation of R&D results on the market, job creation) becomes apparent. A lot can happen during this period that influences the effects of the funding. For example, the innovation biographies included reorganisations and strategic takeovers in companies, insolvencies, staff changes and far-reaching financial and investment decisions. The organisations were affected differently by the impact of the coronavirus crisis and their handling of the pandemic restrictions varied considerably. The method underscores the importance of evaluation principles to listen carefully and consider context.[6]

In addition, the method proved to be a very powerful tool for communicating and reflecting on the evaluation results. Compared to other evaluations without the use of innovation biographies, it led to particularly lively discussions with stakeholders about the cases, the evaluation results based on them, the support measures as such, and the conclusions. Here, too, a story is told, which makes the complexity of innovations comprehensible for the policymaker (away from the abstract 'number' towards a real understanding of an innovation process). In addition to the linear time axis, the non-linear influencing factors are also clarified. From this and from the large number of innovation biographies, the transferability of the results arises, which in turn leads to further development of the funding, as innovation processes are better understood.

The accompanying evaluation setting in Berlin was certainly unique because most common evaluation studies cover a shorter time span and therefore have fewer observation points. The timing of the interviews is variable but should cover 3-4 points in time (at least project start, project end and later with a greater interval). It is also possible to apply the method in full retrospect. The disadvantage of the method is the relatively high cost for an enormous depth in a few cases. The costs can be influenced by the frequency of the interviews.

Innovation biographies can also contain a spatial component, which regions contribute to the generation and dissemination of knowledge. This was not so interesting in a city like Berlin but could play a greater role in larger regions.

The method is thematically open and broadly applicable. This evaluation study focussed on innovation ideas from many fields and sectors as well as key business ideas from start-ups. However, the method has also been

6 This refers especially to Standard G2 Accuracy: Context analysis: "The context of the object of Evaluation should be analyzed comprehensively and in sufficient detail and taken into account in the interpretation of results." DeGEval (2017), p. 44.

applied to the core idea of application labs at research institutions. Innovation biographies of social innovations are also conceivable. Therefore, innovation biographies could be a good qualitative supplement in larger, accompanying evaluations.

CONCLUSIONS

The contribution presents experiences with innovation biographies as a qualitative evaluation method. It clearly shows the interplay of different funding measures (including ERDF and other national and European funds) in the innovation process and the influence of political and regulatory frameworks and other external factors. The narrative style used in the innovation biography makes the impact mechanism of the funding visible, thereby facilitating the communication of evaluation results to stakeholders and a wider audience. The innovation biographies also provide an authentic account of the extended time periods required to realise the outcomes of the funding. In certain cases, this spans across several funding periods.

The benefits of this approach and its applicability to the evaluation of RTI measures have been discussed. The risks involved in implementing and analysing innovation biographies (especially the creation of trust) proved to be manageable. In future, the analysis of the extensive material generated by a larger number of innovation biographies could be facilitated by using AI. As an evaluation method, innovation biographies can be used to capture the complexity of innovation and transformation processes at the micro level. Consequently, the method contributes to the advancement of R&I policy evaluation frameworks and methods.

REFERENCES

Butzin, A. / Rehfeld, D. / Widmaier, B. (Hrsg.) (2012): Innovationsbiographien – Räumliche und sektorale Dynamik.

Crevoisier, O. / Jeannerat, H. (2009): Territorial Knowledge Dynamics. From the Proximity Paradigm to Multi-location Milieus. European Planning Studies. 17 (8), 2009, 1223-1241.

DeGEval – Gesellschaft für Evaluation e.V. (2017): Standards für Evaluation. Erste Revision 2016.

European Commission (2014): Guidance Document on Monitoring and Evaluation.

Funnell, Sue C., Rogers, Patricia J. (2011): Purposeful Program Theory: Effective Use of Theories of Change and Logic Models.

Helmstädter, E. / Widmaier, B. (2001): Wissensteilung und Innovation. Jahrbuch Institut Arbeit und Technik 2001, 115-128.

IfS (2018): EFRE-Förderung im Land Berlin 2014 bis 2020 - Evaluierung der Prioritätsachse 1 „Innovation". Zwischenbericht. https://www.berlin.de/sen/wirtschaft/gruenden-und-foerdern/europaeische-strukturfonds/efre-foerderperiode-2014-2020/der-efre/operationelle-programme/das-operationelle-programm-des-efre-2014-2020/zwischenbericht_evaluierung-pa-1.pdf

IfS (2022): EFRE-Förderung im Land Berlin 2014 bis 2020 - Evaluierung der Prioritätsachse 1 „Innovation". Endbericht. https://www.berlin.de/sen/wirtschaft/gruenden-und-foerdern/europaeische-strukturfonds/efre-foerderperiode-2014-2020/der-efre/operationelle-programme/das-operationelle-programm-des-efre-2014-2020/endberichtevaluierung_pa_1.pdf

Rammert, W. (2000): National Systems of Innovation, Idea Networks, and Comparative Innovation Biographies, TUTS-WP-5-2000. http://www.ts.tu-berlin.de/fileadmin/fg226/TUTS/TUTS_WP_5-2000.pdf

Rogers, P. (2014). Theory of change. Methodological briefs: impact evaluation, 2(16), 1-14.

Snow et al. (2021): Storytelling for Systems Change. Insights from the field. Centre for Public Impact, Dusseldorp Forum and Hands Up Mallee. https://www.centreforpublicimpact.org/partnering-for-learning/storytelling-for-systems-change

AUTHOR

KATHLEEN TOEPEL
IfS Institut für Stadtforschung und Strukturpolitik GmbH
Lützowstr. 93, 10785 Berlin, Germany
Email: toepel@ifsberlin.de
Orcid: 0009-0002-2322-6777

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e13, 1-18
DOI: 10.22163/fteval.2025.704

EXPLORING UNINTENDED CONSEQUENCES IN STI EVALUATIONS AND MONITORING

TOWARDS A FRAMEWORK FOR UNINTENDED CONSEQUENCES AND ITS USE IN EVALUATION AND MONITORING EXERCISES

SARAH SEUS, FLORIAN WITTMANN AND NELE WEIBLEN
DOI: 10.22163/FTEVAL.2025.704

ABTRACT

With the turn to mission orientation and transformative policies, STI funding organisations increasingly aim for impacts beyond the research sector. In consequence, in the STI evaluation community an intensive discussion has emerged on the nature of intended effects of research funding proposing new concepts and methods for modelling and measuring these intended effects. In contrast, the understanding of unintended effects – despite the recognition of their importance - has remained limited. This paper seeks to address this gap by exploring and consolidating the dispersed knowledge on unintended consequences by a multifaceted approach. We perform a systematic literature analysis focusing on the characteristics of unintended consequences and explore how unintended effects have been addressed in different evaluation guidelines and evaluation studies. Based on this review, we conclude that the diversity of unintended consequences is a major bottleneck to advance both theory and practice. To overcome this challenge, this paper develops a reflection tool consisting of guiding questions that can be used to navigate through the diversity of unintended consequences.

Keywords: unintended effects, uninteded consequences, SIPER

1. INTRODUCTION

When conducting (programme) evaluations the focus is first of all on the intended effects that should be achieved by a funding impulse. Especially with the turn to mission orientation and transformative policies (Mazzucato 2018; Diercks et al. 2019), STI funding organisations have aligned with policy demands and have designed programmes aiming at impacts far beyond the research sector (e.g. on European level the "Green Deal Calls" in H2020, in Germany the Strategy for Research for Sustainability (FONA); The Swiss National Research Programmes (NFP) or the Swedish Strategic Innovation Programme (SIP). In recent years, the STI evaluation community has responded to the challenges posed by these new funding programmes: an intensive discussion has emerged on the nature of intended effects of research funding, leading to the development of new concepts and methods for modeling and measuring these effects. (Bruno and Kadunc 2019; Bührer et al. 2022; Spaapen et al. 2011; Dinges et al. 2020; Seus and Bührer 2021).

A side effect of this shift towards transformative policy-making and the interest in longer term effects of research funding is the growing need to understand unintended consequences of funding activities. This is important for the following reasons: Aiming for a more systemic approach and more complex programmes, funding can potentially result in multiple interdependencies and cross-cutting effects, given its broader reach and the involvement of a wider variety of stakeholders. At the same time, against the background of increasing budget deficits and necessary cuts, the question of efficient use of resources gains additional momentum, creating additional pressure to avoid negative effects deviating from initial policy goals. Moreover, moving beyond the focus on policy outputs alone and emphasising the need for formative evaluation and learning (Molas-Gallart et al. 2021) increases the need to better understand the consequences of funding beyond programme priorities.

While there is an ongoing debate among scholars about unintended effects (Braun 2009; Jabeen 2016; Meijer and Sivertsen 2020; Morell 2011; Morell 2005; Turcotte-Tremblay et al. 2021; Zwart 2015), the debate appears rather inconclusive and is characterised by multiple unconnected strands of discussion. At the same time, it remains unclear how unintended consequences are addressed in evaluation practice. Therefore, our contribution seeks to explore and consolidate the dispersed knowledge of unintended consequences dispersed so far, by bringing together scholarly debates as well as STI evaluation practice. For this purpose, we draw on a systematic literature research approach and analysis of specialised evaluation databases (SIPER[1]) as well as evaluation guidelines (OECD, World Bank, Fteval) to analyze both a conceptual as well empirical perspective of unintended effects of STI activities.

Based on the analysis, we conclude that the high variety of unintended consequences is a major bottleneck to advance both theory and practice. In consequence, we argue in favour of a more pragmatic approach towards unintended consequences that focuses on specifying the focus of analysis. For this purpose, we provide some guiding questions that can serve as reflection tools for evaluators and actors in charge of monitoring. We end with a discussion on the future implications for evaluation and monitoring practice.

2. UNINTENDED CONSEQUENCES OR UNEXPECTED EFFECTS – WHAT TERMINOLOGY TO USE?

A first difficulty and source of confusion is the lack of clearly established terminology regarding the concept of unintended consequences. In the evaluation literature, largely related to the field of development studies, the terms 'unintended consequences' or 'unintended outcomes' are used most often (Oliver et al. 2019, p. 63; Bamberger et al. 2016; Jabeen 2016). These terms are commonly used to describe "effects […] not envisaged by the originator of the intervention or policy" (Oliver et al. 2019, p. 63). Concepts such as 'spill-over effects', 'externalities' or 'negative trade offs' (Oliver et al. 2019; Bonell et al. 2019; Jabeen 2016) are often used as synonyms in the same publication are.

This inconsistent use of the terminology is also reflected in the evaluation guidelines and evaluation studies: the terms 'unintended' and 'unexpected' are used interchangeably (e.g. in the fteval evaluation standards (Kohlweg 2019) with 'non-anticipated' and 'unanticipated' without further analytical differentiation (for a more detailed discussion see section 4).

In the following, we will use the term 'unintended consequences' as an umbrella term that encompasses both immediate unintended effects of the intervention

1 www.si-per.eu

and intervention-induced outcomes and impacts. Therefore, we assume that the term ‘consequences’ includes different categories, such as ‘results’, ‘effects’, ‘outputs’ and ‘outcomes’.

3. UNINTENDED CONSEQUENCES: MAIN DIMENSIONS BASED ON THE LITERATURE

To gain a better grasp of the fuzzy concept of unintended consequences, we conducted a literature review to explore the theoretical discourses on the topic and identify the existing definitions and dimensions of unintended consequences. Intentionally, we broadened our scope beyond STI evaluation literature to include general evaluation literature as well.

The inherent problems related to the concept of unintended consequences, especially the variety of terms, the diversity of (scientific) contexts and the lack of clear definitions have been discussed as early as 1936 (Merton 1936). Until today, different authors have put forward their definitions depending on the characteristics they see most important with regard to unintended consequences. In the following review, we strive to summarise the discussions on the nature of unintended consequences by proposing a classification of six dimensions and outlining their main characteristics:

A first dimension concerns the knowability of effects. Unintended consequences, according to this dimension, can be either anticipated, “known by the actor at the time of action” (Jabeen 2018, p. 264), or unanticipated. In the case of unanticipated consequences, a further distinction between foreseeable and unforeseeable outcomes is made. This division relates to the idea that effects can either be predicted or not (Jabeen 2018). Unforeseeable consequences occur where “adaptive and nonlinear phenomena make prognostication impossible” (Morell 2005, p. 445; Braun 2009). In contrast, foreseeable or predictable outcomes might still not be foreseen, leading to unforeseen effects, especially in those cases for which “applicable analytical frameworks and experience were not considered” (Morell 2005, p. 446). A last subcategory of effects in the knowability dimension, is that unintended effects can be overlooked, meaning they are “known but deliberately ignored for practical, political or ideological reasons” (Morell 2005, p. 445).

A second literature strand discusses unintended consequences with regard to the size of the unintended effects. Following Meijer and Sivertsen

(2020), the societal impact of a programme or research can be either normal or extraordinary. Normal societal impacts refers to “the results of active, productive, and responsible interactions between (units of) research organizations and other organizations according to their purposes and aims in society” (Meijer and Sivertsen 2020, p. 67). In contrast, extraordinary societal impacts are defined as “rare incidences where [...] interactions between science and society have unexpected widespread positive or negative implications for society” (Meijer and Sivertsen 2020, p. 67).

Thirdly, unintended consequences are differentiated according to the direction of the effect - based on whether the effect is evaluated as positive, negative or neutral (Bonell et al. 2019; Derrick et al. 2018; Jabeen 2016; Meijer and Sivertsen 2020). By combining the dimensions of size and value, Derrick et al. (2018) provide the example of ‘Grimpacts’, which are impacts of extraordinary size (Meijer and Sivertsen 2020) but with negative implications for society. Braun (2009), however, cautions against generally equating unintended effects with negative ones, as unintended outcomes could also be beneficial.

A fourth dimension concerns the controllability of the unintended effects. Controllability refers to the fact that even if unintended effects are anticipated, it might not be possible to avoid them (Braun 2009). Following Braun’s line of argument, the controllability of unintended effects depends on several different parameters. While simple effects, resulting from individual actions, are controllable, more complex or intricate effects are more challenging to control.

Moreover, another dimension is related to the stakeholders affected by unintended consequences (Jabeen 2018). In most cases, the beneficiaries of the interventions would be the ones affected. However, unintended effects can also occur among groups not directly targeted by the interventions or policies. These groups could include unsuccessful applicants, peers who did not apply for the funding, and organisations within the wider system, such as research organisations hosting grant holders.

Finally, the last dimension to consider when trying to locate unintended consequences is the timing of occurrence: Does the effect already occur during the implementation of the intervention or can it be expected only after the end of the intervention (Jabeen 2018)?

In the following table we summarised these different characteristics discussed above into six distinct dimensions. For each we describe the underlying concepts and refer to the literature in which these characteristics are discussed.

It should be noted that the last two dimensions i.e. "stakeholders affected" and "timing" are crosscutting to the first ones. The describe the possible locations of occurrence, whereas the remaining dimensions describe characteristics of unintended consequences.

Table 1: Six dimensions characterising unintended consequences

Dimension	Characteristics	Literature
Knowability	Anticipated vs. unanticipated Foreseeable vs. unforeseeable	Jabeen 2018 Braun 2009; Morell 2005
Size	Normal vs extraordinary	Meijer and Sivertsen 2020
Direction	Positive vs. negative vs. neutral	Jabeen 2016; Bonell et al. 2019; Derrick et al. 2018; Meijer and Sivertsen 2020; Braun 2009
Controllability	Controllable vs. uncontrollable	Braun 2009
Stakeholders affected	Beneficiaries / external stakeholder / funding organisations / all /	Jabeen 2018
Timing	During funding period / after project ends / ...	Jabeen 2018

4. UNINTENDED CONSEQUENCES IN EVALUATION STUDIES: AN EMPIRICAL REVIEW OF THE ACTUAL USE OF THE CONCEPT

As a starting point of our empirical analysis, we take the conclusion by (Bamberger et al. 2016) that most evaluation studies overlook unintended consequences. In a first step, we investigate how unintended consequences are discussed in evaluation guidelines. Focusing on a selection of main funding organisations provides a first hint at the use of unintended consequences in evaluation activities and their conceptualisation. In a second step, we conduct a systematic review of evaluation studies from the past 15 years. Drawing on the Science and Innovation Policy Evaluation Repository (SIPER) we employ both quantitative and qualitative text analyses to better understand the role of unintended consequences in evaluations and analyse the evaluations against our six dimensions discussed in chapt. 3.

4.1 UNINTENDED CONSEQUENCES IN EVALUATION GUIDELINES OF THE MAIN EVALUATION COMMISSIONING BODIES.

Because our particular interest is in the evaluation in the Science Technology and Innovation (STI) sector, we first looked at evaluation standards of the Austrian Platform for Research, Technology and Innovation Policy Evaluation (fteval) before expanding our screening to the guidelines and handbooks of evaluation societies and multilateral organisations, especially the OECD, World Bank and the UN.

The most relevant guidance is provided by the "Evaluation Standards for Research , Technology and Innovation Policy" (Kohlweg 2019) formulated by the Austrian Platform for Research and Technology Policy Evaluation - fteval. The concept of unintended consequences is mentioned twice in the glossary of the document. The search term "unintended" can first be found in the definition of the term 'Findings / Results': "Output, direct outcomes or longer-term effects of an intervention (intended or unintended, positive and/or negative)" (ibid p. 29). The second reference is related to explaining the term 'programme theory': "Impact modelling based on different presumed effects, showing how the activities and outputs of a programme contribute to the intended (and where relevant, also any unintended) outcomes and impacts." (ibid. p. 32).

Our screening of standards and principles of international evaluation societies, including the American, German and the European Evaluation Societies, yielded no results for search terms such as "unintended" or "unexpected".

Looking beyond STI policy, the guidelines and handbook of the OECD-DAC, the Independent Evaluation Group of the World Bank and the United Nations Evaluation Group (UNEG) mention the term "unintended" in glossary sections to characterise different forms of effects and highlight the positive and negative nature of effects (OECD 2021; OECD Publishing 2023; OECD 2019; Leeuw and Vaessen 2010; United Nations Evaluation Group 2016)

Especially in methodological guidelines, unintended consequences are discussed in a little bit more detail, e.g. to highlight the challenges related to capturing these unintended consequences and short discussions what data collection methodologies could be useful to detect and assess unintended consequences. (see in particular Leeuw and Vaessen 2010; Vaessen et al. 2020; UNEG Methods Use and Appropriateness Working Group 2022).

In sum, the standards, guidelines and handbooks of major evaluation commissioning bodies refer to the concept of unintended consequences as something to keep in mind as a supplementary effect category. Further explanations or guidelines on how to handle unintended consequences are missing.

4.2 UNINTENDED CONSEQUENCES IN EVALUATION STUDIES

The second focus of our empirical analysis was to understand how unintended consequences are analysed in evaluations. In order to get an overview of the use of "unintended consequences" in evaluations so far, the online database SIPER[2] was systematically searched. SIPER provides an online access to a large collection of policy evaluations to enable support for academics but also policy learning for practitioners. In total the SIPER database contains 1127 evaluation studies published between 2006 and 2023 covering OECD countries, with an emphasis on UK, Germany, Austria, France, Canada, Scandinavian countries (if an English publication was available) and Latin American countries for the years between 2017 and 2021. As a comprehensive database of evaluation studies, it thus provides a valuable source to assess the integration of unintended consequences into policy evaluation practice.

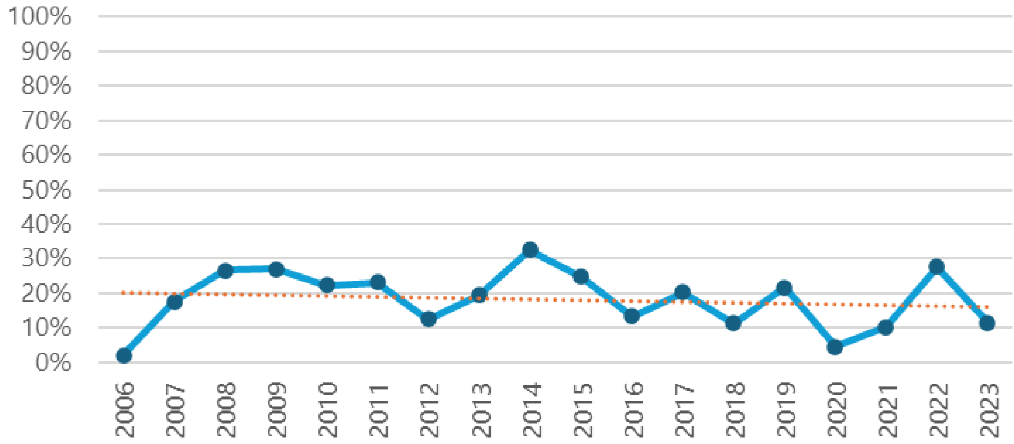
In a first step, we explored the relative frequency of evaluations published in a given year referring to a set of pre-defined keywords for unintended consequences.[3.]

From 2007 until 2023, a total of 208 documents include at least one of the searched keywords, which amounts to 18% of all SIPER evaluations. Besides the rather limited reference to unintended consequences as such, the share of evaluations referring to them is also heavily fluctuating over time, with lowest results around 10% and peaks just over 30%. However, as can be seen from the figure 1, there is no trend towards an increased focus on unintended consequences visible over time. While one can assume a certain time lag with evaluations, the changing policy landscape with its shift towards transformative policy-making (cf. section 1) has, so far, not impacted evaluation practice.

2 www.si-per.eu; SIPER is part of the Research Infrastructure for Science and Innovation Policy Studies (RISIS)and management and maintenance of the database have moved to the Fraunhofer Institute for Systems and Innovation Research ISI.

3 Search terms were: „unintended" and "unexpected" each in combination with "result" "effect" "outcome", "output", "impact", "consequences". Further search terms were: "non anticipated" and "unanticipated". These keywords were selected based on the terminology used in the evaluation guidelines and validated by the literature review. The evaluation studies have been translated into English with the Large Language Models LLaMA (reference: [2307.09288] Llama 2: Open Foundation and Fine-Tuned Chat Models) using the prompt "You are a professional translator. Your task is to translate the text into English accurately. Translate the following text from {language} to English". The subsequent counting and analysis have been conducted with R and Python.

Figure 1 Percentage of evaluation studies per year with at least one keyword related to unintended consequences



Moreover, there is little evidence for a systematic use of specific terms, indicating the co-existence of multiple understandings on unintended consequences Analysing the co-occurrence of 'unintended' (and 'unexpected' as a potential synonym), the analysis revealed a rather unsystematic use instead of a clear pattern. As can be seen from the following table, most hits could be found with the term "unintended" followed by "consequences" and "effect". The term "unexpected" is mostly used in combination with the term"result" but again occurs in combination with a large variety of other terms (outcomes, effects, impacts, etc.).

Table 2: Number of documents in which the keyword occurs

	effect	*outcome*	*output*	*impact*	*consequence*	*result*
unintended	58	14	0	15	50	3
unexpected	7	12	3	9	2	22
unanticipated	69					
Non-anticipated	28					

In order to better understand how the evaluations use these different keywords and what their underlying definition of unintended consequences is, we proceeded with a qualitative text analysis. This analysis was carried out in two steps: Only 58 out of the 208 relevant evaluations contained the keywords at least three times. One may assume that a closer investigation of unintended consequences can be found in these studies, however, this was not the case. No in-depth discussion was found in these 58 evaluations, instead the keywords are mostly used as name-dropping. In most cases, the search terms occur in the introduction, e.g. as a stated goal of the evaluation, or they can be found in the methodology section. Only rarely are unintended consequences discussed in the analysis and finding sections. From those evaluation studies that have a dedicated chapter on unintended consequences, we selected five for an in-depth review.

We screened along the six dimensions and their characteristics discussed in the previous section.

All studies use a logic chart model and a theory of change. In most cases, unintended consequences are detected with qualitative methods, mainly through exchange with different stakeholder groups. The list of unintended consequences can be quite long, leading to the impression that the effect size is rather small, affecting only parts of the target groups. In all cases, the main stakeholder group affected is the beneficiaries. Often, other stakeholder groups are discussed with regard to unintended consequences, in particular organisations supporting the programme or non-successful applicants. Both positive and negative unintended consequences are reported. Even though the screened evaluations have dedicated evaluation questions to unintended consequences, we have the impression that unintended consequences are a residual category for effects reported during the course of data collection but without clear link to an evaluation questions.

In sum, our analysis indicates that despite growing awareness of unintended effects, the concept is only partly enshrined in evaluation guidelines. While the concept is mentioned in guidelines from the past years, it often lacks further hints for practical implication. In a similar vein, the study of unintended consequences remains limited to a minority of evaluations. The combination of quantitative and qualitative text analysis reveals that the concept of unintended consequences is a) still rarely addressed and, in most cases, amounts to a superficial reference and b) when discussed in more depth, there is no systematic use of terminology or consistent application of the concepts.

5. AVENUES FOR DISCUSSION

5.1 SUMMARY OF RESULTS AND CONCLUSIONS

As demonstrated in the literature review, the discussion on unintended consequences suffers from multiple constraints. At the conceptual level, unintended consequences are a heterogenous phenomenon that manifests itself in different dimensions, potentially affecting different stakeholder groups at different points in time. In evaluation practice, despite the claim that unintended consequences are important, we observe only very few in-depth analyses of the topic. Despite regular reference to unintended consequences, these are often treated as a residual category that serves as a catch-all term without further specifying the focus of unintended effects.

From our perspective, this variety is also the key bottleneck for advancing both empirical analysis and conceptual debate, as it subsumes highly heterogenous types of unintended consequences. Against this background, it therefore appears necessary for both evaluation commissioning bodies as well as evaluators to specify the type of unintended consequences more clearly (out of the universe of unintended consequences they are interested in). This is a prerequisite for developing an evaluation design in a way that allows to investigate unintended consequences systematically. Therefore, we propose to disentangle the issue of unintended consequences, suggesting a pragmatic approach that deliberately focuses on conceptually relevant but also empirically feasible unintended consequences. This is in line with recent approaches (Turcotte-Tremblay et al. 2021) that consider a clear definition of terms as a necessity for all evaluations that deal with unintended effects.

5.2 TOWARDS A FRAMEWORK FOR UNINTENDED EFFECTS AND AN EFFECTIVE USE IN MONITORING AND EVALUATION EXERCISES

To better tackle unintended consequences in M&E exercises, there is the need to make the concept more tangible. Furthermore, we argue that different characteristics of unintended consequences are relevant for different M&E activities. We particularly differentiate between evaluation and monitoring exercises. To include the later in the discussion is deemed useful in view of the growing discussion on impact-oriented monitoring (OECD Publishing 2023; Roberts and Khattri 2012).

Drawing on the insights from the literature review in section 3, we argue that one can distinguish six dimensions of unintended effects, requiring evaluators to clarify their priorities: Slight changes have been made with regard to the literature analysis: we have chosen not to include the aspect of "controllability" as it is closely related to the dimension of "knowability". Instead, we included the dimension related to the "scope of the effect", which seems to be more relevant for practical use. To translate it into a usable concept, we transformed the six dimensions into guiding questions for evaluators or other actors dealing with monitoring and evaluation.

iii. **Knowability**: Here we strive for the unintended consequences that could be anticipated, in comparison to those unanticipated and not foreseeable: What unintended consequences would we expect based on tacit knowledge? Tacit knowledge could come from the programme management or from similar types of funding or the evaluator's knowledge of specific contextual factors and challenges that derive from it.

iv. **Stakeholder groups affected** by unintended consequences: What group of stakeholders shall we choose for investigating unintended consequences? As discussed above, stakeholder groups can be beneficiaries of the intervention, but also other groups less directly targeted by the interventions. The selection of stakeholder groups will determine the object of the unintended consequences. Therefore, we suggest using this question as an entry point for analysis of unintended consequences.

v. **Scope of the effect**: Are you interested in unintended consequences affecting the whole group of stakeholders selected (i.e. average effects), or are you rather interested in outliers or effects on a specific subgroup? This comes down to the issue of how many individuals are affected by the unintended consequences.

vi. **Size of the effect**: How strong should the unintended consequence be that you are looking for? Are you interested in knowing about large-scale effects, e.g., with a big implication for the intervention or even beyond, or small-scale effects, e.g., reinforcing the existing trajectory of the intervention only marginally.

vii. **Direction of effects**: Are you primarily looking for negative unintended consequences, or are you also interested in positive or even neutral unintended consequences?

viii. **Timing**: Are you looking at mainly short-term effects that occur during the funding period? Or do you want to know about longer-term effects that materialise only after the end of the intervention?

Taking these guiding questions further, we suggest differentiating between monitoring and evaluation exercises. Depending on what exercise is undertaken, the focus of analysis within these six dimensions will be different. In the following table, we therefore illustrate three archetypical applications of M&E (Monitoring; Summative Evaluation; Formative Evaluation) and argue that different characteristics will be the priority of the analysis. While we are aware that priorities may vary even within these groups (e.g., a stronger focus on learning in monitoring systems), they underline the need to differentiate among different types of unintended effects.

Table 3: Different needs of Monitoring and Evaluation when looking at unintended consequences

	Monitoring	Summative Evaluation	Formative Evaluation
Knowability	Anticipated	Anticipated	Anticipated and Un-Anticipated
Affected stakeholder groups	Beneficiaries	Beneficiaries, unsuccessful applicants	Beneficiaries, unsuccessful applicants, funding organisations
Scope	Larger sub-groups	Sub-Groups	Sub-Groups, Individuals
Size of effects	Medium-sized to large	Medium-sized to large	Including small
Direction of effects	Negative, potentially positive	Negative, positive, potentially neutral	Negative, positive, neutral
Timing of effects	During funding	During and after funding	During (and after funding)

FURTHER DISCUSSIONS – HOW TO STRENGTHEN THE ANALYSIS OF UNINTENDED CONSEQUENCES IN PRACTICE

The most challenging dimension for practical implementation is the knowability-dimension of unintended consequences. The core question here is to what extent unintended consequences may be a priori identified and what context conditions can improve analysis. In the following, we outline the different implications for anticipated, potentially foreseeable and non-anticipated unintended consequences.

Particularly those unintended effects that may be anticipated (e.g., due to known challenges of certain types of funding or context-specific knowledge) might be included from the very beginning in the development of programme theory or in the planning phase of an evaluation, especially through the reconstruction of a theory of change and a set of clear evaluation questions targeting unintended consequences. For these "knowable" unintended consequences, quantitative approaches might be a suitable method to detect their value and especially the size and scope of the effect. Types of knowable unintended consequences can be thought of already in the design phase of the intervention, as assumptions about them are made explicit (Jabeen 2018; Roberts and Khattri 2012; Oliver et al. 2019). As a consequence, taking anticipated unintended consequences seriously would require allocating appropriate (additional) budget and time for the analysis.

In contrast, *potentially foreseeable unintended* effects may require a further strengthening of capacities and knowledge on the subject by actors involved in evaluation and monitoring. Increased knowledge, for instance, could be achieved through a collection and mapping of unintended consequences for specific instrument types or funding mechanisms (e.g., collaborative research). A meta-review of existing evaluations would help to grasp the diversity and guide evaluators. Capacity-building related to the handling of unintended consequences should be provided for both evaluators and programme managers. Particularly evaluators can draw on a wealth of experience from many different situations and interventions. In contrary, programme managers often have (anecdotical) insights from the interventions and can point to the blind spots to look at. But it is only in the interplay of both programme managers and evaluators that potentially foreseeable but intended consequences can be anticipated.

Finally, a particular challenge are those *unintended consequences that are non-anticipated and not potentially foreseeable*, as they cannot be a priori included in a theory of change. Possible avenues here are either a stronger reliance on explorative and qualitative approaches during the course evaluation exercises or relying on beneficiaries and stakeholder engagement (Peterson and Skolits 2019; Bamberger et al. 2016; UNEG Methods Use and Appropriateness Working Group 2022). While the detection of this type of unintended consequences will necessarily partly be coincidental, the choice of methods and thereby the possibility to create space for capturing such effects may make the difference.

6. PUBLICATION BIBLIOGRAPHY

Bamberger, Michael; Hesse-Biber, Sharlene; Tarsilla, Michele (2016): Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. In Evaluation and Program Planning 55, pp. 155–162. Available online at https://doi.org/10.1016/j.evalprogplan.2016.01.001.

Bonell, Chris; Lorenc, Theo; Oliver, Kathryn; Tinkler, Jane (2019): Understanding the unintended consequences of public health policies: the views of policymakers and evaluators. In BMC Public Health 19, Article 1057. Available online at https://doi.org/10.1186/s12889-019-7389-6.

Braun, Dietmar (2009): Evaluation und unintendierte Effekte – eine theoretische Reflexion. In Hildegard Matthies, Dagmar Simon (Eds.): Wissenschaf unter Beobachtung. Effekte und Defekte von Evaluationen. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 103–124. Available online at https://doi.org/10.1007/978-3-531-90863-2_7.

Bruno, Nelly; Kadunc, Martina (2019): Impact Pathways: Tracking and communicating the impact of the European Framework Programme for research and innovation. In fteval Journal for Research and Technology Policy Evaluation (47), pp. 62–71. DOI: 10.22163/fteval.2019.330.

Bührer, Susanne; Feidenheimer, Alexander; Walz, Rainer; Lindner, Ralf; Beckert, Bernd; Wallwaey, Elisa (2022): Concepts and methods to measure societal impacts – an overview. Karlsruhe (Fraunhofer ISI Discussion Papers Innovation Systems and Policy Analysis, 74).

Derrick, G. E.; Faria, R.; Benneworth, P.; Budtz-Petersen, D.; Sivertsen, G. (2018): Towards characterising negative impact: Introducing Grimpact. In : Proceedings of the 23rd International Conference on Science and, pp. 1199–1213. Available online at http://hdl.handle.net/1887/65230.

Diercks, Gijs; Larsen, Henrik; Steward, Fred (2019): Transformative innovation policy: Addressing variety in an emerging policy paradigm. In Research Policy 48 (4), pp. 880–894. DOI: 10.1016/j.respol.2018.10.028.

Dinges, M.; Meyer, Susanne; Brodnik, Christoph (2020): Key Elements of Evaluation Frameworks for Transformative R&I Programmes in Europe. In fteval Journal for Research and Technology Policy Evaluation, pp. 26–40. DOI: 10.22163/fteval.2020.486.

Jabeen, Sumera (2016): Do we really care about unintended outcomes? An analysis of evaluation theory and practice. In Evaluation and Program Planning 55, pp. 144–154. Available online at https://doi.org/10.1016/j.evalprogplan.2015.12.010.

Jabeen, Sumera (2018): Unintended outcomes evaluation approach: A plausible way to evaluate unintended outcomes of social development programmes. In Evaluation and Program Planning 68, pp. 262–274. Available online at https://doi.org/10.1016/j.evalprogplan.2017.09.005.

Leeuw, Frans; Vaessen, Jos. (2010): Impact evaluations and development: NONIE guidance on impact evaluation. Edited by World Bank Group. Washington D.C. Available online at http://documents.worldbank.org/curated/en/411821468313779505/Impact-evaluations-and-development-NONIE-guidance-on-impact-evaluation.

Mazzucato, Mariana (2018): Mission-Oriented Mission-Oriented Research & Innovation in the European Union. A problem-solving approach to fuel innovation-led growth. Edited by European Commission Directorate-General for Research, checked on 5/21/2022.

Meijer, Ingeborg; Sivertsen, Gunnar (2020): Normal versus extraordinary societal impact: how to understand, evaluate, and improve research activities in their relations to society? In Research Evaluation 29 (1), pp. 66–70. Available online at 10.1093/reseval/rvz032.

Merton, Robert K. (1936): The Unanticipated Consequences of Purposive Social Action. In American Sociological Review 1 (6), pp. 894–904. DOI: 10.2307/2084615.

Molas-Gallart, Jordi; Boni, Alejandra; Giachi, Sandro; Schot, Johan (2021): A formative approach to the evaluation of Transformative Innovation Policies. In Research Evaluation 30 (4), 431–442. DOI: 10.1093/reseval/rvab016.

Morell, J. A. (2011): Evaluation in the Face of Uncertainty: Anticipating Surprise and Responding to the Inevitable: Guilford Publications. Available online at https://books.google.de/books?id=sOZSu8vN2eMC.

Morell, Jonathan A. (2005): Why Are There Unintended Consequences of Program Action, and What Are the Implications for Doing Evaluation? In American Journal of Evaluation 26 (4), pp. 444–463. Available online at 10.1177/1098214005281354.

OECD (2019): Better Criteria for Better Evaluation. Revised Evaluation Criteria Definitions and Principles for Use. Available online at https://doi.org/10.1787/15a9c26b-en.

OECD (2021): Applying Evaluation Criteria Thoughtfully. Edited by OECD Publishing. Available online at https://doi.org/10.1787/543e84ed-en.

OECD Publishing (Ed.) (2023): Glossary of Key Terms in Evaluation and Results-based Management in Sustainable Development (Second Edition). Paris.

Oliver, Kathryn; Lorenc, Theo; Tinkler, Jane (2019): Evaluating unintended consequences: New insights into solving practical, ethical and political challenges of evaluation. In Evaluation 26, pp. 61–75. Available online at https://doi.org/10.1177/1356389019850847.

Peterson, Christina; Skolits, Gary (2019): Evaluating unintended program outcomes through Ripple Effects Mapping (REM): Application of REM using grounded theory. In Evaluation and Program Planning 76, p. 101677. DOI: 10.1016/j.evalprogplan.2019.101677.

Roberts, Dawn; Khattri, Nidhi (2012): Designing a Results Framework for Achieving Results: A How-To Guide. World Bank Group. Washington D.C. Available online at https://documents.worldbank.org/en/publication/documents-reports/documentdetail/331541563854787772/designing-a-results-framework-for-achieving-results-a-how-to-guide.

Seus, Sarah; Bührer, Susanne (2021): How to Evaluate a Transition-Oriented Funding Programme? Lessons Learned from the Evaluation of FONA, the German Framework Programme to Promote Sustainability Research. In fteval Journal for Research and Technology Policy Evaluation (52), pp. 10–18. DOI: 10.22163/fteval.2021.514.

Spaapen, J.; van Drooge, L.; Propp, T.; van der Meulen, B.; Shinn, T.; Marcovich, A. (2011): Social impact assessment methods for research and funding instruments through the study of productive interactions (SIAMPI). Final report on social impacts of research. Edited by Royal Netherlands Academy of Arts and Sciences. Amsterdam.

am Turcotte-Tremblay; Gali, I. A.G.; Ridde, V. (2021): The unintended consequences of COVID-19 mitigation measures matter: practical guidance for investigating them. In BMC MEDICAL RESEARCH METHODOLOGY 21 (1). DOI: 10.1186/s12874-020-01200-x.

UNEG Methods Use and Appropriateness Working Group (Ed.) (2022): Compendium of Evaluation Methods Reviewed – Volume II. Real time learning Remote data collection Stakeholders engagement Assessing causal relationship, checked on 1/13/2025.

United Nations Evaluation Group (Ed.) (2016): Norms and Standards for Evaluation. New York.

Vaessen, Jozef Leonardus; Lemire, Sebastian Thomas; Befani, Barbara (2020): Evaluation of International Development Interventions : An Overview of Approaches and Methods. World Bank Group. Washington D.C. Available online at https://documents.worldbank.org/en/publication/documents-reports/documentdetail/942511608652015232/evaluation-of-international-development-interventions-an-overview-of-approaches-and-methods.

Zwart, Frank de (2015): Unintended but not unanticipated consequences. In Theory and Society 44, pp. 283–297. Available online at 10.1007/s11186-015-9247-6.

AUTHORS

SARAH SEUS
Fraunhofer Institute of Systems and Innovation Research ISI
E-mail: sarah.seus@isi.fraunhofer.de
ORCID: 0000-0003-0791-5919

FLORIAN WITTMANN
Fraunhofer Institute of Systems and Innovation Research ISI
E-mail: florian.wittmann@isi.fraunhofer.de
ORCID: 0000-0002-9890-6091

NELE WEIBLEN
Fraunhofer Institute of Systems and Innovation Research ISI,
University of Konstanz
E-mail: nele.weiblen@isi.fraunhofer.de
OCRID: 0009-0006-8762-3527

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e14, 1-17
DOI: 10.22163/fteval.2025.706

DESIGNING A THEORY OF CHANGE IN LESS THAN AN HOUR

ENABLING RAPID, COLLABORATIVE DEVELOPMENT WITH A CONTEXT-INDEPENDENT CARD-SET

ERIKA HAJDU, GIOVANNA LIMA AND STEFAN DE JONG
DOI: 10.22163/FTEVAL.2025.706

ABSTRACT

The increasing focus on societal impact in academia calls for effective tools that can help to address grand challenges through interdisciplinary and transdisciplinary collaboration. While Theories of Change are widely used to plan and evaluate the societal impact of research projects, applying this approach often proves to be time-consuming and resource-intensive, especially in the academic context with rigid structures and competing priorities.

To address some of these challenges, the Evaluating Societal Impact team at Erasmus University Rotterdam (the Netherlands) applied user-centred design to develop a card game that supports the development of Theories of Change in a more accessible, efficient and engaging way. The cards can be used in different contexts and without prior training or knowledge to understand envisioned change processes, significantly reducing the time needed to create a first Theory of Change and making the process fun through gamification.

Trialled with a range of partners in academia, government organisations and municipalities, it has proven to be effective in fostering co-creation, overcoming power imbalances in a group setting and helping to accelerate the development of a shared vision. The tool is gaining widespread interest in the Netherlands and internationally as it offers a context-independent, time-efficient and user-friendly approach to embedding societal impact practices within academia. By providing a concise and engaging experience, the game is used to introduce diverse groups to Theories of Change, fostering interest and engagement with the method. This journal contribution describes our approach and experiences in developing and using the card game. We discuss the possibilities and limitations of the Journey of Progress card set with the aim of inspiring future comparable approaches and solutions in policy evaluation.

Keywords: Theory of Change, user-centred design, card game, project development, impact evaluation, impact planning, impact strategy

1. INTRODUCTION

Academic organizations play a pivotal role in shaping society through becoming more responsive to societal challenges. Researchers are encouraged to address and demonstrate their contribution to solving large societal challenges (de Jong et al., 2022; Global University Network for Innovation, 2017; Perkmann et al., 2020;). Around 7.500 grant applications are submitted to the Dutch Research Council (NWO, 2024) and 35.000 to the European Commission (European Research Council, 2024) annually, which often include societal impact as a cornerstone requiring academics to carry out impact planning and evaluation activities.

Rather than focusing exclusively on research excellence rooted in scientific and commercial impact, literature suggests that focusing on societal impact may benefit from a process-oriented approach based on shared learning and reflection (see for instance Spaapen & Van Drooge (2011) and D'Este et al., (2018)). However, the more complex a project becomes, the harder it is to bring people together and ensure shared ownership and responsibility. Different viewpoints and conflicting priorities need to be managed to ensure that everyone can effectively contribute to a common goal (see Cundill et al., 2018; Kalinauskaite et al., 2021). The endless meetings and dry, bureaucratic processes that are supposed to guarantee synergy can make people lose motivation (Snooks et al., 2023). Facilitating impact activities and collaborating with partners from other disciplines (interdisciplinarity) and beyond academia (transdisciplinarity) calls for alternative methods and practices in many organisations (see D'Este et al., 2018; Perkmann et al., 2020).

In this praxis-oriented article, we demonstrate how we are applying user-centred design to the context of impact evaluation in the academic context. By developing tools that facilitate engaging work processes, the uptake of evaluation methods can be promoted, as we see in an example of a context-independent and hands-on card game based on the Theory of Change method.

2. INTEGRATING THEORETICAL AND PRACTICAL PERSPECTIVES

Many frameworks and methods have been proposed that provide the theoretical background to plan, enact, evaluate and demonstrate one's positive contribution to society (see Smit and Hessels, (2021) for a review of such frameworks e.g., SIAMPI (Spaapen & van Drooge, 2011), ASPIRA (Joly et al., 2015) or see Design for Social Innovation, Transition Design (Irwin et al., 2020)). However, despite the growing demand, the use of theoretical methods across disciplines seems to fall behind their potential (de Jong et al., forthcoming). Although many scientists are motivated to solve societal problems, applying impact evaluation methods can be difficult and time-consuming. Many obstacles, such as financial and time constraints, competing priorities and the lack of available resources (Hughes et al., 2016) make these activities difficult in a context riddled with rigid structures and evaluation cycles.

2.1 THEORIES OF CHANGE IN HIGHER EDUCATION

Within the context of research projects, we see Theories of Change (ToC) (e.g. Belcher & Claus, 2020) as a dominating formative evaluation method. It is widely used in evaluations (Mayne, 2017) and in research proposals, like the Impact Pathway in Horizon Europe application forms (European Commission, 2024). Many public research organisations, funding organisations and consultancy firms in the higher education and research sector as well as independent trainers and facilitators use ToC. At Erasmus University Rotterdam (the Netherlands), we have also seen increased interest towards applying this method. Members of different organisational units and initiatives are looking to develop their ToC in the context of strategy formulation, grant applications or even general project planning.

2.2 EXPERIENCES FROM THE FIELD

A common approach to develop ToCs includes a workshop setting where participants 'build' their ToC; a shared narrative of how and why an intervention is expected to lead to a desired change (Belcher et al., 2020). While there are different ways to approach the development of a ToC (see Mason & Barnes, 2007) we have chosen to do so by means of a joint back-casting exercise, that links a sequence of outcomes and outputs back to activities and inputs.

Many challenges arise when developing a ToC in a group setting. Our experience has shown that the workshops were perceived to take too much time and effort and felt as an additional chore to the research teams, leading to low turnout numbers. The terminology confused participants as terms such as 'theory' and 'assumptions', have a different meaning within the context of the ToC method and the everyday working lives of researchers. The words 'output' and 'outcome' are often a cause for confusion for those who are not versed in impact theory. Additionally, senior researchers tended to dominate discussions, resulting in lower support for the resulting ToC from junior researchers. While these observations are based on the authors' context, these issues seem to be a shared experience across higher education institutions as we have found during peer-learning and networking conversations with many professionals in the field, for instance during a panel at the conference of the European Forum for Studies of Policies for Research and Innovation in 2023.

Thus, the ToC framework is seen as complex, confusing with its jargon and the development requires excessive time investment. This can lead to misalignment, gaps in stakeholder relations, inefficient use of resources or even failing to attract funding. Alternatives to ToC workshops, such as templates and dedicated software's (e.g. TOCO (n.d) or Changeroo (n.d.)) are available yet many of these tools pose similar or additional challenges such as relatively high costs or a steep learning curve which can have similar demotivating effects on potential users. How can we engage people more fruitfully, and align their limited availability with the high demands of the ToC approach to enable its benefits?

2.3 USER-CENTRED DESIGN WITHIN ORGANISATIONS

Design thinking is making its way into public policy due to its ability to approach 'wicked problems' (such as complex societal questions and challenges) from a creative perspective (Van Buuren et al., 2019). Public organisations such as higher education institutes with a focus on societal relevance can benefit from applying designedly approaches to be more responsive to emerging challenges (Muñoz et al., 2023; Vaugh et al., 2020). According to Herbert Simon (2019), design thinking is a process that leads to the creation of any type of intervention that changes existing situations into preferred ones. The particular strand of design known as human-centred design puts people's wants and needs at the centre and aims to fully understand the problems and experiences of those involved in a particular context (van der Bijl-Brouwer & Dorst, 2017). With an emphasis on the users and usability, we can enable the development of new tools and impact evaluation processes that align with people's needs, making these activities more convenient for academics. Taking this approach one step further, gamification (applying elements of game design in a non-game context (Deterding et al., 2011)) has the potential to enhance engagement with diverse tasks and processes (Gupta & Gomathi, 2017).

Several initiatives explore the application of design practices and principles at Erasmus University Rotterdam (EUR), an organisation that does not traditionally offer a degree in design sciences. Design methods are applied in teaching and research, as well as in the operation of the university. Within the Evaluating Societal Impact (ESI) project the role of user-centred design was explored in developing tools that enable the EUR community to maximize their (positive) societal impact. The project ran from 2020 to 2024 within the Strategy Office.

3. APPROACH

Organizations have an important role and the power to shape the impact practices of their community (de Jong & Balaban, 2022). The ESI project at EUR proposed that hands-on tools and gamified solutions can facilitate impact activities within large traditional organizations and contribute to organizational change. By incorporating usability principles and creating new and alternative work processes, the project was hoping to increase the uptake of impact-related methods such as Theory of Change.

3.1 JOURNEY OF PROGRESS – A CARD GAME FOR RAPID TOC DEVELOPMENT.

The ESI team at EUR has taken a user-centred approach to translating the ToC method into a hands on tool for the context of inter- and transdisciplinary projects. 'Journey of Progress' (a card game based on the ToC method, see

Figure 1), addresses the challenges of transformative collaborations and the traditional workshops that we have encountered within our practice. The development of the card game followed the design thinking approach, incorporating frequent iterations based on the feedback and testing with experienced workshop facilitators and future users at the university. The aim of the tool is to empower and enable projects, initiatives and various organisational units to develop a Theory of Change autonomously, discussing their impact and desirable long-term changes (Evaluating Societal Impact, 2024b).

Figure 1: Photo of the card game Journey of Progress, showcasing the content of the game

Using the Journey of Progress card set, members of a group can take a first step towards formalising their understanding on how and why change is expected to occur within their specific context (Evaluating Societal Impact, 2024b). The card game is designed to streamline the ToC process, allowing groups of two to six people to develop an initial ToC in just 40 minutes. Through backcasting and collaborative discussions about a shared goal and the different ways to reach it, participants construct pathways linking their desired future to the specific actions necessary for change.

The game provides a structure for engagement with short, to the point instructions that guide 'players' step-by-step through the construction of their ToC. In the first phase, each player receives six cards to start with, deliberately limiting the number of cards to highlight the resource constraints of real-life practice. Everyone starts with one blank card for a vision statement and three 'change cards', which are outcomes that support their vision. One 'how card' is used to describe an activity to set the desired changes in motion, and one ‘what card’ to write down what is the output of this activity. In the first ten minutes, each player works out their individual contribution by backcasting (See Figure 2). Simple, jargon-free questions and examples on the back of the cards help players articulate their vision for the future, envision future outcomes and identify necessary actions. For example, the ‘how card’ prompts the players by asking *“What do you need/can you do to realise the required change?”*. The back of the card reads: *“A resource, an action, or an intervention within your control that can contribute to change.”* with the example *“Making employees aware of current developments within the organisation.”* to help them envision potential activities.

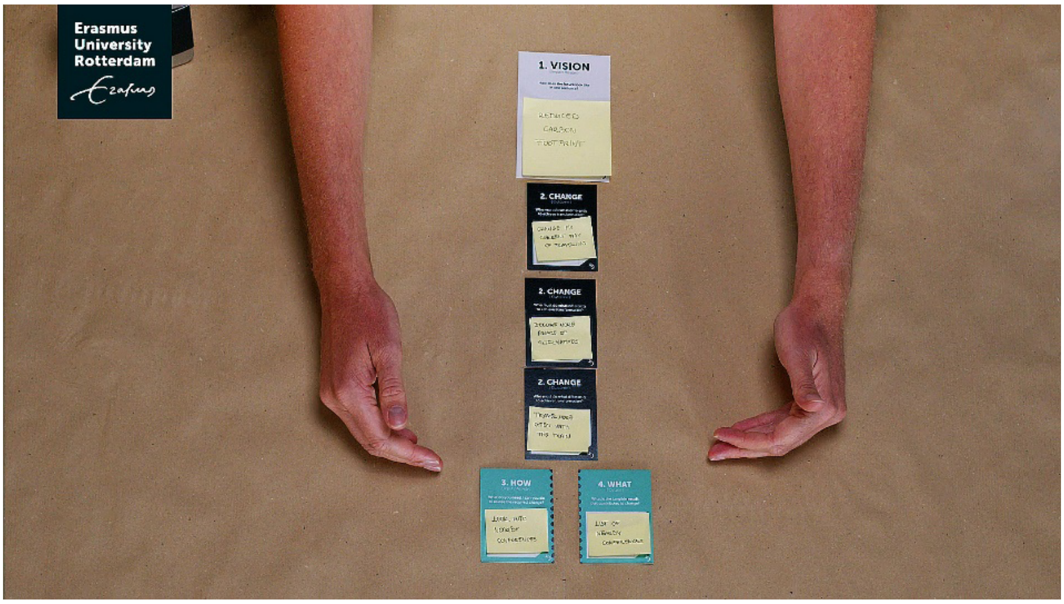


Figure 2: Still from the video ‘How to play Journey of Progress’ (Evaluating Societal Impact, 2024c) showing an individually built pathway after the first phase.

In the second phase, players review everyone’s input and integrate different perspectives through a structured conversation. By systematically reviewing the different cards (first reading out all vision cards, then all change cards, etc.) players build on each other’s input and merge their contributions, discharging cards that do not fit the joint ToC. The structure of individual work and group discussion encourages contributions from everyone around the table, as ideas are first made explicit on the cards. This approach is designed to avoid the phenomenon of groupthink, where everyone accepts the ideas of the first speaker without ownership of the discussion and its results. With carefully determined time limits and using sticky notes on the cards, creating the first draft of a ToC becomes a time-bound yet productive and flexible activity.

The standard game tackles the basics of autonomously building a ToC including a vision statement, different outcomes and outputs, inputs, and activities. Accompanying short videos explaining the basics of the ToC approach and the use of the card game support the players. Additionally, a 'how-to guide' with instructions and tips is available for project leaders and facilitators ('game masters'). While the game does not explicitly go into the details of power relations, trade-offs and other contextual issues relevant to the development of a ToC, those wanting to refine the results of the standard 40-minute session can do so by extending the duration of the session. Further extensions are envisioned, such as cards for defining stakeholders or bringing specific assumptions to the surface. Another extension of the game helps users to focus on developing specific indicators for monitoring and assessing their planned changes.

The card game can be used in various contexts to understand envisioned change processes. While the primary target audience are people with limited to no experience in impact evaluation, the cards can be used by experts and facilitators of impact evaluation to introduce and apply the ToC approach with various target audiences in a short, effective way. The card game, just as other ToC tools, can serve as a basis for developing relevant indicators for measuring societal impact of research projects, organisational strategies, as well as social innovation within transformative R&I policies.

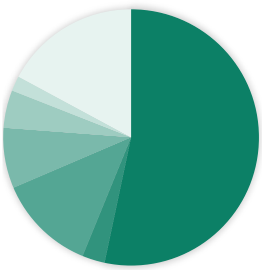
4. RECEPTION AND FEEDBACK

The tool has been extensively trialled and tested across multiple types of use cases. The writers have used Journey of Progress in different projects, with a variety of stakeholders (Figure 3). Examples include various academic and administrative organisational units within the university, municipalities such as Amsterdam, the Hague and Rotterdam, and governments and government agencies, such as the Dutch Ministry of Science and the Dutch Research Council.

Figure 3: Impressions of the Journey of Progress card game in use.

Feedback has been positive and the demand for the card game indicates a wide interest. By the end of April 2025, over 290 boxes of Journey of Progress have been distributed to more than 105 organisations in 20 countries (see Table 1 for an overview). Users include the Swiss National Science Foundation, Poland's National Science Centre, the Italian Presidency of Ministries and Germany's Standing Scientific Commission on Education Policy, using this tool for organisational change management and discussions on societal impact. We see use cases ranging from project level applications to departmental and organisational level discussions on strategy, planning and evaluation. However, we lack detailed data on how all these users apply the tool in terms of their level of facilitation and the use of supporting materials. Evidence of the tool's effectiveness is primarily drawn from qualitative feedback based on more than 20 sessions by ESI team members, where we experimented with different levels of facilitation of the players.

People experience the card game as "*enjoyable*" and recommend the game to others in their network. One player described the session as "*a good combination of individual contribution and collective discussion*" and the quick format seems to be appreciated by users. "*The best part was that it never felt like a compromise. The co-creation process left us more energized than at the start.*", as commented by a member of a cross-European educational network. The tool has been highly sought after, with colleagues from higher education institutions and public research organizations wishing to purchase the tool, as well as requesting workshops using the card game, which indicates its perceived usefulness by users.



Type of organisation	Main purpose of application[1]
53% Education institutions *(higher education institutions)*	Stimulation impact-thinking *(individuals)*
	Research support *(individuals and projects)*
	Developing funding proposals *(projects and programs)*
	Strategic development *(departments)*
17 % Research and scientific entities *(research centres, institutes, foundations. collaborations, networks, associations, and consultancies)*	Strategic development and goal identification *(organisation)*
	Stimulation impact-thinking *(individuals)*
	Supporting collaborations *(projects and programs)*
12% Government and public sector *(ministries, municipalities, health organisations and utilities)*	Strategic development *(organisation, teams)*
	Change management *(organisation)*
	Supporting collaborations *(projects)*
	Policy evaluation
8% Networks and collaborative structures *(network organisations, university networks and transdisciplinary collaborations)*	Developing Theories of Change *(programs)*
	Supporting collaborations *(projects and programs)*
5% Non-profit and philanthropic entities *(funding agencies, fundraising agencies and other non-profit organisations)*	Supporting collaborations *(projects and programs)*
	Developing Theories of Change *(not specified)*
	Monitoring *(not specified)*
2% Private sector *(private consultancies, commercial companies)*	(Impact) training and consultancy
3% Else	Supporting collaborations
	Developing funding proposals
	Personal/family planning

Table 1: Overview of the type of organisations and their main purpose of using the game based on orders and workshops given to 105 organisations between May 2024 and 2025.

5. DISCUSSION

Reflecting on the first-hand experiences of our team as well as early adapters of the game, we foresee that user-centred tools such as the Journey of Progress card game have potential in the impact evaluation space. The feedback of players as well as facilitators confirms our initial observations that the ‘traditional’ ToC workshops, across different providers, are experienced as ‘dry’ and ‘boring’ as opposed to a gamified approach that accounts for the user’s needs. Our experiences indicate positive outcomes in terms of quickly developing a shared vision with different groups and in complex environments, overcoming power differences between participants and providing immediate value to researchers and other users with quick results, as well as to ToC facilitators by increasing participation and engagement. The card game can be an effective way to collect required input for funding applications or systemically integrate impact thinking in an organisation.

When compared to other tools the Journey of Progress card set presents some practical benefits including accessibility, time efficiency, adaptability and cost efficiency. The game can be used autonomously and does not require users to have previous knowledge of the ToC method which may support broader uptake, though this simplicity might limit the depth of discussions and reflection typically expected in ToC development. The game is context independent and suitable for a wide range of domains, unlike fixed-context alternatives, (e.g. CESVI’s (2020) card set or the Theory of Change Game by Tribaldos and Schneider (2021)). Journey of Progress also allows for easy iterations, as opposed to fixed formats (such as the CUBISS worksheet (Jans et al., 2019)). Reducing the time requirement can potentially enhance participant’s motivation while also reducing costs, as the hourly fees for facilitators and participants can rapidly add up during traditional, multi-hour workshops. These observations are based on our self-assessment of the strengths and limitations of JoP compared to these other approaches and we have not conducted an in-depth comparison with users.

5.1 LESSONS LEARNED AND LIMITATIONS

Translating theoretical methods into hands on tools for impact evaluation brings its unique challenges. The ESI project has developed multiple tools for enabling impact, for different organisational levels and phases of impact-related activities (Evaluating Societal Impact, 2024a). The development of such tools using user-centred design requires engagement from the community ('end-users') to ensure their usefulness. Stimulating co-creation from within while avoiding research fatigue of the target audience and negotiating one's span of control (Should we ensure the use of such tools?) can be a challenge. Drawing upon the international community of impact and policy evaluation has been useful to gather similar experiences, inspiration and feedback.

Both in regular workshop settings as well as when using the Journey of Progress card game, there are many factors influencing the quality of the session results. We acknowledge that using a card game in just 40 minutes, especially without an experienced facilitator, might not produce the same depth or quality results as for example a two-day long facilitated workshop, where each contribution is carefully reviewed. The real merit of the game is giving a 'taste' in a concise, interactive way so that research groups and initiatives are more likely to engage with the ToC approach. Players are advised to photograph and revisit their ToC after a game session and to extend and build upon it – just as with a ToC constructed in another ways.

We found the total autonomy of players being a difficult goal to reach. We understand total autonomy in this context as a group of non-experts organising themselves and applying a tool without external guidance (an expert facilitator). On the one hand, workshop participants expect guidance when an external party is present, as they are used to have so in the context of a workshop. Usually someone takes charge to lead the group at a table, but we encourage assigning a 'game master' who's explicit role is to keep an eye on the time and the scope of the discussion. On the other hand, we see improved results when participants are introduced to the approach through the accompanying videos - or better, through a presentation where they could ask questions about the method. A session also benefits from a joint reflection on the results, therefore, as of current we do not support fully autonomous approaches and advise the presence of a facilitator or an experienced 'game master'.

5.2 NEXT STEPS

The game was developed as part of a strategic project, with no additional funding allocated for further validation. At present we cannot say that the use of Journey of Progress would lead to more successful or higher quality grant applications, better evaluation practices, let alone societal impact. A systemic comparison across the effect of different methods of constructing a ToC is an interesting future venue. Furthermore, we acknowledge that using the above-described card game might not offer an advantage in every single setting and the game's usefulness in a much broader context (such as policy evaluation) could be a potential future research topic.

Journey of Progress has the potential to be expanded and scaled to different applications. The above-mentioned extensions (focusing on stakeholders, assumptions or indicators), while they might extend the timeframe of a session, could be beneficial in many settings. Moreover, in situations in which all partners cannot be physically present (which is often a case with inter- and transdisciplinary, cross-border projects and initiatives), users can benefit from a digital version of the game. While the physicality of the exercise allows for a more profound experience and fun interaction, we have successfully trialled Journey of Progress in an online collaborative environment (Miro). Furthermore, the tool can be also supported with interviews that allow for the perspectives of those who cannot be in the (virtual) room: they still get represented in the design of the shared ToC.

6. CONCLUSION

Theory of Change (ToC) workshops remain a popular method of formative impact evaluation of research projects, with many challenges that can hinder the application of this approach. We have seen that observing the issues that people face during ToC workshops and providing solutions for those issues can result in novel solutions that not only pique people's interest but contribute to the use of methods that are otherwise seen as dull or too theoretical. The card set described in this article allows for a more interesting, structured and time efficient approach to developing a ToC. The results have been positive as reported by multiple teams who have used the game to get acquainted with the method or to develop their ToC, positioning this solution as a valuable asset in a broader set of tools and activities related to impact evaluation. In this praxis-oriented journal contribution we describe our experiences to invite and inspire academics and practitioners of impact- and policy evaluation to consider alternative methods, such as the use of the Journey of Progress card game in their work that can allow members of a wide range of projects and initiatives to talk about complex processes in a concise, effective way.

REFERENCES

Belcher, B., & Claus, R. (2020). *Theory of change*. Td-net toolbox profile (5). Swiss Academies of Arts and Sciences: td-net toolbox for co-producing knowledge. www.transdisciplinarity.ch/toolbox. doi.org/10.5281/zenodo.3717451

Belcher, B. M., Davel, R., & Claus, R. (2020). A refined method for theory-based evaluation of the societal impacts of research. *MethodsX*, *7*, 100788. https://doi.org/10.1016/j.mex.2020.100788

Van der Bijl-Brouwer, M., & Dorst, K. (2017). Advancing the strategic impact of human-centred design. *Design Studies*, *53*, 1–23. https://doi.org/10.1016/j.destud.2017.06.003

Van Buuren, A., Lewis, J., Guy Peters, B., and Voorberg, W. (2020) Improving public policy and administration: exploring the potential of design,Policy & Politics, vol 48, no 1, 3–19, DOI: 10.1332/030557319X15579230420063

Cesvi. (2020, January 23). The theory of change action cards. CESVI Learning in Palestine. Retrieved January 17, 2025, from https://cesvilearningpalestine.home.blog/the-theory-of-change-cards/

Changeroo. (n.d.). [Software]. Business for Development. https://changeroo.com/

Cundill, G., Harvey, B., Tebboth, M., Cochrane, L., Currie Alder, B., Vincent, K., Lawn, J., Nicholls, R. J., Scodanibbio, L., Prakash, A., New, M., Wester, P., Leone, M., Morchain, D., Ludi, E., DeMaria Kinney, J., Khan, A., & Landry, M. (2018). Large Scale transdisciplinary collaboration for adaptation Research: Challenges and Insights. *Global Challenges*, *3*(4). https://doi.org/10.1002/gch2.201700132

D'Este, P., Ramos-Vielba, I., Woolley, R., & Amara, N. (2018). How do researchers generate scientific and societal impacts? Toward an analytical and operational framework. *Science and Public Policy, 45(6),* 752–763. https://doi.org/10.1093/SCIPOL/SCY023

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness. *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, *11*, 9–15. https://doi.org/10.1145/2181037.2181040

European Comission (2024). Horizon Europe Programme: Standard Application Form (HE RIA, IA). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/af/af_he-ria-ia_en.pdf

European Research Council. (2024, December 13). *Dashboard of ERC funded projects and evaluated proposals*. Retrieved January 16, 2025, from https://erc.europa.eu/projects-statistics

Evaluating Societal Impact. (2024a). *Impact Toolbox* (By Erasmus University Rotterdam). Retrieved January 10, 2025, from https://www.eur.nl/en/research/research-services/societal-impact-evaluation/impact-toolbox

Evaluating Societal Impact. (2024b). Journey of Progress – Theory of Change card game. Erasmus University Rotterdam (EUR). https://doi.org/10.25397/EUR.25311478.V1

Evaluating Societal Impact. (2024c, February 22). *How to play Journey of Progress* [Video]. YouTube. https://www.youtube.com/watch?v=YQFCySkvp-0

Global University Network for Innovation. (2017). Higher education in the world, 6: Towards a socially responsible university; balancing the global with the local. GUNI. https://unesdoc.unesco.org/ark:/48223/pf0000248150

Gupta, A., & S, N. G. (2017). A Review on Gamification and its Potential to Motivate and Engage Employees and Customers. *International Journal of Sociotechnology and Knowledge Development*, *9*(1), 42–52. https://doi.org/10.4018/ijskd.2017010103

Hughes, A., Lawson, C., Salter, A., Kitson, M., Bullock, A., & Hughes, R. (2016). *The Changing State of Knowledge Exchange: UK Academic Interactions with External Organizations 2005-2015*. NCUB. https://eprints.lancs.ac.uk/id/eprint/145906

Irwin, T., Tonkinwise, C., & Kossoff, G. (2020). Transition Design: An educational framework for advancing the study and design of sustainable transitions. *Cuadernos Del Centro De Estudios De Diseño Y Comunicación*, *105*. https://doi.org/10.18682/cdc.vi105.4188

Jans, L., Somers, E., & Cubiss. (2019). *Impact toolkit: De Theory of Change*. https://www.cubiss.nl/impact-toolkit

Joly, P., Gaunand, A., Colinet, L., Larédo, P., Lemarié, S., & Matt, M. (2015). ASIRPA: A comprehensive theory-based approach to assessing the societal impacts of a research organization. *Research Evaluation*, *24*(4), 440–453. https://doi.org/10.1093/reseval/rvv015

De Jong, S. P. L., & Balaban, C. (2022). How universities influence societal impact practices: Academics' sensemaking of organizational impact strategies.

Science and Public Policy, 49(4), 609–620. https://doi.org/10.1093/SCIPOL/SCAC012

De Jong, S. P. L., Balaban, C., & Nedeva, M. (2022). From 'productive interactions' to 'enabling conditions': The role of organizations in generating societal impact of academic research. Science and Public Policy, 49(4), 643–645. https://doi.org/10.1093/SCIPOL/SCAC015

De Jong, S. P. L., Diaz-Faes, A. A., Woolley, R., & Molas-Gallart, J. (forthcoming). Investigating the use of qualitative and process oriented societal impact assessment approaches. In M. Gulbrandsen, C. Donovan, & S. Tellman (Eds.), *Handbook on the societal impact of research*. Edward Elgar Publishing.

Kalinauskaite, I., Brankaert, R., Lu, Y., Bekker, T., Brombacher, A., & Vos, S. (2021). Facing Societal challenges in living Labs: Towards a conceptual framework to facilitate transdisciplinary collaborations. *Sustainability*, *13*(2), 614. https://doi.org/10.3390/su13020614

Mason, P., & Barnes, M. (2007). Constructing theories of change. *Evaluation*, *13*(2), 151–170. https://doi.org/10.1177/1356389007075221

Mayne, J. (2017). Theory of Change Analysis: Building Robust Theories of Change. *Canadian Journal of Program Evaluation*, *32*(2), 155–173. https://doi.org/10.3138/cjpe.31122

Muñoz, C. B., Nanclares, N. H., Zamorano, L. R. M., & Sánchez, J. Á. L. (2023). Design thinking in higher education. In *Routledge eBooks* (pp. 37–61). https://doi.org/10.4324/9781032675558-3

Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). (2024). *NWO Jaarverslag 2023*. https://www.nwo.nl/sites/nwo/files/media-files/NWO%20Jaarverslag%202023.pdf

Perkmann, M., Salandra, R., Tartari, V., McKelvey, M., & Hughes, A. (2020). Academic engagement: A review of the literature 2011-2019. *Research Policy*, *50*(1), 104114. https://doi.org/10.1016/j.respol.2020.104114

Prinsen, G., & Nijhof, S. (2015). Between logframes and theory of change: reviewing debates and a practical experience. *Development in Practice*, *25*(2), 234–246. https://doi.org/10.1080/09614524.2015.1003532

Simon, H. A. (2019). *The Sciences of the artificial*. https://doi.org/10.7551/mitpress/12107.001.0001

Smit, J. P., & Hessels, L. K. (2021). The production of scientific and societal value in research evaluation: a review of societal impact assessment methods. *Research Evaluation*, *30*(3), 323–335. https://doi.org/10.1093/reseval/rvab002

Snooks, H., Khanom, A., Ballo, R., Bower, P., Checkland, K., Ellins, J., Ford, G. A., Locock, L., & Walshe, K. (2023). Is bureaucracy being busted in research ethics and governance for health services research in the UK? Experiences and perspectives reported by stakeholders through an online survey. *BMC Public Health*, *23*(1). https://doi.org/10.1186/s12889-023-16013-y

Spaapen, J., & Van Drooge, L. (2011). Introducing "productive interactions" in social impact assessment. *Research Evaluation*, *20*(3), 211–218. https://doi.org/10.3152/095820211x12941371876742

TOCO (Theory of Change Online). (n.d.). [Software]. ActKnowledge. https://www.actknowledge.org/toco-software/

Tribaldos, T., & Schneider, F. (2021). Enabling players to develop theories of change for sustainable development: a serious game. *Simulation & Gaming*, *52*(5), 664–678. https://doi.org/10.1177/10468781211022399

Vaugh, T., Finnegan-Kessie, T., Donnellan, P., & Oswald, T. (2020). The potential of design thinking to enable change in higher education. *All Ireland Journal of Higher Education*, *12*(3).

AUTHORS

ERIKA HAJDU
Erasmus University Rotterdam
Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
Email: erika.hajdu@eur.nl
ORCID: 0009-0005-3492-8522

GIOVANNA LIMA
Erasmus University Rotterdam
Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
Email: giovanna.lima@sfdora.org
ORCID: 0000-0003-2797-0034

STEFAN DE JONG
Erasmus University Rotterdam
Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
Email: s.p.l.dejong@essb.eur.nl
ORCID: 0000-0001-5145-4393

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e15, 1-14
DOI: 10.22163/fteval.2025.707

TOWARDS AN EVALUATION FRAMEWORK FOR INTERNATIONAL STRATEGIC PARTNERSHIPS BETWEEN UNIVERSITIES

A CO-CREATIVE APPROACH

LEONIE VAN DROOGE AND NIKI VERMEULEN
DOI: 10.22163/FTEVAL.2025.707

ABSTRACT

Formalised international strategic partnerships between universities are relatively new. Such agreements include both research and education and cover a range of departments across the partner universities. As these partnerships are expected to contribute to strategic goals and have great impact, the question of evaluation becomes prominent.

This paper presents a project dedicated to the development of a framework for the evaluation of international strategic partnerships. The project was a collaboration between international officers of six universities (of which five in Europe) and academics of two of these universities. It was decided to co-create a framework. The academics guided the international officers through an evaluation of a strategic partnership and developed and tested the framework on the go.

The result is an evaluation framework that is very different from what the international officers initially envisaged. Yet it has changed the way in which evaluation thinking is integrated in the practice of international partnerships in these universities.

Keywords: Strategic partnership, internationalisation, research collaboration, evaluation, co-creation, transdisciplinarity

1. INTRODUCTION

University researchers collaborate across borders and continents. Students go on exchanges and study abroad. A more recent development in the internationalisation of universities is the establishment of formalised international strategic partnerships between institutions. These formal agreements include both research and education and they cover a range of departments. As these partnerships are expected to contribute to strategic goals of universities and have great impact, the question of management and evaluation becomes prominent.

International officers of six universities[1] observed a growth in these strategic partnerships, and a lack of consistent and aligned evaluation practices. They aimed to develop a framework for the evaluation of international strategic partnerships. The framework needed to be useful during all phases of a partnership, in any institutional setting and for a variety of partnerships. Moreover, the framework had to be based on existing literature. Also, the framework needed to be aligned with the evaluation practices in their respective universities. In the end, the framework had to contribute to the quality and sustainability of the partnerships. They initiated the EVALUATE project.

After the start of the project, the internationalisation officers extended the core project-group by including researchers with expertise in collaboration, internationalisation and evaluation. The development of the EVALUATE framework (EVALUATE project, 2022) thus became a transdisciplinary project, in which researchers and practitioners collaborated. The core question of the project and this paper was "how to develop a framework for the evaluation of international university partnerships, that is useful for the daily practice of internationalisation officers and that is based on the state-of-the-art literature". The researchers proposed a co-creative approach to make sure that the framework was embedded in both theory and practice.

The project went through several phases. From orientation, through extension & design of the process, to design & development of the framework. The dissemination phase started after the formal conclusion of the project. The remainder of the paper describes all these phases except the dissemination phase.

1 Coordinator University of Edinburgh plus strategic partners University of Sydney, University of Copenhagen, University of Helsinki, University College Dublin, and Leiden University.

It should be noted that the project took place during the Covid19 pandemic, i.e. 2020-2022. All but two meetings took place online. The second and final in person meeting took place in October 2022, after which the dissemination phase started.

2. ORIENTATION

During the first phase of the project the international officers focused on the definition of international strategic partnerships. They could not find a straightforward definition in the literature. Moreover, they identified a broad range of activities that they themselves covered in these partnerships. They therefore developed their own working definition:

“A strategic partnership is a formal relationship between two or more universities. It is centrally supported and takes the form of a top-down engagement that depends on a bottom-up approach. A strategic partnership is university-wide, covers a range of departments and includes both research and education. A strategic partnership often demands a high level of engagement from the involved parties and can deliver greater impact than the sum of the individual parts” (EVALUATE project, 2022, p. 7).

The project team learned that there were few relevant studies on the topic of international strategic partnerships. They reached out to researchers working at their universities and asked for support in the development of the framework.[2]

3. EXTENSION & DESIGN OF THE PROCESS

After they joined the project, the researchers focused on three aspects: (1) a substantial literature review; (2) the daily practice and the core issues of international strategic partnerships; and (3) the design of the process that would lead to an evaluation framework.

From the literature review (EVALUATE 2022, 112-170) they learned there is ample literature on internationalisation, mobility, and environmental impacts of international collaboration. Yet it is not straightforward to find literature

2 Science Technology and Innovation Studies (STIS) at University of Edinburgh and Centre for Science and Technology Studies (CWTS) at Leiden University.

immediately relevant for international strategic partnerships between universities, nor for the evaluation of such agreements. Moreover, the term of strategic partnership has different meanings. Plus, these partnerships are treated differently in each of the participating universities.

Consequently, the notion of international strategic partnerships was unpacked, to understand the different forms and formats it can take. This approach delivered keywords for a broader literature review, which put forward various, separate bodies of literature that are all relevant to international partnerships. The literature review provided the evidence base to develop the framework. However, it also became clear that there is no precedent for the evaluation of strategic partnerships.

Next to the literature review, the researchers acquainted themselves with the practice of working with international strategic partnerships. They participated in the online meetings of the project groups and identified a number of aspects to take into account. Despite the definition developed in the first phase, the international officers did not use “strategic” and “partnership” in a consistent way. Partnerships could include non-academic partners, a limited number of departments only, or would not have a direct link with the university strategy. Moreover, some of the international offices are responsible for dozens, if not hundreds, of agreements and memorandums of understanding. As a consequence, the management of most of these partnerships was light.
And finally, even though the international offices were responsible for the evaluation of the partnerships, most of them had little experience and capacity.

The researchers proposed to develop a framework from scratch. They chose a co-creative approach to make sure that the framework was embedded in both theory and practice. They opted to develop a flexible framework, that could be used for different evaluation phases, questions and partnerships. They aimed to deliver a framework that was easy to use, given the limited capacity and experience. They also suggested that the framework would include a number of questions and options. This deviated from the initial desire for an evaluation framework or device that would lead to a clear evaluation result, without too much effort. One of the researchers had developed a similar framework before (Isabelle van Elzakker & Leonie van Drooge, 2019) and this helped convince the international officers to proceed.

The proposed approach was as follows. Each international office (six in total) was asked to do an evaluation of a strategic partnership. The researchers would guide the staff through the evaluation while simultaneously and iteratively developing the evaluation framework. They would adjust it based on the feedback and responses of the international office staff. More specifically, there would be four online sessions which each international office, one online plenary workshop with presentations and one live workshop with reflections.

4. DESIGN & DEVELOPMENT OF THE FRAMEWORK

Each international office chose one partnership to evaluate. Five of the six evaluations are included in the handbook (EVALUATE project, 2022, p. 46-111):

1. the evaluation of a first international partnership of the university, in order to learn for future partnerships;
2. the evaluation of the value of an international partnership with academic as well as non-academic partners;
3. the midterm enhancement-led evaluation of a small partnership, in which both universities participated;
4. the evaluation of a partnership with one university, with the focus on the value of the alignment of teaching and research;
5. the evaluation of a partnership with one university, with the focus on the commitment and participation of internal stakeholders, within the university.

These examples illustrate how the case studies extend the definition of strategic partnerships. Example 2 included non-academic partners, example 3 consisted of a very small partnership of two research departments per university only and from example 4 it became clear that international strategic partnerships sometimes only focus on education or research, instead of both.

The researchers developed an agenda for the development of the framework and the support of the evaluation, divided over four meetings:

1. Everything about the strategic partnership. Result: Context + evaluation question + evaluation form
2. Data collection. Result: information
3. Analysis of data. Result: evidence

4. Interpretation of evidence, reflection, conclusion.
   Result: Assessment (+case report).

Each of the meetings was online, including representatives of the research team and of the specific case study team. During the meetings, the researchers learnt that the topic of the first meeting, "everything about the strategic partnership" remained on the table throughout the evaluation, including the history of the partnership, the content of the agreement, the overall goals, the specific targets, the implementation plan, the inclusion of stakeholders (especially in the own organisation), and results. Moreover, they learned that basic information about evaluation methods was required. International office staff was keen to use bibliometric tools but were unfamiliar with recent developments. Plus, international office staff wanted to do interviews and surveys, yet they were not used to design and conduct those.

In addition, the researchers learned from the meetings that the use of a script or a series of questions would be useful. This was confirmed at the first in person meeting of the project. Several of the international officers mentioned that evaluation "is about asking the right questions." They also decided to remain using the word "evaluate" (which is also the name of the project). Using the word helped to stress that the efforts should be systematic and dedicated to judge merit, worth or significance by combining evidence and values (Better Evaluation, n.d.).

The final framework had a different order than the initial agenda set for the meetings. The framework consisted of a series of questions, grouped per topic. The two basic questions underlying the evaluation are:

1. What is the partnership about?
2. What is the evaluation about?

The framework consists of four categories of questions and suggestions

1. The evaluation and its context (figure 1)
2. The central evaluation question (figure 2)
3. The partnership and its context (figure 3)
4. Methods

1. THE EVALUATION AND ITS CONTEXT

Questions to consider when planning an evaluation

- Why evaluate now?
- Who asks for the evaluation?
- Who is involved and in what role?
- What is at stake? What are the consequences?
- What will the follow up look like? What will happen with the results of the evaluation?

Questions that address the more technical aspects of the evaluation

- **When should the evaluation take place?**
  - Prior to formalizing the partnership
  - During the partnership
  - Towards the end of an agreement term
- **What is the goal of the evaluation?**
  - To decide: "do we want to partner with university X?"
  - To reflect and improve: "How does the partnership develop?"
  - To understand outcome: "What are the results of the investment?"
  - To monitor: "What are the investments, activities and results?"

How is the evaluation/decision organised?

- What will be used as evidence?
- What methods will be used to collect and analyse information?
- Who will decide? Who assesses?

Figure 1: The evaluation and its context

2. THE CENTRAL EVALUATION QUESTION

Questions regarding the relationship with certain strategies / policies / aspirations

- To what extent does the partnership contribute to strategy X / policy Y / impact Z?
- How do we ensure the partnership will contribute to strategy X / policy Y / impact Z?

Phase-specific questions

- **Phase one: before the partnership**
  - Who do we want to partner with and why?
  - What can the partnership deliver? To us (and who is us?), the partner, society?
  - What are the aims/goals and how can they be reached?
  - What are the potential risks? How are they mitigated?
- **Phase two: during the partnership**
  - What can be done to improve implementation?
  - What can be learned for other partnerships?
- **Phase three: Late in, or after, the term of an agreement**
  - Do we want to renew the partnership?
  - What has been achieved?

Figure 2: The central evaluation question

3. THE PARTNERSHIP AND ITS CONTEXT

Questions focused on the partners

- What external partners are involved?
- What is your institution's history with this partner?
- What departments are involved in the university?
- Are there any champions that play a key role in the partnership?

Questions focused on the goals and ambitions of the partnership

- What is the rationale for the partnership?
- How is (/will be/has been) the partnership formalized?
- What are the goals of the partnership?
- What are goals and expectation of the partner? Has this been discussed?
- How is the partnership implemented?

Questions focused on the partnership activities

- What activities are part of the partnership?
- What is the starting situation?
- Who is involved in the partnership?
- What are strategies for the partnership?
- Is there an implementation plan?
- Does the university make funding available?

Questions focused on expectations regarding the partnership

- What does success/value mean? And at what cost?
- What results are expected?
- What further impact is foreseen?
- What are potential risks? Perceived by whom? How about risk mitigation?

Other questions about the partnership

- What is the history of this partnership / collaboration?
- What funding opportunities are available?
- What regional/national/supranational policies are relevant?
- What institutional policy arrangements and contexts are relevant?

Figure 3: The partnership and its context

The project resulted into a handbook (EVALUATE project, 2022) that is available online (Open Access). It includes the evaluation framework with questions and examples; five of the case studies; and the literature review. The handbook also includes a section on lessons learnt, written by the international office staff. The lessons are:

1. Know your stakeholders and their (naturally diverse) interests. This includes the recommendation to invest upfront in the relationship with the partner, and arrive at a mutual understanding.

2. Integrate evaluation with existing data and systems. This includes practical recommendations, such as deciding on targets, the inclusion of internal stakeholders that are responsible for the implementation of the partnership and the collection of data.

3. Get to know evaluation methods. This includes the recommendation to make evaluation part of the management of a partnership and the option to use interactive and participatory methods, if only to come to mutual understanding.

4. Expect change in ideas about evaluation - and to invest time and effort. The officers concluded that their understanding about the power of evaluation changed radically as a result of the intensive discovery process they went through in the course of developing case studies: "In common with many experiences of radical change, our preconceptions were disrupted. We strove to find meaning. And finally, we came to terms with a new reality. It's also worth reflecting on the significant scale of this intensive discovery process - and ensuring that those involved have the time and resources they need". (EVALUATE handbook, 2022, p. 45).

The handbook was launched at the second and final in-person meeting during the International Association of Universities General conference. The conference took place at University College Dublin, October 2022.

5. CONCLUSION & DISCUSSION

The EVALUATE project was a learning journey for all. The governance philosophy of the various universities, the partnership activities, the goals of the partnerships and the relation to university strategies differ between the partners and cases. The lack of consistent and aligned evaluation practices was confirmed throughout the project. Consequently, partners realised that a rigid framework with clear measures or benchmarks is not realistic or useful. Evaluation is best integrated from the start to the end of partnerships. If integrated well, evaluation is a cyclical activity returning in every phase of the partnership, underpinning decisions, and new actions. This was also recognised by the international officers.

Only through mutual exploration of the practice of international strategic partnerships, did it become possible to find ways to mobilise existing insights from the evaluation literature and integrate this in a meaningful way. As such, the creation of an evaluation framework was an experiment in co-creation, developing evaluation in practice with those using the framework. By working on the development of the framework from the start, it was ensured that the resulting framework fits existing practices and that it can more easily be integrated into ongoing partnership work. This enhanced the capability for implementation and allowed project participants to communicate results and spread evaluative thinking.

Our project to co-create an evaluation framework was innovative in a number of ways. Firstly, the framework provides a solution for university staff working in the areas of partnership development and evaluation. This is an area of need, given recent and current prioritisation of, as well as debates about international partnerships in both institutional, national and international strategies that concern university education and research. In addition, the co-creation with a project team consisting of academic evaluation specialists and international officers was key. The combination of academic and professional inquiry enabled the project to apply robust academic methodologies and scrutiny alongside professional experience and expertise in the field. Moreover, the project was complimentary to a wide range of other initiatives and can support the higher education sector to forge greater understanding of the impact of university partnerships. For example, the UK Research Excellence Framework (REF) highly values the impact of research. The impact of research often happens through partnerships – and this project provides a framework to understand the value of partnerships.

Most importantly, the co-creation approach fits with current calls for reforming evaluation culture (CoARA, 2022). First of all, the project steered away from purely quantitative approaches. Although some universities had a good overview of investments on the one hand, and outputs in terms of publications and funding on the other hand, it was agreed that good evaluation requires more than measuring what can be quantified. At the start, the evaluative cycle was introduced, that shows how good evaluation practice is integrated throughout the life cycle of a partnership, and that a range of methods can be used to answer evaluative questions, developing mixed-method approaches (Better Evaluation, n.d.). In line with this, the project connected to current shifts in evaluation culture, moving away from accountability towards a formative role for evaluation (Molas-Gallart et al. 2021, Dinges et al. 2020). Moreover, it was stimulated to evaluate with the partner. It meant a shift from evaluation for accountability towards evaluation as communication device and mutual learning between partners (Spaapen 2015, Joly and Matt 2022). As such, the framework and its development contribute to novel approaches to evaluation policy that challenge current practices. In fact, the process approach can be viewed as an experiment, which can inspire the development of other types of frameworks and contribute to policy learning.

6. EPILOGUE

The framework has been developed with five European universities (four in EU member states) and one Australian university. There is, without a doubt, a bias in the framework. The six universities operate in a similar realm and can all be seen as global North, despite one clearly being located in the Southern hemisphere. English is the official language in three of the countries included and in the other three it is an important, if not dominant, language in academia. Universities representing the global South, and/or universities based in countries where other languages are dominant (Spanish, Chinese), were not part. Consequently, we are currently working to make the framework suitable for more global contexts, including new partners and case studies to continue the co-creation process.

Since the start of the project, there have been global developments that affect many people and aspects of life, including universities and their partnerships. Collaboration with universities from certain regions have become contested and there have been calls and actions to support colleagues from research organisations under threat. These developments stress the importance of international collaboration and solidarity between universities as well as the careful evaluation of international strategic partnerships.

LITERATURE

Better Evaluation (n.d.) Getting started, assessed 19 January 2025. https://www.betterevaluation.org/getting-started/what-evaluation

CoARA (2022) Agreement on reforming research assessment, assessed 19 January 2025. https://coara.eu/agreement/the-agreement-full-text/

Michael Dinges, Susanne Meyer, Christoph Brodnik (2020) “Key Elements of Evaluation Frameworks for Transformative R&I Programmes in Europe.” fteval Journal for Research and Technology Policy Evaluation (51), 26-40. DOI:10.22163/fteval.2020.489

EVALUATE project (2022) Harnessing the power of evaluation to build better international strategic partnerships between universities - The EVALUATE framework and handbook. Zenodo. DOI:10.5281/zenodo.7330795

Pierre-Benoit Joly, Mireille Matt (2022) “Towards a new generation of research impact assessment approaches.” The Journal of Technology Transfer, 47 (3), 621-631. DOI:10.1007/s10961-017-9601-0

Jordi Molas-Gallart, Alejandra Boni, Sandro Giachi, Johan Schot (2021) “A formative approach to the evaluation of Transformative Innovation Policies.” Research Evaluation, 30 (4), 431–442. DOI:10.1093/reseval/rvab016

Jack B. Spaapen (2015) “A New Evaluation Culture Is Inevitable.” Org. Farming, 1 (1), 36-37. DOI:10.12924/of2015.01010036

Isabelle van Elzakker, Leonie van Drooge (2019) “The political context of research infrastructures: consequences for impact and evaluation.” fteval Journal for Research and Technology Policy Evaluation (47), 135-139. DOI:10.22163/fteval.2019.342

AUTHORS

LEONIE VAN DROOGE
ZonMw, PO Box 93 245, 2509 AE The Hague,
the Netherlands
Email: drooge@zonmw.nl
ORCID: 0000-0002-7177-4586

NIKI VERMEULEN
Science, Technology and Innovation Studies (STIS),
University of Edinburgh; Chisholm House, High School Yards
EH1 1LZ, Edinburgh, UK
Email: niki.vermeulen@ed.ac.uk
ORCID: 0000-0001-8080-855X

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e16, 1-15
DOI: 10.22163/fteval.2025.708

THE SITUATION OF R&I POLICY EVALUATION IN COST INCLUSIVENESS TARGET COUNTRIES

KLAUS SCHUCH, NINO GOGADZE, VITALII GRYGA, ATTILA HAVAS, RENATA ANNA JAKSA, OLHA KRASOVSKA, ARZU KURŞUN, SOFIA LIARTI, ERIONA SHTËMBARI, MIROSLAV ŠTEFÁNIK, LENA TSIPOURI AND ALBENA VUTSOVA
DOI: 10.22163/FTEVAL.2025.708

ABSTRACT

The paper summarises the main findings of a discussion between professional evaluators and scholars on the situation of R&I evaluations in the so-called COST Inclusiveness Target Countries. It shows that the situation depends to a large extent on the state of the respective evaluation markets in the countries and sectors concerned. Since the implementation of EU evaluation rules and practices can give the R&I evaluation sector a modernisation boost, it makes a difference how often one is confronted with European evaluation requirements and thus also a difference whether a COST Inclusiveness Target Country is a member of the EU or not. The paper addresses the consequences of a missing or fragmented evaluation market, offers some alternative approaches (and their limitations) to overcome them and, based on the identified challenges, formulates recommendations for both governments and R&I policy administrations as well as R&I policy evaluators. The aim of this opinion paper is to shed light on the often-critical situation of R&I policy evaluation in the COST Inclusive Target Countries and to identify possible solutions for improvement.

Keywords: R&I evaluation, evaluation practices, COST Inclusiveness Target Countries, evaluation market, R&I policy intelligence

1. INTRODUCTION

Evaluations in the field of research and innovation (R&I) constitute a standard instrument in many European countries to review different issues at different points of time (ex-ante, interim, terminal and ex-post) such as the design, relevance, efficiency, processes and dynamics, effectiveness, coherence, impact and sustainability of various R&I policy interventions. These interventions can be new or existing programmes, policy instruments or other monetary and non-monetary measures. Organisations such as universities, research organisations and research funding agencies, as well as systems such as the competitive research system of a country, can also form the object of R&I policy evaluations.

Evaluations are an important element of the intelligence portfolio available for reflective and evidence-informed R&I policy making. However, it should not be assumed that the established, mostly sectoral (e.g. related to social policy, public health policy, development assistance or other policy areas) or national evaluation "systems", to use an approximation term, are similar to one another. On the contrary, they differ not only in terms of scope and scale, but also in terms of customary practices and in their respective stage of development, the (legal) and practical degree of commitment (including follow-up activities of the assessment) and the underlying evaluation culture.

At the *"REvaluation '24"* conference, organised by the Austrian Platform for Research and Technology Policy Evaluation (fteval), Fraunhofer ISI, IFRIS, Joanneum Research and the COST Action PROFEEDBACK in Vienna in December 2024, a group of scholars and evaluators discussed the different evaluation practices in the so-called COST Inclusiveness Target Countries. These include candidate[1] countries for EU membership, as well as the EU Member States with less developed R&I systems. The workshop on '*R&I Evaluation Systems in COST Inclusiveness Target Countries*' was organised by PROFEEDBACK. The results of the presentations and the discussions are summarised in this communication. They were 'distilled' by the first author drawing on the contributions of the other workshop participants and endorsed by the contributors. It is not intended to provide a detailed assessment of the situation in the reviewed countries, but rather to appeal to all those responsible in the national governments and agencies devising and implementing R&I policies to pay more attention to evaluation.

1 https://european-union.europa.eu/principles-countries-history/eu-enlargement_en; accessed on 6th February, 2025.

The following countries were represented at the discussion: Albania, Bulgaria, Georgia, Greece, Hungary, Slovakia, Türkiye and Ukraine. The aim of the workshop, and hence this paper, was to develop recommendations for improving evaluation practices based on the common problems identified by evaluation practitioners in the field of research and innovation.

2. BEING IN THE EU (OR NOT) MAKES A DIFFERENCE

GOVERNANCE STANDARDS

Although the results of the discussion clearly showed that there are problems almost all countries have to struggle with, there is one fundamental difference: whether or not a country is a member of the EU. EU membership goes hand in hand with a certain degree of *'new public management'* focusing on efficiency, efficacy and impact, which includes also the accountability of public interventions (and spending) in terms of good governance practice or, rather, 'good governance requirement'. However, good governance is not a uniformly implemented standard.

FUNDING AND RELATED EVALUATION PRACTICES

In terms of evaluation practice, the use of EU funds and the utilisation of EU programmes have led to a clear push towards mandatory and more frequent evaluations, often applying similar evaluation criteria. By contrast, evaluations of purely national interventions still tend to be rather infrequent in most reviewed COST Inclusiveness Target Countries. This is, of course, also related to the number of policy interventions in the R&I area. In countries with small portfolios of support measures for research, technological development and innovation, there is generally less need for evaluation, especially if one relies on established measures (which should, however, also be subject to regular reviews). All of the COST Inclusiveness Target Countries are characterised by below-average R&D spending as a share of GDP (compared to the EU average).

The situation is even harsher for non-EU countries. Although they also benefit from a few EU programmes, these interventions are less frequent and often less comprehensive in terms of scope and scale. In addition, evaluations of national measures are rare.

Overall, this means that in the EU Member States, especially due to the prevalence of projects funded by the Recovery and Resilience Facility (RRF), the European Social Fund (ESF) and the European Regional Development Fund (ERDF), mandatory evaluations have to be undertaken regularly, while in the candidate countries this recurring regularity is lacking and evaluations demanded by EU funding are not always connected to the respective national or local governments or administrations. This also has implications for the trickle-down effect on national arenas of action and administrative competences. While EU practices come across as a modernisation and reform agenda with normative power in the EU[2], in the candidate countries it remains in essence an external matter that may be sometimes inspiring, but as a rule does not change regular administrative practices.

CHALLENGES IN NON-EU COUNTRIES

When it comes to EU expenditures on projects carried out in candidate countries, the evaluations made are sometimes seen as inappropriate and exaggerated. The efforts required to conform to EU evaluation standards seem sometimes particularly demanding by the local authorities. This is also caused by a lack of professional evaluation departments in the national or sectoral administrations, in particular of qualified evaluation administrators, and unevenly distributed knowledge about the purpose, the added-value, and the ‘rules of the game’ of evaluations. Moreover, external capacities that could carry out evaluations are difficult to identify, which increases transaction costs because often there is no national ‘evaluation market’, expressed by a lacking database of ‘certified/qualified’ evaluators, in which customers and providers of evaluations can find each other easily. Sometimes *Voluntary Organisations for Professional Evaluation* (VOPEs) can act as intermediaries, if they are known to the occasional evaluation enquirers.

2 The requirements of the RRF (Recovery and Resilience Facility) are considered to become a change maker in the EU Member States due to their focus on evaluating performance. In terms of inspiring practices, the ex-ante impact assessments of policy interventions regularly exercised at European level or the European Commission’s consideration of broader social impacts have also been mentioned.

3. THE CONSEQUENCES OF A MISSING OR FRAGMENTED EVALUATION MARKET

THE 'IDEAL' EVALUATION MARKET

An ideal evaluation market would be characterised by sufficient demand for evaluations that can be met by a sufficiently large number of evaluation providers. What is 'sufficient' can be approximated through several factors. These include, among others, that evaluation providers are actually in competition with each other, do not enter into market-distorting agreements, are economically, legally and personally independent of the respective clients and can quickly make the necessary capacity and expertise available for the various evaluation requirements in a specific policy area at a competitive market price.

CONSEQUENCES OF NON-IDEAL EVALUATION MARKETS

Unfortunately, 'ideal' evaluation markets are encountered only rarely. National evaluation markets are often too small and fragmented across different policy areas, which is why evaluation expertise from abroad is sometimes also drawn upon. However, since evaluations require a high level of contextual knowledge, e.g. about the national actors and their characteristics and relationships to each other, in a particular sectoral R&I system, the choice of international evaluation providers is not always expedient. Further, information and data sources that need to be considered are usually only available in the respective national language(s). On top of these, in some EU Member States with less developed R&I systems and especially in the candidate countries, prices determined by national living standards are insufficient to be able to pay international providers. Access to international evaluators works better where information and data sources are available in English or in a larger cross-border common language area, such as the German-speaking area in the EU (Germany, Austria, Switzerland). This is also the reason why there are more evaluation providers in the R&I sector in Austria available than the small country itself can provide as a market.[3]

3 Streicher, J., Polt, W. and Unger, M. (2020). Eine Untersuchung der Marktsituation im Bereich der FTI-Evaluierung in Österreich. fteval Journal for Research and Technology Policy Evaluation (50). pp. 72-81. ISSN 1726-6629; DOI: 10.22163/fteval.2020.472

ALTERNATIVE APPROACHES AND THEIR LIMITATIONS

Buying in foreign evaluation expertise may solve the ad hoc challenge of commissioning and getting a good evaluation, but it does not necessarily contribute to national or local capacity and competence building. Therefore, mixed evaluation teams that involve national and international evaluators could be prioritised when the national market is insufficient. This is easier for large consulting firms that are based in different countries to comply with, but they are not always sufficiently experienced in evaluating R&I activities.

Another approach sometimes is, to not source via an evaluation market but rely on in-house capacity inside the government (ministries, agencies) to carry out evaluations themselves. However, this has several disadvantages, such as a higher risk of compromised independence and underutilisation or overly narrow specialisation. It is costly to stockpile evaluation expertise without being able to guarantee sufficient demand for evaluations from the public sector. An advantage of such approach is a potentially better integration of the evaluation results in subsequent policy-design decisions.

To minimise or shift the cost pressure to some extent, a basic supply of evaluation expertise can also be built up via universities (e.g. by means of evaluation courses that should be offered on a recurring basis). Although evaluation is part of the scientific canon and policy evaluations in particular make use of empirical social and economic research methods, expertise built up in this way without sufficient practical experience runs the risk of remaining too theoretical and of not being able to provide the necessary contextual knowledge.

THE SELF-REINFORCING CYCLE OF EVALUATION PRACTICE

A final major shortcoming of a poorly developed evaluation market to be highlighted here is the lack of established good practices. Evaluation as an ad hoc business implies that the risk of procedural errors increases on both, the client and contractor side. On the client side, this can affect the identification and commissioning of external evaluation providers by making formal errors such as disregarding the relevant publicity and deadline requirements for public procurement. Furthermore, problems with the creation of meaningful and clear Terms of Reference (ToR) and the estimation of realistic price-quantity structures for the requested services are becoming more frequent. The creation of good ToR, the estimation of realistic budgets for requested services and the implementation of efficient and correct procurement are skills that must be learnt.

The problem of creating realistic and competitive price-quantity structures is also encountered on the provider side of evaluations. This is due to inexperience with the evaluation object and the context in which an evaluation takes place or misalignment with the client's expectations (especially in the case of unclear ToR). The design of the requested products itself, i.e. the format and focus of meaningful evaluation reports and interim presentations, also differs from conventional scientific work. Here too, tacit knowledge that is accumulated in the course of practice is vital.

Both clients and providers also face shared uncertainties, such as expectations dealing with critical points and ensuring the use and usability of evaluation results. Ethical issues, in particular, can quickly become a problem in underdeveloped evaluation markets if appropriate agreements and standards of good evaluation practice are not already in place and known and need to be negotiated from scratch. Evaluation practice creates an evaluation culture, which in turn reflects back on evaluation practice. This can become a vicious or a virtuous cycle.

4. MAIN CHALLENGES

Insufficient evaluation markets in the field of R&I policy evaluation affect all countries. As mentioned above, being part of the EU or not makes an initial profound difference. But other factors, such as the size of a country and the importance of R&I as a policy field in the national or regional system, play a role too. All the countries, presented at the workshop, are at different levels. Nevertheless, problem areas were identified that played a role for most of the evaluators from the participating countries. These include, for example, the following:

STRUCTURAL CHALLENGES

- **Underfinancing and lack of regular evaluations:** Lack of financing and infrequent evaluation tenders can cause underdeveloped evaluation markets and a lack of available qualified evaluation experts. Moreover, the price specifications for evaluation contracts are often too low, which leads to under-budgeting. This is often due to a lack of a clearly defined range of services, a lack of understanding on the part of clients about methodological efforts and the challenges of data collection, which is why price-quantity frameworks are often significantly underestimated.

- **Conceptualisation and enforcement:** In many countries, strategies, concepts, programmes and even laws developed for certain interventions in the field of R&I are significantly more ambitious than their actual implementation and enforcement. This gap between aspiration and realisation is often a consequence of inadequately secured budgets and lack of human resources that quickly render the best intentions obsolete, while political changes can further disrupt implementation. Therefore, even the measures that sound best on paper must be critically scrutinised through evaluations ('reality check').

- **Data availability and accessibility:** If evaluations can make use of existing data, this is a golden opportunity that must be seized. In particular, monitoring data and longitudinal data on specific R&I policy intentions are of great advantage in this regard. Unfortunately, this is not often the case. Access to administrative data should be ensured in any case.

EVALUATION PROCESS CHALLENGES AND POLITICAL BARRIERS

- **Planning:** Evaluations should fit into the policy cycle, which, however, is often not the case. Thus, their normative power is limited and the results rarely used.

- **Political pressure:** In some countries, constructive feedback is confused with politically motivated criticism, which puts evaluators under stress. However, the only concern of evaluators should be to produce a good and useful evaluation and to act otherwise free of vested interests (including their own) and political influence.

- **Neglecting evaluation results:** Evaluations are a central element of 'policy intelligence'. Failure to use the results of good evaluations is a failure of policy. Follow-up steps and feedback loops for redesigning funding instruments are sometimes missing.

5. RECOMMENDATIONS

The participants of the Workshop call on those responsible in politics and the administration to improve evaluation practices. The following recommendations are offered to them at the end.

RECOMMENDATIONS FOR GOVERNMENTS AND ADMINISTRATIONS OF R&I POLICIES

INSTITUTIONAL FRAMEWORK AND DEVELOPMENT

1. Evaluation should be a central, mandatory component of any substantial policy intervention. This should be enshrined in law or at least have the character of administrative instructions, which can be included in call regulations. Sometimes, however, the problem is not the legal obligation, but enforcement. More frequent evaluation requirements help developing a competent national supplier market. But ensure competition and avoid a dominant provider, otherwise the market runs danger to become disrupted in the mid- to long-term.

2. The establishment of a functional evaluation practice and culture needs time.

3. Nominate one person in each of the ministries and funding agencies responsible for R&I policy making and R&I delivery to drive forward and centrally manage the evaluation agendas for the R&I activities. This person should also be the internal contact person for other colleagues when evaluation issues arise.

4. At a general policy level, establish jointly elaborated evaluation guidelines that apply to several fields of policy (and across departments). Of course, these must be adapted and supplemented for each ministry. Especially with regard to terminology, tendering procedures, assessment standards for the offers received, ethics, transparency, processing and accountability, a jointly elaborated guiding framework helps reducing uncertainties and transaction costs. As far as terminology is concerned, it is best to use the OECD Nomenclature as a starting point.[4]

4 OECD (2023), Glossary of Key Terms in Evaluation and Results Based Management for Sustainable Development (Second edition), OECD Publishing, Paris .https://doi.org/10.1787/632da462-en-fr

INTEGRATING EVALUATION INTO POLICY CYCLES

5. Every substantial R&I policy measure should be evaluated in terms of its relevance, efficiency, effectiveness, coherence, impact and sustainability. This applies to R&I programmes, R&I instruments, R&I organisations, agencies, systems (such as a country's entire research funding system), regulatory frameworks, etc. Start by addressing institutional weaknesses, because these usually consume a large amount of public spending and focus on organisations that can have an impact and could act as agents of change for the whole system, such as persistent innovators.

6. Evaluations are not only scientific work but require also judgment. Thus, take them seriously. Good evaluations come at a price, while under-funded evaluation projects generate inferior results and are therefore unsuitable for political legitimisation and guidance. A specific evaluation project is only being carried out once by the commissioned team and it should be provided with the best possible working conditions. This includes not only a sufficient budget, but also sufficient time and access to data to do the job well, as well as the opportunity for consultation with the responsible persons in the commissioning authority regarding any professional questions that may arise.

7. Make use of evaluation results. Integrate evaluations into the policy cycles so that their results are available when needed. Timely preparation of relevant calls for tenders for external evaluation is crucial in this regard, so that evaluations can start in good time and also work long enough to deliver useful results. An evaluation does not necessarily have to end with the delivery or acceptance of the evaluation report. Continue to use the knowledge gained by the evaluators, formally or informally. Clients should provide so-called *'management responses'* to the recommendations presented by the evaluators, stating whether and how they intend to proceed with the recommendation.

8. Create transparency. This applies to both, planned or tendered evaluation procedures and the results of evaluations. Publishing evaluation reports generally strengthens credibility and fosters dialogue within the affected community, but it also forces evaluators to deliver higher quality evaluations and evaluation reports, as the public nature of the process ensures accountability and minimizes the risk of embarrassment.

9. Gradually venture into more systemic evaluations or portfolio evaluations to avoid losing sight of the big picture when evaluating individual measures only.

EMPIRICAL EVIDENCE IS NEEDED

10. For the most important public R&I interventions monitoring systems should be established, which regularly collect the requested information[5], such as number of students or number of staff, number and type of publications and patents etc. disaggregated by meaningful categories like fields of research, organisations, departments etc. or number of granted projects disaggregated by fields of research and so on. Monitoring data are important sources for R&I policy evaluations to build on, but they do not replace evaluations.

11. When creating price-quantity frameworks, both clients and evaluators should be aware that empirical research incurs costs. Sometimes, however, relevant data are actually collected by government departments (quite possibly in the course of work not directly related to the policy intervention under review). Such databases or repositories should be made available to evaluators for aggregated analyses, subject to confidentiality clauses. Sometimes databases are also commercially curated (e.g. company data, publication data). The corresponding costs for using these databases must be budgeted for. Already existing licences, e.g. in ministries or agencies, should be made available to the evaluators for the time of the evaluation and the evaluation purpose. If necessary, specific adjustments in official statistics regulation should be made to ensure the availability of disaggregated data, if this is not prohibited for other more relevant reasons (such as martial law in Ukraine).

12. Overall, the value of having functional science, research and innovation statistics that are based on OECD and EUROSTAT standards should not be underestimated. They provide valuable data for capturing national, sectoral or regional research and/or innovation systems, even though they are usually too aggregated for specific R&I policy evaluation purposes. To obtain additional and more specific data, governments should also endeavour to participate in international or European surveys, such as the European Innovation Survey or SheFigures or use

5 An inspiring example is TUBITAK's new grant management platform that uses the advantages and functionalities of advancing digitalisation and can be used for monitoring purposes.

international support offers such as the Policy Support Facility of DG Research and Innovation for more ambitious R&I policy reviews.

CAPACITY BUILDING AND TRAINING

13. Utilise the evaluation requirements arising from EU interventions and programmes (e.g. RRF) in order to gradually generate learning effects for national evaluation practices. Do not see this primarily as an additional burden, but as an opportunity for your own reform efforts and the development of national evaluation expertise. There are already many useful guidelines and training programmes, which are also available in English.

14. A possible avenue for building up national evaluation capacity is to create a database of evaluators with practice in the evaluation of international programmes. Their registration should be on a voluntary basis, but the experience of these experts, especially those with long-term practice, is a good basis for building a national pool of experts, profiled by thematic areas.

15. To counteract the lack of suitable evaluation, personnel in the long term, evaluations as a subfield of empirical economics and social science research should be more strongly promoted in academic education or in specific trainings[6]. In addition, foreign evaluation providers should be encouraged to include local staff in their teams (local content policy). Creativity in developing public-private partnerships that link the academic and consulting sectors is called for.

RECOMMENDATIONS FOR EVALUATORS

16. Evaluation is neither rocket science, nor an easy scientific endeavour, because it requires profound context-related knowledge as well as robust methodological and social skills. As regards advanced evaluation, global networks such as OECD or UNESCO and certain academic organisations provide pertinent materials and tools to learn from (e.g. bibliometrics, patent analysis). In addition, being well versed in the methods of empirical social sciences and economic research is essential. Additional support

6 To give an example: https://www.mioir.manchester.ac.uk/study/short-courses/evaluation-of-science-and-innovation-policies/; accessed on 10 February 2025.

and knowledge sharing of best evaluation practices can also be used within specialized networks and platforms.[7]

17. Choose your performance indicators wisely. Sometimes R&I policy interventions expect too much and overburden the performance or overestimate the incentive mechanisms of their measures. Occasionally, they are simply poorly designed. Unfortunately, the evaluation questions often have excessive expectations, which is why it seems useful to supplement them with more finely grained or alternative indicators. Additionally, exploring alternative data sources, such as mining websites, can be more effective than of solely relying on self-declared data from funded projects.

18. On the part of the evaluators, it is recommended to expand the occasionally dominant self-image of their role as critical assessors and to stronger promote, in addition to accountability aspects, learning and steering effects. However, this does not mean that evaluations should be uncritical. Particularly in smaller countries with few evaluators, there is a risk of conformism or of being co-opted by the commissioning bodies to obtain favourable evaluation results. To avoid possible dependencies and to allow fresh perspectives, it can make sense to involve colleagues from abroad who are unaffected by local networks and cliques.

19. Try to express your evaluation results, including the recommendations made, as clearly as possible in order to be able to guide action. Avoid vague and superficial statements or recommendations that may not even address anyone in particular.

20. Explore opportunities beyond government contracts, if this is possible. Take advantage of the – admittedly not always numerous – demand for evaluations from the private or civil society sectors as well. This will enable you to apply and develop your methodological skills, giving you a better sense of what learning from evaluation or steering through evaluation means. Additionally, it will provide insights into the practical usability of evaluation and how you can contribute to this process.

7 Examples: Austrian Platform for Research and Technology Policy Evaluation (https://fteval.at/en/; accessed on 10 February 2025). It runs R&I policy evaluation conferences held every three years. Another example is DeGEval (https://www.degeval.org/en/working-groups/research-technology-and-innovation-policy/; accessed on 10 February 2025), which runs a dedicated working group for research, technology and innovation policy.

21. Insist on being allowed to publish your evaluation reports. It benefits your professional portfolio and CV. More importantly, it fosters an evidence-informed dialogue on science, research and innovation policy, contributing to the advancement of the field.

AUTHORS

KLAUS SCHUCH
ZSI – Centre for Social Innovation
ORCID: 0000-0002-7603-4384

NINO GOGADZE
Centre for Science and Technology Studies (CWTS), Leiden University
ORCID: 0009-0006-4498-9481

VITALII GRYGA
Institute for economics and forecasting of the NAS of Ukraine
ORCID: 0000-0003-0026-3591

ATTILA HAVAS
Institute of Economics, Hungarian Academy of Sciences Centre of Excellence
ORCID: 0000-0002-8571-8817

RENATA ANNA JAKSA
HÉTFA Research Institute
ORCID: 0009-0007-9299-5402

OLHA KRASOVSKA
Ukrainian Evaluation association
ORCID: 0009-0006-3230-684X

ARZU KURŞUN
Giresun University
ORCID: 0000-0001-9560-7689

SOFIA LIARTI
OPIX P.C.
ORCID: 0000-0002-2525-6411

ERIONA SHTËMBARI
University of New York Tirana
ORCID: 0000-0003-1390-8301

MIROSLAV ŠTEFÁNIK
Institute of Economic Research, Slovak Academy of Sciences
ORCID: 0000-0002-9993-7885

LENA TSIPOURI
National and Kapodistrian University of Athens & OPIX P.C.
ORCID: 0000-0002-7454-6271

ALBENA VUTSOVA
Sofia University “ St. Kliment Ohridski”
ORCID : 0000-0001-8223-6727

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e17, 1-18
DOI: 10.22163/fteval.2025.709

EVALUATION OF R&D AND INNOVATION POLICY IN UKRAINE: MISSING ELEMENTS

VITALII GRYGA, OLHA KRASOVSKA AND YULIIA RYZHKOVA
DOI: 10.22163/FTEVAL.2025.709

ABSTRACT

The paper discusses a comprehensive approach suggested for the evaluation of the research, development and innovation (RDI) domain in Ukraine aiming to design evidence-based policy making. It is built around four main elements, namely the evaluation of research and development (R&D), the evaluation of innovation performance, performance the assessment of research institutions, and policy evaluations. The novelty of the approach lies in considering the complexity of the evaluation of the RDI domain through the prism of its elements. This paper analyses existing evaluation approaches for each element trying to identify the 'missing elements' needed for evidence-based RDI policy in Ukraine.

It is suggested to approach the evaluation of R&D performance using a model that examines the long-term correlation between the dynamics of scientific personnel and the science intensity of GDP. In turn, the authors consider composite indices as a proper way to analyse innovation performance, despite the controversial issues described in the paper.

The paper also highlights the absence of the unified approach to the assessment of research institutions' performance in Ukraine, despite the attempt to unify the assessment which recently has been undertaken. The paper argues that it's too early to assess the relevance of the approach.

The conducted analysis leads to the conclusion that Ukraine demonstrates good potential for ensuring only two of the four elements of the complex evaluation of the RDI domain, namely evaluation of R&D performance and institutions, while evaluation of innovation performance and policy evaluations are lagging. Policy evaluations remain the weakest element of the complex evaluation system of RDI policy predominantly because of the absence of an independent evaluation culture.

Keywords: R&D and innovation policy, innovation performance, policy evaluations, assessment of research institutes, evaluation framework

INTRODUCTION

'Evidence-based' policymaking has become not only a common trend but rather a 'must have' for ensuring effective policy design (Newman et al, 2016). Meanwhile, «most academic research on public policy achieves little influence in government» (Mead, 2015), raising questions about the sources of evidence that fuel policy-making processes. Extensive research exists evaluating RDI policies and instruments. Many studies address financial instruments, particularly subsidies and taxes (Negassi and Sattin, 2019; Czarnitzki and Lopes Bento, 2010; Shim and Shin, 2022), using different econometric techniques. A more general approach to the evaluation of R&D policy was proposed by E. Arnold (2004), which focused on different levels of the system. Transitioning from the theory to practice, there is a comprehensive document outlining an evidence framework on monitoring and evaluation of the EU's research and innovation programme. It addresses nine key impact pathways, emphasizing more on impact rather than merely tracking inputs and outputs (EC, 2023), thus placing impact evaluations at the core of evidence-based policymaking.

Ukraine inherited a well-developed R&D system but has been unable to economically benefit from it. The country's R&D potential has been deteriorating for a prolonged period. However, recent reforms in RDI domain included updating the legislation, deeper EU integration, the establishment of a new R&D funding body based on international experience. An update of the legislation framework in 2015-2016 offered some hope for improving the situation in science and innovation. Nevertheless, these changes did not increase either the demand for regular RDI-policy evaluation or policymakers' awareness of its necessity.

At the same time, the active roles of civil society and international organisations – particularly the European Union (EU) and its member states – in supporting Ukrainian reforms have highlighted the need to ensure evidence-based policymaking across various domains, including RDI.

For example, the new law on official statistics, adopted in 2023, explicitly suggests that government authorities utilize the produced data for decision-making. In 2016-2017, upon request from the Ministry of Education and Science of Ukraine (MESU), the European Commission conducted a peer review of the Ukrainian research and innovation system. The review emphasized the necessity of setting up "...a system to continuously monitor the development of STI policies in Ukraine and introduce a full-fledged evaluation culture and system for research and innovation policy" (EC, 2017).

Bringing evidence to the core of policymaking process becomes even more critical during economic crises, when public fundings are limited. Although the full-scale war in Ukraine is ongoing, discussions on the different modalities for the recovery process have already become quite active. A series of recovery conferences held after the invasion brought to the agenda the necessity to have a clear vision of multi-optional recovery strategies. The Ukraine Facility Plan, which serves as the basis for the implementation of the EU's financial support programme for Ukraine in 2024-2027, replaced the Ukraine Recovery Plan, previously characterized by loosely connected ideas and proposals rather than constituting a coherent strategic document. Therefore, the authors believe this is a crucial moment to raise awareness among policy makers and policy implementers as well as civil society regarding the importance of evidence-based policymaking and to conceptualize and RDI evaluation framework by revising previous experience through contemporary evaluation approaches.

METHOD /APPROACH

In this study, we rely on Arnold's approach to R&D policy evaluation which should have three levels (Arnold, 2004):

- traditional evaluation of individual interventions, such as programmes;
- assessment of the overall 'health' of RDI systems (including performance, connectivity, capabilities);
- subsystems evaluation or 'bottleneck analysis', which explores the systems role of institutions, classes of actors and clusters. It is based on the results from previous levels and proposes performance improvements of RDI system parts.

Based on the described approach, we explore the system of RDI evaluations in Ukraine, which is supposed to serve as the core of a comprehensive evidence-based policymaking process in the mentioned domain. For this purpose, we consider four primary elements of a comprehensive evaluation system in RDI domain (see Figure 1):

1. Evaluation and assessment of research institutions and researchers.
2. Evaluation of R&D performance.
3. Evaluation of innovation performance.
4. Policy evaluations.

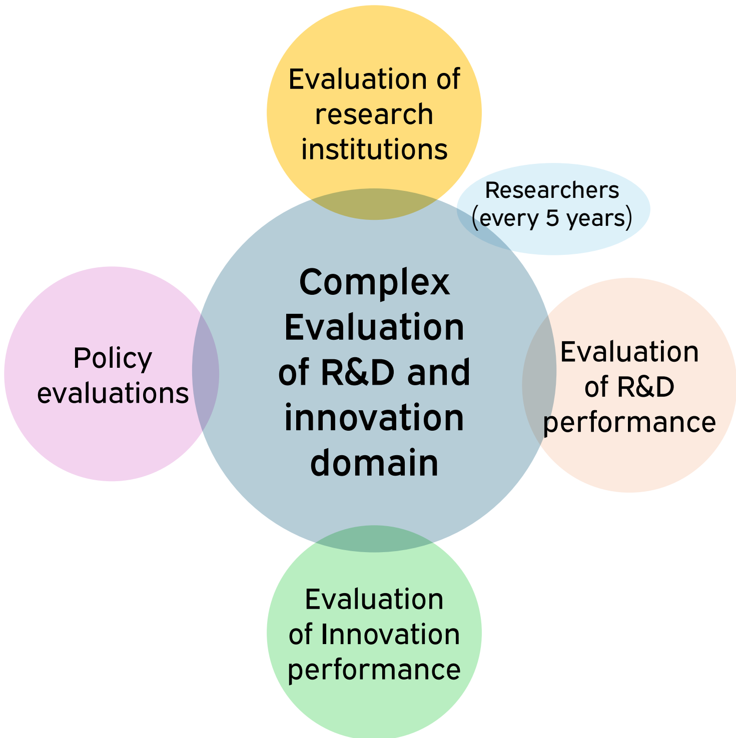


Figure 1. Building elements of the complex evaluation system in R&D and innovation domains
Source: developed by authors

Through a combination of various methods (desk research and secondary data analysis, including statistical analysis and econometric techniques), we investigated current evaluation practices in each component to identify the 'missing elements' necessary for evidence-based RDI policy in Ukraine. The data includes official statistics on RDI and economic development, official legislation, regarding the evaluation of research institutes and researchers as well as analytical reports, produced by international experts within the UNECE, the EU Policy Support Facility and other initiatives.

The study is based on the assumption that properly evaluation all aforementioned components which are essential for effectively assessing a complex domain such as RDI. Consequently, this contributes evidence-based policy, which in turn drives sustainable economic development.

A BRIEF OVERVIEW OF THE UKRAINIAN RDI SYSTEM

Since its independence, Ukraine has inherited a Soviet-type RDI ecosystem characterised by a high R&D intensity of about 2% of GDP and approximately 450,000 research-oriented personnel. However, the prolonged economic transformation, the disruption of previous ties with Soviet partners and inefficient governance and policymaking have led to the deterioration of the R&D sector. Figure 2 presents the long-term dynamics of the research intensity of GDP and the share of the labour force involved in R&D (per 1000 of employees aged 15-70). It shows that even decades, after the shock following the collapse of the Soviet Union, the R&D sector is still struggling and shrinking. Currently, R&D funding is nearly seven times lower (0.33% of GDP) than the EU average and the number of R&D personnel is four times smaller.

The low level of RDI funding is attributed to persistent budget constraints and the prevailing attitude of the government, especially the Ministry of Finance of Ukraine, which has traditionally viewed the RDI sector as a fiscal burden. They have consistently demanded evidence of the value added generated by science, including its economic or social impact. Formal compliance of public R&D funding with fixed in budget plans indicators has not been sufficient for the government to consider increasing R&D funding. At the same time, no substantial evaluations at medium or large scale have been carried out for RDI programmes, strategies, or policy instruments.

However, a few nationwide evaluation-like exercises, supported by international institutions upon request from the MESU, are worth mentioning. The first is the UNECE Innovation performance, review of Ukraine (UNECE, 2013). It provides an examination of the RDI system, the institutional framework for innovation policy and the various mechanisms and instruments of public support for innovation in the country, along with valuable policy recommendations. However, the study did not evaluate the performance or impact of existing policies. The second exercise is the EU Horizon 2020 Policy Support Facility Peer Review of the Ukrainian Research and Innovation System (EC, 2017). EU experts developed 30 recommendations to raise the quality and relevance of the science base. Most of them were considered by MESU, gradually apart from those associated with an additional increase in public spending on RDI.

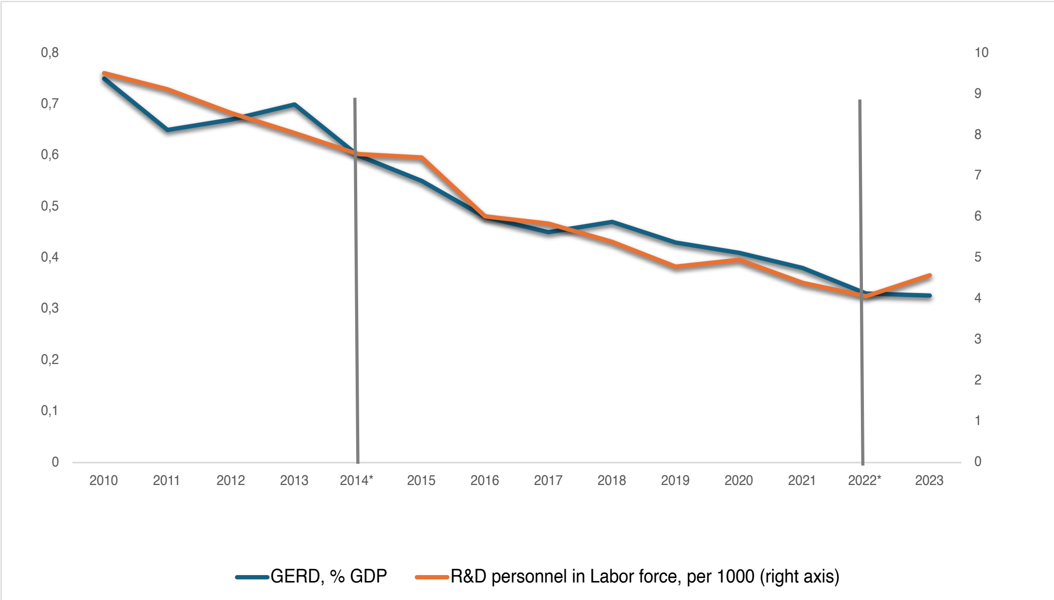


Figure 2. Main indicators of the R&D system dynamic of Ukraine
*) punctured lines reflects Russian invasions and military actions.
Source: compiled by authors, based on data from Ukrainian Statistic Service (Ukrstat).

Currently, the RDI system of Ukraine consists of diverse players, including the National Academy of Sciences of Ukraine (NASU), sectoral academies of sciences, higher education institutions (both public and private universities), other R&D institutions and private research-oriented companies (see Annex I). To build stronger links between science and business, initiatives such as science parks, startup incubators, acceleration programmes and innovation-oriented educational programmes have also been established. The main channels of public R&D funding are the MESU, NASU and the National Research Foundation of Ukraine. An important source of RDI funding is foreign sector, particularly European programmes and initiatives, notably Horizon Europe and its predecessors. Innovations are supported through the Ukrainian Startup Fund, the State Finance Institution for Innovations, etc.

EVALUATION AND ASSESSMENT OF RESEARCH INSTITUTIONS

Currently, there is no unified approach for evaluating and assessing research institutions in Ukraine. The assessment of the effectiveness of research institutions conducted by the NASU uses its methodology updated in 2023 (NASU, 2023). The methodology draws on evaluation criteria and procedures used in the evaluation of scientific institutions in countries such as Germany, Austria, the United Kingdom, Poland and the Czech Republic. Based on the evaluation results, research institutes are assigned to categories that correspond to recommended future actions—such as receiving a 'green light' to continue operations, enhancing international cooperation, undergoing reorganisation or being closed.

In parallel, the MESU has its own methodology for the state certification of scientific entities. In 2024, the Ministry developed a new approach to evaluate the R&D effectiveness of research institutions and universities. As a result, all public research-performing organisations are scheduled to be evaluated in 2025. MESU introduced new criteria that consider research contributions to global science, economic growth, national defence, and the overall benefit to Ukrainian society (impact assessment), alongside compliance with open science principles. It is planned that the evaluations will be conducted simultaneously in all research and higher education institutions within specific scientific fields. Based on the results of the assessment, research-performing organisations will be assigned to one of four categories: A) world-class and leading positions in their scientific area, B) high quality research, C) satisfactory research performance, lacking an active international profile, D) low research performance: the institution fails to meet state certification standards. MESU reserves the right to make R&D budgetary decisions, based on this categorisation.

To ensure transparency and efficiency, the National Electronic Scientific Information System (URIS) supports the evaluation process through its suite of digital tools. URIS is a multifunctional IT system that provides the collection, formation, processing, storage, and use of data and information in the field of scientific and science and technology (S&T) activity of Ukraine. The system was created to combine information on the results of scientific research, the activities of research institutions and higher education institutions (HEIs), as well as Ukrainian researchers[1]. In the future, it is expected that URIS will be used to provide evidence for decision-making, ensuring accessibility of data from the Ukrainian science system, including research data and information, equipment, services and resources for research, grant management (application and reporting), etc. URIS will serve as a Current Research Information System, a modern tool for managing scientific data and making strategic decisions in the field of science.

1 https://dntb.gov.ua/completed-projects/urisinfo

At this point, it is premature to evaluate the relevance of the new methodology or to determine the consistency of its application across research institutions. However, several drawbacks have already been identified. First, the approach requires institutions to manually insert information that could be automatically retrieved - for example, journal quartiles or publication titles via DOI2 - resulting in unnecessary additional effort and time for the staff. Second, the list of accepted evidence for impact is limited to only three items, whereas research organisations typically have a broader range of documents to demonstrate the impact of their R&D activities. Third, the methodology has not yet been tested on real-world cases, and not all indicators are sufficiently justified. For example, many indicators are based on formulas that disproportionately favour PhD students and university researchers, thereby giving HEIs an unjustified advantage over other public research organisations (PROs), which typically employ fewer staff:

$$\mathbf{(I)} = \frac{\text{scientific output}}{\textit{Researchers} + \mathbf{0.5} \times \text{Scientific and pedagogical personnel} + \mathbf{0.1} \times \textbf{PhD students} + \textit{Doctoral sudents}}$$

In addition to the evaluation of research institutions, the government requires PROs and HEIs to perform examination (evaluation) of individual researchers every three to five years. The duration depends on the outcome of the previous assessment: researchers with strong performance are granted five years, while those with weaker results receive a three-year period. The data researchers provide during the evaluation process include a list of publications, information on participation in R&D projects, national and international cooperation, a description of scientific results, and other research-related activities. In fact, information for the evaluation of R&D institutions and researchers is similar, but the data formats differ significantly, which creates additional pressure on researchers. As research institutions typically gather data from researchers on an annual basis, conducting individual evaluations adds minimal value. Evaluation results can affect personal careers, but they may also help PROs/HEIs to improve their performance.

In order to consider the assessment of research institutes as a relevant element of evidence-based policymaking we suggest eliminating the outlined issues, continuing the alignment of methodologies for the assessment PROs and HEIs, and developing policy options with a funding mechanism for each R&D performing category.

2 Digital object identifier

EVALUATION OF R&D PERFORMANCE

The Ukrainian statistical office produces statistical data on S&T and innovation development, based on OECD manuals and Eurostat methodology. In contrast to the EU, many indicators in Ukraine have been subject to frequent changes over the past decades, hindering the development of consistent long-term datasets for in-depth analysis. Unfortunately, the government of Ukraine has not paid sufficient attention to the RDI, and as a result, no framework for RDI performance evaluation has been developed. Instead, the government predominantly depends on technical assistance from the EU, including the 2016 Peer Review of the Ukrainian Research and Innovation System (EC, 2017) and ongoing support for Ukraine in research infrastructure policies (2025), among other initiatives.

Analytical reports, produced by organisations affiliated with the MESU are limited due to their primarily descriptive nature and focus on budget expenditures. As a result, they do not reveal how RDI indicators are interconnected, whether policy instruments have affected the health of the RDI system and other critical insights.

Following the approach of Saltelli and Giampietro (2017), authors conclude that the practice of R&D evaluation in Ukraine lacks quantitative storytelling. Therefore, we suggest a mixed-method approach for the evaluation of R&D performance, using both quantitative and qualitative methods. The quantitative analysis should go beyond descriptive data and be enriched with econometric techniques to identify the strength of the links between the indicators. For example, we investigated the correlation of the dynamics of scientific personnel with the research intensity of GDP in the long-term perspective. The panel data, based on a heterogeneous sample of countries, confirmed the existence of a direct proportionate relationship between the indicators. Our research proves the validity of the proposed model with a high degree of statistical significance.

In addition, the panel data analysis reveals different patterns of S&T development across countries. Some demonstrate low elasticity of research personnel intensity, while in others the elasticity is considerably higher. Importantly, elasticity is not constant and varies according to the level of S&T development and each country’s policy approach to science, technology and innovation development.

In the case of Ukraine, the relationship between R&D funding and research personnel intensity is relatively straightforward over the period 2010–2020 (see Figure 3). Moreover, between 2016-2020, the decline of R&D funding triggered an even stronger response in the reduction of research staff than in the previous period, emphasizing systemic failures of Ukrainian R&D policy.

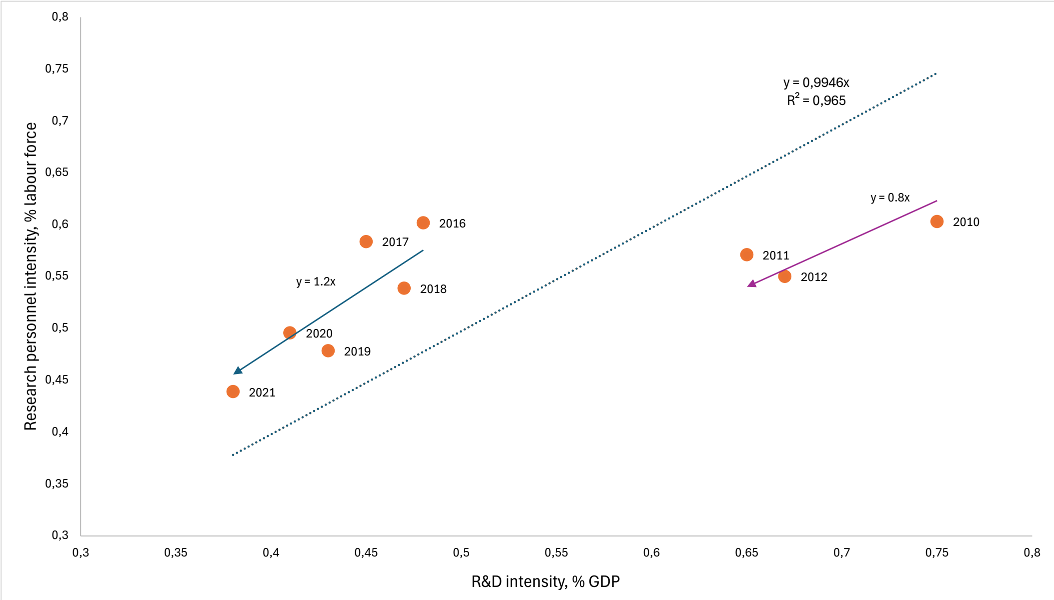


Figure 3. Comparison of indicators of GDP science intensity and the workforce in Ukraine.
Source: Authors calculations based on data from the State Statistics Service.
URL: https://ukrstat.gov.ua/

The practical value of the conducted analysis lies in the justification of target indicators to be considered for the development of strategic documents and key targets. The model allowed us to calculate the necessary level of R&D funding in Ukraine by 2030 to achieve at least 40% of the EU-27 level of 2021. According to our calculations, to reach this goal, Ukraine needs to increase R&D funding by at least EUR 200 million by 2030.

EVALUATION OF INNOVATION PERFORMANCE

The evaluation of innovation performance was not a priority for the government of Ukraine, although some data were produced by the national statistical office. The lack of interest in assessing innovation performance was evident in the fact that no dedicated public funding for innovation was provided for nearly two decades. The State Innovation Fund, established in 1992 to distribute innovation grants and soft loans, lost credibility, due to opaque and allegedly politicised award decisions. After most of its core functions were suspended in 2000, the fund was nominally replaced by the State Innovation

Financial-Credit Institution. However, for years, the new institution received only symbolic budget allocations and did not launch any competitive funding programmes.

As a result, for almost two decades, Ukraine lacked effective national innovation policy instruments - grants, tax credits or co-investment schemes - to support innovative firms. This further undermined the incentive to track innovation outcomes systematically. The situation is set to change with the creation of the Innovation Development Fund, also known as the Ukrainian Startup Fund, which provides grants for innovative start-ups.

Ukraine applies the EU's Community Innovation Survey (CIS) methodology to gather data on innovation activity that are comparable to EU standards. In parallel, a national methodology is used to assess innovation performance in the industrial sector exclusively. The data are shared with international statistical institutions, which use them to calculate various innovation indices, e.g. the Global Innovation Index, the European Innovation Scoreboard, as well as in reports like the UNECE Innovation Performance Review of Ukraine (UNECE, 2013). Meanwhile, domestic demand for innovation statistics comes primarily from the researchers themselves (e.g. Zhernovyi, 2024) and partially from regional authorities who need to deal with smart specialisation, namely, to identify regional strengths in innovation activity. Smart specialisation was introduced into regional policymaking in 2019. Currently, Ukrainian regions are updating their strategic and operational objectives. However, the update is expected to reflect the economic impact of the war, rather than assessing the effectiveness of smart specialisation implementation and its role in driving innovative transformation.

Relying on composite indices for the evaluation of innovation performance does not appear sufficient to gain a comprehensive understanding. The relevance of this approach has sparked debate among scholars, given that innovations are inherently unpredictable and often depend on the interactions and relationships between stakeholders (Granger, 2020). In addition, composite indices are rather 'static', meaning they do not consider the innovation process per se, which changes over time. This limitation is particularly problematic in rapidly changing environments, where policy must respond promptly. For countries at an early stage of developing RDI evaluation systems, it is crucial to take into account the general limitations of composite indices, as outlined in Nardo et al. (2005).

The comparability of indicators remains a significant challenge. Although the indicators themselves are designed for comparison, the data and procedures used for their collection and interpretation differ across countries and are not standardised across all fields of science, technology, and research. One example of this is Cyprus. According to Eurostat, innovation activity in Cyprus was reported over 65% in the CIS 2018 and 2020, but it suddenly dropped to 40% in the CIS 2022. Similar statistical inconsistencies are observed in Ukraine: after a modest rise to 28% in the CIS 2018, innovation activity dropped to less than 9% in the CIS 2020. These drastic changes appear to have been driven by several factors that warrant thorough investigation. Without proper interpretation, innovation data can be easily misread by policymakers, potentially leading to flawed policy decisions.

Furthermore, these indices rarely capture the relative importance of individual factors, the relevance of input data, the causal relationship between input and output or the frameworks and conditions under which innovation emerges. The link between investments and results is particularly unclear and underresearched: investments in innovation cannot easily traced to specific outcomes and their attributability diminishes over time. Such as indicators fail to reflect the time lag between investments in innovative activities input and their eventual output. This time lag is not only undefined but also likely to vary across different types of innovative activity.

Despite the limitations and even though indicators can, at best, only identify strengths and weaknesses rather than explain them, composite indices offer a broad overview of a country's innovation system and may therefore be considered a useful tool for evaluating innovation performance over time. However, from a short-term perspective, countries with underdeveloped innovation ecosystems often require alternative evaluation methods, such as targeted surveys, to track progress and enable timely interventions at early stages.

POLICY EVALUATIONS

Today, there is no explicit strategy for the development of science in Ukraine. The attempt by MESU to develop a National Strategy for Education and Science was unsuccessful in 2023, partly due to a lack of institutional capacity to reconcile and align the hundreds of ideas and measures proposed by more than 1,700 experts. Meanwhile, the Ukrainian government has shown greater willingness to approve documents, associated with the European integration process, such as the National Plan for Open Science, which was adopted in 2022.

There were also other documents related to science and innovation. In 2023, MESU updated the Roadmap for EU Integration in Education and Science, with a focus on aligning Ukrainian legislation with the EU acquis and expanding Ukraine's participation in EU programmes. However, the fragmented approach to policymaking continues to undermine the coherence and effectiveness of RDI policy evaluations in Ukraine. An analysis of this domain highlights that Ukraine yet develops a common framework or set of guidelines for conducting policy evaluations. Some policy areas such as culture and regional development do include explicit legal provisions for evaluation, including basic modalities and provisions for external evaluations. Although these are not enforced, the existence of such legislative framework at least provides a potential foundation for the future introduction of policy evaluations. In contrast, in RDI domain, the legislative framework for evaluation is rather limited. Several strategic documents (e.g. Strategy of Innovation Development till 2030[3], Strategy for Digital Development of Innovation Activity till 2030[4]) contain target indicators alongside policy measures. However, these documents do not include provisions for independent evaluation and instead envisage a simplified form of accountability rather than rigorous policy evaluation process.

The analysis of key legislation revealed the following shortcomings in the governance of research and innovation:

- a misalignment between outcomes and indicators, the strategies' tasks and measures, thereby undermining the intervention logic;
- poor coordination of policy documents in the RDI domain;
- low enforcement and implementation of the policy documents, and
- permanent underfinancing of the policy measures' implementation.

UNECE experts studied Ukrainian innovation policy during the COVID time and reached similar conclusions. According to them, poor coordination and complementarities with small and medium enterprises development and industrial policies, inadequate institutional and legal frameworks, and a miscoordination at the central government level are the weakest point of innovation policy of Ukraine (UNECE, 2020).

A key requirement for ensuring evidence-based policymaking is, to initiate a new policy cycle only after a thorough assessment of the effectiveness and lessons learned from the previous one. It is worth noting that this represents a common challenge in the Ukrainian policymaking context. While most policy documents include indicators to monitor implementation, they often

3 https://zakon.rada.gov.ua/laws/show/526-2019-%D1%80#Text (in Ukrainian)

4 https://zakon.rada.gov.ua/laws/show/1351-2024-%D1%80#Text (in Ukrainian)

lack a clear intervention logic or a well-defined theory of change. This issue is commonly attributed to the absence of a well-established evaluation culture, which remains in the early stages of development in Ukraine.

Accordingly, the absence of independent evaluations of the policy measures and their implementation renders policy evaluations virtually absent within the evaluation system for the RDI domain. To improve the situation, greater efforts are needed to develop a national RDI evaluation framework that incorporates both, a solid theoretical foundation and international best practices, for example, the Horizon Europe evaluation framework for RDI programmes and large-scale policy instruments.

CONCLUSIONS

In this study, we attempted to examine the RDI evaluation framework of Ukraine in order to support the development of a unified approach to evaluation and avoid the duplication of efforts by linking different elements (or layers) of the RDI system. The findings revealed significant asymmetries across the four pillars of evaluation: R&D performance, innovation performance, research institution assessment and policy evaluations.

While Ukraine demonstrates promising potential in R&D performance and institutions' evaluation, supported by methodologies that are harmonised with international and EU standards, both innovation performance evaluation and policy evaluation remain underdeveloped. Most critically, the absence of a culture of independent and regular policy evaluations hinders the integration of evidence-based decision-making into strategic planning processes.

Despite ongoing reforms of the RDI system, the full potential of evaluation efforts has yet to be realised. Ukraine must address structural challenges – including data consistency, methodological biases, and the lack of integration across evaluation components – to build a robust, evidence-driven innovation ecosystem. The introduction of new policy measures and instruments should be inseparably linked with proper *ex ante*, interim, and *ex post* evaluations.

Ukraine's experience offers valuable lessons that can be transferred to other transition countries. The key priority is to establish a balanced and comprehensive RDI evaluation framework encompassing institutional assessment, R&D performance, innovation performance, and policy evaluation, in order to mitigate evidence asymmetries and reduce the reporting burden.

Secondly, the long-term value of statistics hinges on their quality, reliability, the compatibility of indicators, and the documentation of any data discontinuities. Thirdly, composite indices should be complemented by mixed method ‘quantitative storytelling’ to capture local dynamics that static benchmarks often obscure. Additionally, embedding independent, cyclical policy evaluations and implementing an automated national research information system — rather than relying on manual data entry are crucial safeguards against an ‘evidence-poor’ policy cycle.

LIMITATIONS

The study has several limitations that readers should bear in mind. First, some findings rely on official statistics, whose definitions and collection procedures have changed repeatedly. Such breaks may distort long-term trends, most notably in the CIS-based innovation and science indicators. Second, the econometric test addresses only the bivariate link between R&D-personnel intensity and GERD to GDP, without controlling other factors, so the reported elasticities are descriptive rather than causal. Third, the conclusion, regarding the forthcoming URIS-supported institutional review and the new MESU evaluation approach remain provisional, as they have not been completed yet.

REFERENCES

Arnold E. (2004) Evaluating research and innovation policy: a systems world needs systems evaluations, Research Evaluation, Volume 13, Issue 1, April 2004, Pages 3–17, https://doi.org/10.3152/147154404781776509

Czarnitzki D., Lopes Bento C. (2010) Evaluation of public R&D policies: A cross-country comparison, ZEW Discussion Papers 10-073, ZEW - Leibniz Centre for European Economic Research.

Granger R. (2020) Innovation Explained: Inside the Blackbox. Accessed at https://www.academia.edu/44413517/INNOVATION_EXPLAINED_INSIDE_THE_BLACK_BOX

European Commission (2017). Peer review of the Ukrainian research and innovation system (KI-AX-16-008-EN-N). Publications Office of the European Union. https://projects.research-and-innovation.ec.europa.eu/sites/default/files/rio/report/KI%2520AX%252016%2520008%2520EN%2520N_UKR_0.pdf

European Commission (2023). Evidence framework for the monitoring and evaluation of Horizon Europe (SWD(2023) 132 final). https://research-and-innovation.ec.europa.eu/document/download/e78eceb1-0859-4192-9117-5bdf4b5cf594_en?filename=swd-2023-132-monitoring-evaluation-he.pdf

Mead, L.M. (2015) Only connect: Why government often ignores research. Policy Sciences 2015, 48, 257–272 https://link.springer.com/article/10.1007/s11077-015-9216-y

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2005). Handbook on constructing composite indicators: Methodology and user guide (OECD Statistics Working Papers, No. 2005/3). OECD Publishing. https://doi.org/10.1787/533411815016

NASU (2023) Methodology for evaluation of the effectiveness of scientific institutions of the National Academy of Sciences of Ukraine. URL https://files.nas.gov.ua/text/Documents/Metodika_ocinuvanya_2023.pdf

Negassi, S. and Sattin, J. (2019) Evaluation of Public R&D Policy: A Meta-Regression Analysis. Technology and Investment, 10, 1-29. doi: 10.4236/ti.2019.101001.

Newman, J., Cherney, A., Head B. W. (2016) Policy capacity and evidence-based policy in the public service. Public Management Review, 19(2), 157–174. https://doi.org/10.1080/14719037.2016.1148191

Saltelli A., Giampietro M. (2017) What is wrong with evidence based policy, and how can it be improved?, Futures, Volume 91, P.62-71, https://doi.org/10.1016/j.futures.2016.11.012.

Shim H., Shin K. (2022) Empirical Analysis of Evidence-Based Policymaking in R&D Programmes. Sustainability, 14(1), 311. https://doi.org/10.3390/su14010311

UNECE (2013) Innovation performance review of Ukraine. Retrieved from https://digitallibrary.un.org/record/751366

UNECE (2020) Sub-regional Innovation Policy Outlook 2020: Eastern Europe and the South Caucasus. Retrieved from https://unece.org/sites/default/files/2021- 06/UNECE_Sub-regional_IPO_2020_Publication.pdf

Zhernovyi D. (2024) CDM framework for R&D–innovation–productivity relationship: theoretical and empirical perspectives. Economy and Society, (68). https://doi.org/10.32782/2524-0072/2024-68-173

AUTHORS

VITALII GRYGA
Institute for Economics and Forecasting NAS of Ukraine
Email: v.gryga@gmail.com
ORCID: 0000-0003-0026-3591

OLHA KRASOVSKA
Ukrainian Evaluation Association
Email: olha.krasovska@gmail.com
ORCID: 0009-0006-3230-684X

YULIIA RYZHKOVA
Institute for Economics and Forecasting NAS of Ukraine
Email: yu.ruzhkova@gmail.com
ORCID: 0000-0001-7944-2247

Annex I. The conceptual model of Ukraine's RDI system

fteval JOURNAL
for Research and
Technology Policy Evaluation

July 2025, Vol. 57, pp. e18, 1-25
DOI: 10.22163/fteval.2025.710

INSIDE THE FUNDING PROCESS: USING GENERATIVE AI TO ASSESS REVIEWERS' CRITERIA PRIORITISATION IN MULTI-STAGE APPLICATION ASSESSMENTS

PETER KOLARZ AND DIOGO MACHADO
DOI: 10.22163/FTEVAL.2025.710

ABSTRACT

When evaluating funding schemes with multiple aims (expressed through multiple assessment criteria such as quality, novelty, relevance, collaboration, etc.), there is an inherent challenge in assessing what role different assessment criteria play in the selection and awarding process. As part of a process evaluation of the Austrian FWF's Emerging Fields programme, we used generative AI to analyse peer reviewers' reports on applications submitted to the scheme. The purpose of this analysis was to understand how various assessment criteria were being operationalised in the review process. Specifically, the Emerging Fields scheme has two separate written application assessment stages: a short outline-proposal stage, followed by a full application review stage. Background research on the scheme's design led us to a hypothesis that reviews in the first of these two stages should emphasise and reward innovative potential and novelty of the proposed project ideas, while reviews in the second stage should place a greater emphasis on scientific quality of the research plans. Our analysis of reviews at both assessment stages used generative AI to assess, first, what priority various criteria had in terms of the amounts of text devoted to each criterion. Second, the analysis assessed the extent of positive or negative sentiment with which reviews covered each criterion (specifically, the criteria covered were scientific quality, novelty, feasibility, and team qualifications). We then compared the results in all combinations between passed applications, failed applications, stage 1 applications and stage 2 applications. Results largely substantiated our hypotheses, most notably with novelty having a significantly higher priority at stage 1 and scientific quality having a significantly higher priority at stage 2. This analysis added substantial value to the Emerging Fields evaluation and has potential to become an invaluable tool to test the effectiveness and appropriateness of research funding processes, particularly in complex multi-stage designs with multiple scheme objectives.

Key words: Research funding, peer-review, generative AI, evaluation

INTRODUCTION

Research funding schemes may have several different aims. These include (but are not limited to), funding basic exploratory research, research that is societally relevant (including to specific societal challenges), research that has particular promise for industrial application, research that is highly innovative and breaks with established paradigms, research that presents new interdisciplinary or transdisciplinary perspectives, or research that seeks to foster collaboration between previously unconnected individuals.

Evaluating funding schemes with diverse objectives requires understanding of how well their processes address complex and varied assessment criteria. Especially in cases of schemes with multiple aims – and consequently, multiple assessment criteria – process evaluations of funding schemes face the challenge of having to assess to what extent application selection processes consider and reward these various aspects of proposed research projects. Put differently: how well do the review processes actually operationalise the aims of the funding instrument?

In recent years, widespread modifications have been made to standard research grant assessment processes, which traditionally involve peer and expert panel reviews. These include short pre-proposals, the inclusion of non-academic reviewers or panellists, and in-person presentations (Kolarz et al., 2023). Funders often introduce such modifications to ensure that the assessment captures a range of different aspects of applications submitted to a scheme. These changes aim to improve the alignment of review processes with the diverse aims of funding schemes, such as fostering interdisciplinarity or rewarding high-risk, high-reward proposals and address the limitations of conventional peer review, such as biases against highly innovative or interdisciplinary work (OECD, 2021; Criscuolo et al., 2017; Laudel & Gläser 2014; Luukkonen, 2012; Boudreau et al., 2012; Langfeldt, 2006). These evolving practices highlight the growing importance of adapting review processes to better reflect and support the multifaceted goals of research funding schemes.

At the same time, advances in artificial intelligence (AI) present new opportunities to further support and refine the assessment process for funding schemes. Generative AI tools have recently been explored for their potential to enhance peer review processes, from streamlining the review workflow to providing adaptive instructions and automating aspects of assessment. These tools can improve the standardisation and transparency of reviews, particularly when used to identify patterns or summarise reviewer feedback, though they also raise challenges, including confidentiality and ethical considerations (Su et al., 2023; Wong, 2023). Beyond peer review, generative AI has demonstrated value in managing large volumes of data, synthesising evidence, and producing detailed reports, as seen in applications like health technology assessments (Fleurence et al., 2024).

Funders and evaluators must assess how different criteria are featured in application reviews, and whether the process prioritises the funding instrument's aims. For instance, the effectiveness of early-stage assessments in filtering for originality has been well-documented in funding schemes that emphasise novel and unorthodox research approaches (Kolarz et al., 2023; Morgan et al., 2020). Similarly, AI-based tools can play a complementary role in capturing nuanced reviewer priorities, enhancing transparency, and reducing the likelihood of overlooking critical dimensions during assessments. Ultimately, understanding whether the assessment process aligns with the objectives of the funding instrument – and how emerging technologies like AI can support this alignment – is becoming increasingly important in shaping and improving the practices of peer review and evaluation.

CONTEXT: EVALUATION OF THE FWF'S EMERGING FIELDS SCHEME

The Austrian Science Fund's (FWF) Emerging Fields (EF) programme seeks to fund particularly original, innovative, paradigm-shifting research. Launched in 2022 as part of the Excellent=Austria initiative, the first call of the Emerging Fields (EF) Programme attracted 45 applications from a range of multidisciplinary Austrian research teams and five projects were ultimately funded.

The call for applications for the EF programme was explicitly open to inter- and transdisciplinary proposals. However, these were not strict eligibility requirements because inter- or transdisciplinary ideas are not inherently novel. Proposals may draw on multiple disciplines that reflect established or already integrated approaches without offering a particularly original or emergent combination. Therefore, the programme only encouraged inter- and transdisciplinarity but did not require it, and it asked applicants and reviewers to focus on how each proposal's approach contributed to the novelty of the research.

Following the programme's decision about its primary focus, our analysis did not treat interdisciplinarity as a standalone evaluative dimension. Instead, the elements of interdisciplinary and transdisciplinary indirectly integrate broader categories such as novelty or team qualifications – particularly where reviewers noted the diversity of perspectives, methodological blending, or team composition.

This evaluation accompanied the first call of the EF scheme end-to-end (Excluding the initial application submission window, so the study itself ran from February 2023 to March 2024). It was tasked critically to review the entire process, to identify strengths and weaknesses in its design, and to provide evidence-based recommendations to the FWF and its supervisory bodies on how to improve these processes for the next EF call. The full evaluation report and methodological annex materials are available on the FWF's website.

THE EF SELECTION PROCESS INVOLVES THE FOLLOWING STEPS:

- First, review of 3-page "synopses", i.e. short outlines of the project idea, by an international Jury of 16 experts. Interviews and other scoping work conducted as part of the early stages of the evaluation showed that stage 1 was intended primarily to assess and reward the novelty of the proposed research ideas

- Second, peer review by at least 3 external reviewers. The evaluation's scoping work showed that stage 2 was intended primarily to assess scientific quality in a conventional sense (feasibility, robustness, etc)

- Third, a presentation by shortlisted applicants to the international jury. The evaluation's scoping work showed that stage 3 was intended primarily to assess team composition. We note that the third assessment stage is out of the scope of this paper as it did not involve written review material in the way the other two stages did, and was too small in terms of application numbers to lend itself to meaningful analysis

The intentions behind each stage obtained through the scoping research on the scheme's design led us to a hypothesis that reviews in the first of these two stages should emphasise and reward innovative potential and novelty of the proposed project ideas, while reviews in the second stage should place a greater emphasis on scientific quality of the research plans.

Prominent international scientists reviewed the stage 1 synopses and stage 2 full proposals for the EF programme and produced evaluation documents detailing their judgements. In total, we had access to 140 records: 87 peer-reviews of synopses and 53 reviews of full proposals.

Given the number, heterogeneity and complexity of the review documents, generative AI was particularly useful in facilitating a systematic assessment. The multidisciplinary nature of the EF programme and the fact that EF applications are at the frontier of science make the review documents far from digestible for a general audience. They are heavy on scientific jargon and technical details, which are hard to understand for someone without deep expertise in each topic.

APPROACH/METHOD

GENERATIVE AI CLASSIFICATION AND REVIEW

We used generative AI to analyse peer reviewers' reports on applications submitted to the FWF's Emerging Fields (EF) programme. In total, we processed 140 review reports: 87 from the 3-page synopses in the first assessment stage and 53 from the full proposals in the second stage. We used a rich and comprehensive generative AI model that navigated these technical details to find relevant individual insights about each review document and stylised facts about the selection process in its two stages. The model was OpenAI's GPT4 large language model, accessed programmatically via a dedicated API. This access mode ensures the privacy and confidentiality of the underlying data.

It also enabled us to explore the capabilities of the GPT4 model in large-scale automation (querying all the documents programmatically without manually inputting and querying each review document individually).

Our focus in this analysis was explicitly on the reviewers' perspective. We did not analyse the proposals' text – whether synopses or complete applications – but examined how reviewers articulated their assessments of key criteria. As such, our results capture patterns in how novelty, scientific quality, feasibility, and team qualifications were perceived and expressed during the review process without making claims about the underlying content of the proposals.

To test our hypothesis, we conducted a comprehensive analysis using topic detection, sentiment analysis, and priority detection. These methodologies used the text of peer review documents from both, the first (synopses) and second (full proposals) assessment stages, as input data. Each approach targeted specific aspects of the review process, enabling a detailed exploration of the evaluative dimensions.

Topic detection aimed to identify and categorise distinct sections of the review text, focusing on four core dimensions: each proposal's novelty, risk or feasibility, scientific quality or rigour, and the research team's qualifications or suitability. The topic detection ensured a systematic examination of key aspects addressed by reviewers. This process enabled a systematic examination of the areas, most relevant to the assessment of proposals, providing insights into how reviewers prioritised and discussed these dimensions. For instance, identifying text specific to "novelty" allowed us to assess the emphasis reviewers placed on the innovative potential of the proposal, while sections addressing "risk or feasibility" highlighted their concerns or confidence in the project's practical execution. Similarly, isolating commentary on "scientific quality or rigour" shed light on reviewers' perceptions of the methodological soundness of the work, and text discussing "team qualifications or suitability" provided insights into the perceived capability of the research team to deliver on the proposed objectives.

Sentiment analysis evaluated the tone of reviewers' feedback on each dimension. It determined whether their observations were positive, negative, or neutral, providing an additional layer of understanding regarding the reviewers' perspectives. This analysis was complemented by explanations and direct quotes from the review documents, ensuring transparency and enabling cross-validation with the original text. The sentiment results were particularly valuable for sense-checking the robustness of the analysis. A fundamental assumption was that successful applications should be associated with more positive reviewer sentiments across these dimensions. If this relationship did not hold – if, for example, successful proposals received consistently negative or neutral sentiment – it would raise concerns about the quality and robustness of the model's outputs. We verified that the AI's classifications aligned with expected patterns by comparing sentiment scores to application outcomes, further validating the approach. This process ensured that the sentiment analysis was descriptive and instrumental in evaluating the trustworthiness and accuracy of the generated insights.

Finally, priority detection quantified the emphasis placed on each dimension by measuring the text length or word count dedicated to it. This served as a proxy for the time and effort reviewers allocated to discussing each aspect, revealing their implicit priorities during the evaluation process. By examining the relative word count devoted to each dimension, priority detection revealed the implicit priorities of reviewers during the evaluation process. For example, a higher word count for "novelty" might indicate that reviewers considered the proposal's innovative potential particularly significant. At the same time, more extensive commentary on "risk or feasibility" could suggest a greater focus on practical concerns. Similarly, longer sections addressing "scientific quality" or "team qualifications" highlighted areas where reviewers invested the most effort to assess the proposal's merit.

The priority detection method provided insights into reviewers' implicit weighting of evaluation criteria and served as a complementary tool to topic detection and sentiment analysis. By triangulating these results, we better understood the review process and how reviewers allocated their attention to various dimensions. Priority detection also enabled comparisons across proposals, offering a standardised way to interpret and analyse the focus of reviewer feedback.

Our prompts were designed to guide the model's responses, ensuring the outputs' precision, clarity, and accountability. For topic sentiment analysis, the model made clear judgements (positive, negative, neutral, or not discussed) for each dimension. We requested concise explanations of up to 100 words to support these judgements, accompanied by direct quotes from the reviewers' text. This approach ensured transparency and allowed for human cross-validation. For priority detection, the prompts instructed the model to rank topics by their importance based on the number of words dedicated to each. The output included a clear ranking and numerical word count estimates for each topic, following a predefined format to ensure consistency and interpretability. The following boxes contain the structured prompts we deployed to ensure the model's outputs were consistent and interpretable, enabling robust downstream analyses while maintaining accountability.

BOX 1. SYSTEM MESSAGE (UNIVERSAL)

Users will provide text from scientific reviews. These reviews follow a common section template but may include different responses in each section. Answer questions using only information provided in the reviews.
Expect users to use the following format:
Review: ***text from the review here***
Question: ***user question here***

Source: Technopolis-Group

BOX 2. TOPIC SENTIMENT PROMPT

Is the reviewer positive, negative, or neutral regarding the proposal's [topic]?

Response format:
Positive/Negative/Neutral/Not discussed. Explanation (maximum 100 words, always include quotes from the reviewer's text).

Source: Technopolis-Group

BOX 3. PRIORITY DETECTION PROMPT

Sort the following topics by importance based on the number of words the reviewer dedicates to each topic. Provide the ranking alongside an estimate of the total number of words per topic. Topics are:
** A. Novelty,
** B. Risk/Feasibility,
** C. Scientific Quality/Rigour,
** D. Qualifications/Suitability of the Team.

##Response format:

[A/C/B/D (104/51/50/20)]
##Note:

** Letters indicate the rank of topics from most to least important.
** Parentheses contain the estimated word count for each topic.
** Provide only the output in brackets; no additional text or commentary.

Source: Technopolis-Group

To address the potential for hallucinations inherent in the black-box nature of AI, we prioritised accountability and transparency in every stage of the analysis. The unpredictable tendency of generative AI models to produce outputs not grounded in the input data required rigorous safeguards to maintain the reliability of results. For each response generated by the model, we explicitly requested reasoning that provided a clear rationale for the decision, accompanied by direct quotes from the original review text, used as the basis for the output. This approach ensured that the model's decisions were grounded in the reviewers' actual statements, avoiding speculative or fabricated content.

We also conducted systematic sense-checking of the outputs by comparing them to real-world examples, such as those presented in Table 1 and Table 2. These examples demonstrate how the model accurately generated explanations for positive and negative assessments of the proposals' novelty, ensuring that its outputs aligned with the source text. Positive examples (e.g. Proposals X, Y, Z) illustrate how the model identified novelty and transformative potential in reviewers' comments. Conversely, negative examples (e.g., Proposals D, E, F) show the model identifying a lack of novelty or incremental contributions in the reviewers' assessments.

The model provided interpretable and verifiable outputs by incorporating direct quotes from the reviewers and framing its responses in a structured format. We manually cross-validated a 20% sample of these outputs, enabling us to agree or disagree with the AI's interpretations and intervene when necessary. In this process, we confirmed that in all sampled cases, the resulting quotes were not hallucinations, and all of the results were reasonable, with just a couple of cases which could be non-obvious/open to interpretation but not patently wrong. The dual-layered process of automated output generation and our human validation ensured robust alignment between the model outputs and the underlying data, significantly mitigating the risk of errors or unsupported conclusions. The tables of examples below further illustrate this process, showcasing how the model's structured approach maintained accuracy and transparency in its outputs.

Table 1. Sense-checking outputs from the LLM (positive examples)

Document	Is the reviewer positive, negative or neutral regarding the proposal's novelty? Answer in the following format: Positive/Negative/Neutral/Not discussed. Explanation (max. 100 words and always include quotes from the reviewer's text).
Proposal X	Positive. The reviewer states that the proposal is "innovative and extremely difficult but totally right" and that the researchers are 'trying to give birth to a new method of analysis which they call the '[method X]' (…), not by simply carrying out a conventional multidisciplinary research with three different disciplines, but by creating a new research field, a sort of new (and needed) interdisciplinary field". This indicates that the reviewer sees the proposal as bringing novelty to the field.
Proposal Y	Positive. Explanation: The reviewer states that the proposal "has the potential to completely transform [topic 1], [topic 2], and perhaps [topic 3]", indicating a belief in its novelty and transformative potential. They also mention that "the groundwork has already been laid over the past decade", suggesting that the proposal is building on recent advances to potentially create a paradigm shift. The reviewer's enthusiasm is evident in the statement, "This is an impressive proposal. I rank it as the best proposal I have ever had the opportunity to review".
Proposal Z	Positive. Explanation: The reviewer states that the project "has a tremendous potential to generate a quantum leap in [topic X]" and commends the applicants for an idea "that has the potential to fundamentally challenge the field of [topic X] and established notion of research". The reviewer also mentions that "a paradigm shift is likely to emerge", indicating a strong belief in the novelty and transformative potential of the proposal.

Source: Technopolis-Group

Table 2. Sense-checking outputs from the LLM (negative examples)

Document	Is the reviewer positive, negative or neutral regarding the proposal‘s novelty? Answer in the following format: Positive/ Negative/Neutral/Not discussed. Explanation (max. 100 words and always include quotes from the reviewer‘s text).
Proposal XX	Negative. Explanation: The reviewer states, “The idea of [research question X] is not a novel idea, it is in fact one of the most sought-after challenges being pursued at an international level”. They also mention that “many of the proposed topics come across as incremental to each of the respective PIs” and that “the high-risk high gain for the project as a whole is missing”. These statements indicate that the reviewer does not find the proposal’s novelty to be particularly strong or transformative.
Proposal YY	Negative. Explanation: The reviewer states that the goal of [work package] “is an ambitious goal, albeit not novel in itself”, indicating that the idea is not new. Furthermore, the lack of a clear focus in the proposal as a whole makes it difficult for the reviewer to comment on its novelty, suggesting a negative view on the proposal’s potential for novelty.
Proposal ZZ	Negative. Explanation: The reviewer states that “the general idea is not of particular novelty, as [topics X/Y] are now widely applied to [field Z]”, and that “It is difficult to imagine how they will lead to a paradigm shift”. They also mention that “this proposed project will not be much different” from other projects at large in [field Z] centers, suggesting that the proposal lacks the originality required for a significant impact on the field.

Source: Technopolis-Group

We combined these methodologies to capture qualitative and quantitative insights into the review process. This multifaceted approach enabled us to understand better how reviewers balanced their assessments across different proposal dimensions, such as novelty, risk/feasibility, scientific quality/rigour, and team qualifications/suitability. By systematically integrating the outputs from topic detection, sentiment analysis, and priority detection, we could identify trends and patterns in evaluative criteria and their relative emphasis.

TRANSFORMATION OF QUALITATIVE INSIGHTS INTO QUANTITATIVE SCORES

To facilitate quantitative analysis, we transformed the outputs of the sentiment and priority detection processes into numerical scores. This allowed for a more structured comparison across proposals and reviewers.

In the sentiment indicators, for each dimension d (novelty, risk/feasibility, scientific quality, and team qualifications), the final metric consisted on the following rule:

$$S_{i,d} = \begin{cases} 1, & if\ sentiment\ is\ positive\ for\ dimension\ d\ in\ topic\ i \\ 0, & otherwise \end{cases}$$

Where:

- $S_{i,d}\ is\ the\ sentiment\ score\ for\ dimension\ d\ in\ document\ i, and$
- $d \in \{novelty, risk, quality, team\}$

This binary scoring approach ensured that only positive sentiments were counted, clearly distinguishing between favourable and neutral/negative evaluations.

Priority detection quantified each dimension‘s prominence by inverting the topics’ ranking based on the word count dedicated to them. The highest-ranked dimension (i.e. the one with the most words) received the maximum value, with subsequent dimensions receiving progressively lower scores. The scoring rule was defined as:

$$P_{i,d} = 4 - R_{i,d}$$

- $P_{i,d}$ *is the priority score for dimension d in document i, and*
- $R_{i,d}$ *is the rank of dimension d based on word count in document i (1 for the most words, 4 for the least words).*

As a result, the dimension with the highest prominence received a score of 3, the second most prominent received 2, the third received a score of 1, and the least prominent scored 0.

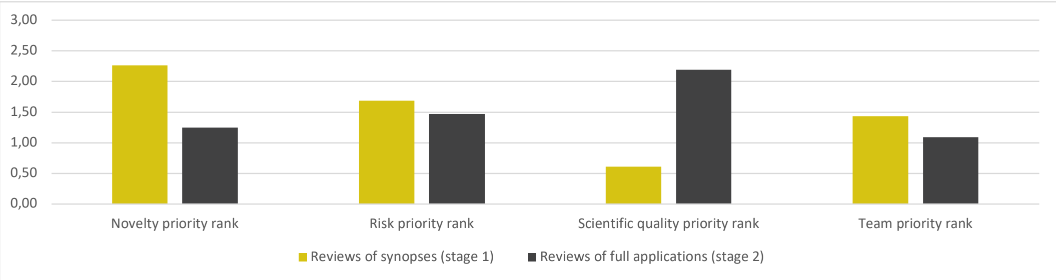
This dual-layered scoring system allowed us to perform statistical analyses and identify patterns, in how reviewers emphasised and evaluated key dimensions of the proposals. Combining sentiment and priority indicators enabled us to develop a detailed framework that facilitated a robust, data-driven understanding of the review process and the evaluation criteria prioritised by reviewers. These scores also enabled comparisons across proposals, further enhancing the interpretability of the results.

RESULTS

Our findings support our main hypotheses, highlighting significant differences in the priorities reviewers assign to proposal dimensions across the two evaluation stages. Most notably, novelty emerges as a significantly higher priority during stage 1 (synopses review), while scientific quality becomes the dominant focus in stage 2 (full proposals).

The analysis of priority rankings reveals a clear shift in focus between the stages. At the synopses stage, reviewers prioritise the novelty of the proposals, reflected by an average ranking score of 2.26 out of a possible 3. In contrast, novelty receives less emphasis during the full proposal stage, where its average ranking score drops to 1.25. Conversely, scientific quality receives minimal attention at stage 1, with an average score of 0.61, but becomes the highest-ranking dimension at stage 2, scoring 2.19. These results align with the expectations that stage 1 primarily evaluates proposals' originality and innovative potential. In contrast, stage 2 involves a more detailed examination of the proposal's scientific robustness and methodological quality (see Figure 1).

Figure 1 Average priority ranking scores for synopses (stage 1) and proposals (stage 2)

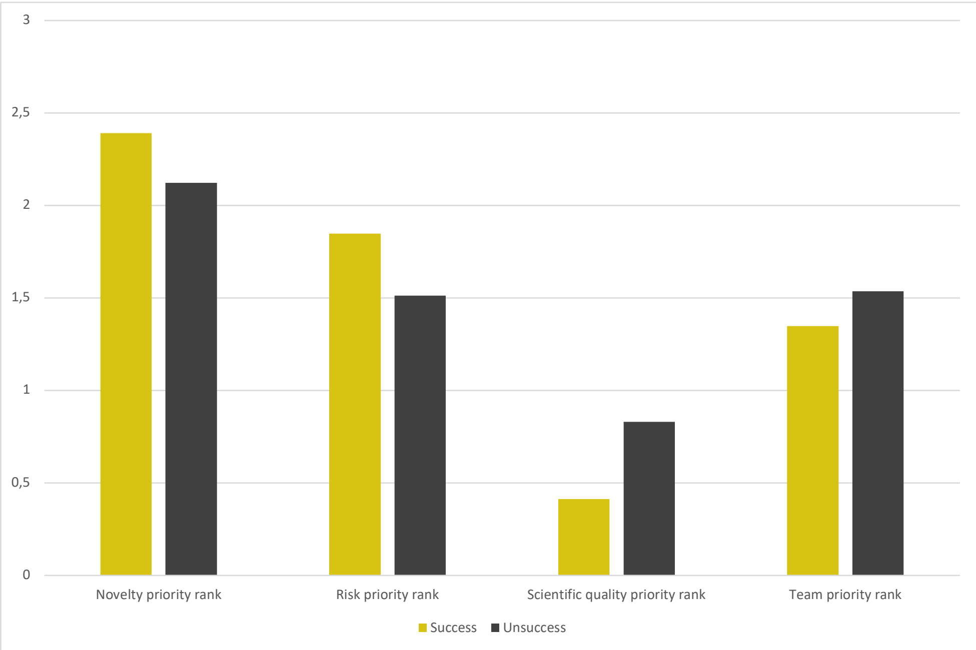


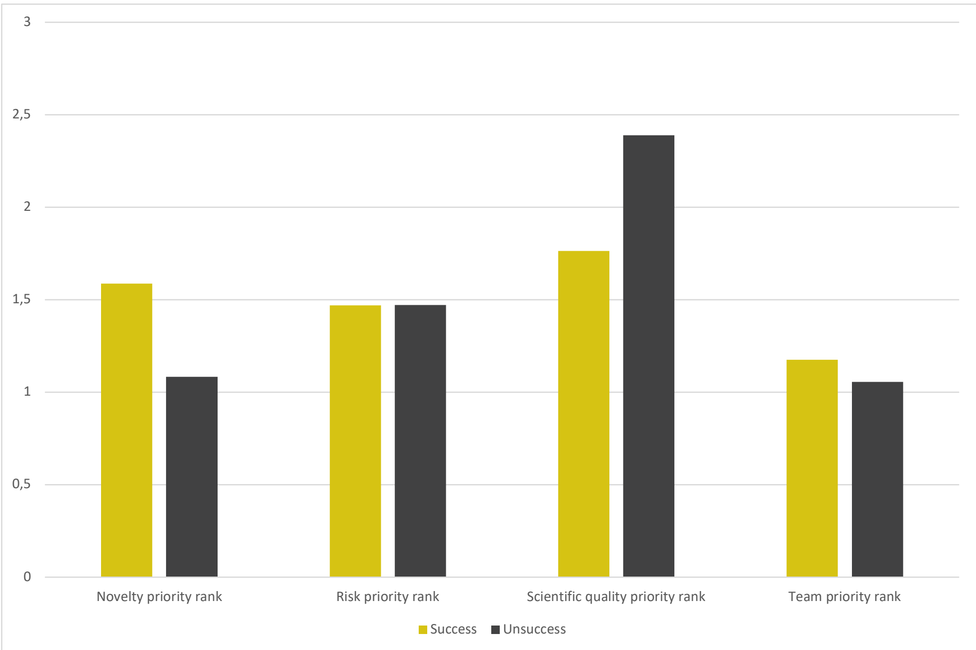
Source: Technopolis Group analysis based on peer-review documents

The dimensions of risk and team qualifications also receive greater attention at stage 1 than at stage 2, albeit with less pronounced differences compared to novelty and scientific quality. This suggests that reviewers dedicate more effort to assessing the feasibility and team capabilities during the initial stage, likely to gauge whether the proposal warrants further consideration.

Additional patterns emerge when we examine the breakdown of priority rankings between successful and unsuccessful applications. Figure 2 illustrates that successful applications consistently score higher in novelty and team qualifications, while unsuccessful applications exhibit slightly higher risk and scientific quality scores. This suggests reviewers may emphasise novelty and team strength when advancing proposals through the selection process. In contrast, unsuccessful applications are likely to undergo more rigorous scientific quality and feasibility scrutiny, possibly reflecting a focus on identifying methodological shortcomings. These observations further support the idea that the two evaluation stages serve distinct functions, with the synopses stage acting as a filter for innovative and promising proposals. In contrast, the full proposal stage ensures that the selected projects meet high scientific standards.

Figure 2. Average priority ranking scores for successful and unsuccessful synopses and proposals





Source: Technopolis Group analysis based on peer-review documents

The sentiment analysis provides insight into how reviewers assess successful versus unsuccessful proposals. As illustrated in Figure 3, successful proposals consistently exhibit higher positive sentiment across all dimensions, with particularly notable differences for novelty and team qualifications. For novelty sentiment, successful proposals score an average of 0.7, significantly higher than the 0.3 average for unsuccessful applications. Sentiment towards risk also shows a marked difference, with successful proposals scoring 0.5 compared to 0.2 for unsuccessful ones. Similarly, scientific quality sentiment reveals a substantial gap, with successful proposals achieving an average score of 0.67 while unsuccessful applications score only 0.24. Finally, the review of team qualifications shows similar patterns, where successful proposals score an average of 0.82 compared to 0.58 for unsuccessful applications.

These findings suggest that reviewers consistently view successful proposals more favourably across all dimensions, emphasising the importance of novelty, risk management, and team strength in advancing proposals through the selection process. The pronounced differences in sentiment highlight these dimensions' critical role in shaping evaluation outcomes and align with the observed priorities for successful applications.

Figure 3 Average sentiment score for synopses (stage 1)

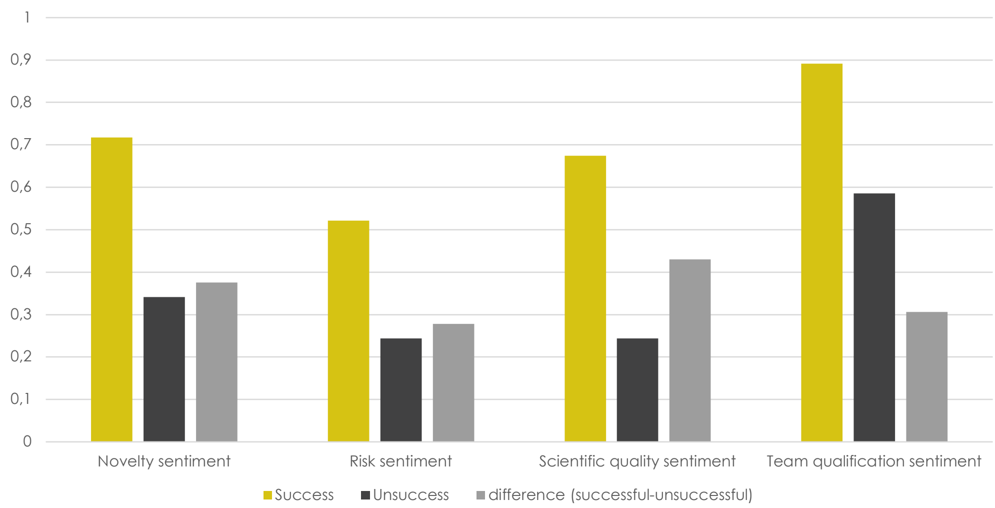
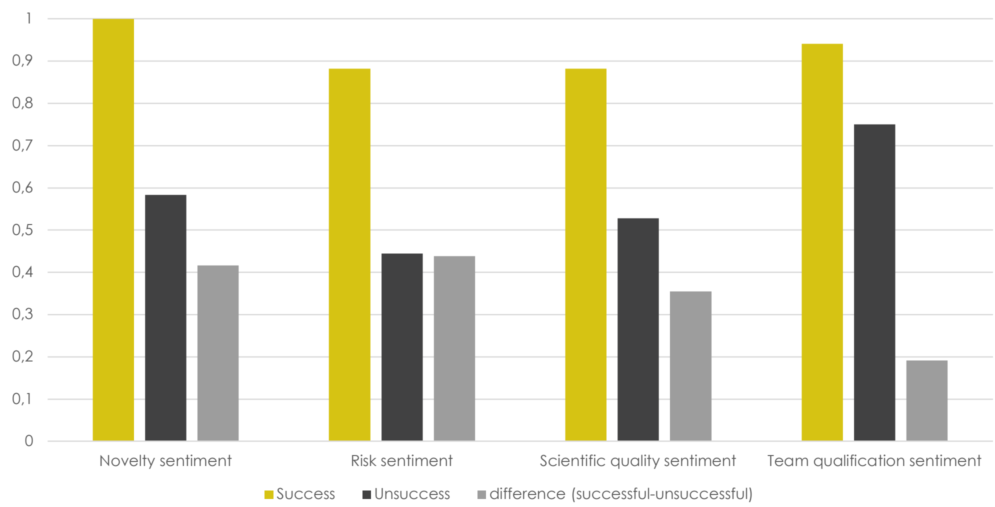


Figure 4 shows how reviewers assess successful versus unsuccessful proposals during the proposal stage (phase 2). Successful proposals consistently generate more positive sentiment across all dimensions, with particularly notable differences in novelty and risk. For novelty sentiment, successful proposals score an average of 1.00, significantly higher than the 0.60 average for unsuccessful applications. Similarly, sentiment towards risk shows a marked difference, with successful proposals scoring 0.93 compared to 0.48 for unsuccessful ones. Scientific quality sentiment also reveals a substantial gap, with successful proposals achieving an average score of 0.93, while unsuccessful applications score only 0.55. Finally, team qualification sentiment shows the smallest difference but remains significant, with successful proposals scoring an average of 0.96 compared to 0.76 for unsuccessful applications.

In comparing the sentiment ratings between the synopsis and the full proposals, the overall trend is that successful proposals receive higher sentiment scores. However, the magnitude of these differences varies. For instance, the gap in novelty sentiment is slightly larger when evaluating the full proposal (0.40) compared to the synopsis (0.38), and the difference in risk sentiment shows an even more pronounced increase from 0.28 in the synopsis to 0.45 in the full proposal. In contrast, scientific quality sentiment is greater when only the synopsis is assessed (0.43) than when the full proposal is reviewed (0.38). Similarly, team qualification sentiment has a larger gap for the synopsis (0.31) than for the full proposal (0.20). While successful proposals consistently score higher across both stages, the specific extent to which success and failure separate in terms of novelty and risk sentiment becomes more pronounced at the full proposal stage. In contrast, differences in scientific quality and team qualification sentiments are more prominent when evaluators focus on the synopsis alone.

Figure 4. Average sentiment scores for proposals' evaluation (stage 2)



Source: Technopolis Group analysis based on peer-review documents

The comparison of priority rankings and sentiment differences between the synopsis and proposal stages highlights a clear evolution in the evaluation focus.

Novelty is central to reviewers' assessments during the synopses stage, as reflected by its high priority rank (2.26), but it diminishes in relative importance at the proposal stage (1.30). This shift suggests that reviewers emphasise other dimensions more once the initial novelty threshold is met. By contrast, scientific quality experiences a substantial increase in prioritisation, with its rank rising from 0.61 in the synopses stage to 2.17 in the proposal stage. This finding aligns with the observed sentiment patterns and underscores the importance of deeper scrutiny of methodological robustness and scientific rigour as proposals progress through the evaluation process. While important across both stages, risk shows a relatively stable ranking (1.69 at the synopsis stage and 1.46 at the proposal stage) despite an increase in sentiment differences at the proposal stage. This indicates that risk considerations remain a consistent component of the review process, with sentiment differences reflecting the level of engagement with feasibility and risk mitigation as proposals advance. Finally, while receiving high positive sentiment at both stages, team qualifications are comparatively less prioritised, especially during the proposal stage. This suggests that while team strength is acknowledged as an important factor, it plays a more supportive role relative to dimensions such as novelty, risk, and scientific quality.

These findings demonstrate a structured progression in the review process, with novelty playing a crucial filtering role at the synopsis stage and scientific quality emerging as the dominant consideration during the proposal stage. Risk remains a steady focus throughout, while team qualifications, despite positive sentiment, receive comparatively less prioritisation. These shifts reflect the nuanced and evolving priorities of the evaluation process, ensuring that innovative potential and scientific rigour are appropriately assessed at different stages.

TRIANGULATION WITH OTHER EVALUATION APPROACHES

We note that the evaluation of the Emerging Fields scheme also included other method components, some of which we can draw on to further enhance confidence in our results. The full evaluation methodology can be found in the annex sections of the evaluation (Kolarz et al., 2024). However, the following two are relevant to our findings:

The evaluation included in-person observation of the FWF Scientific Board meetings that followed stages 1 and 2. In each case, the Scientific Board discussed the reviews of the synopses/full applications and the evaluation team coded each discussion point thematically and in terms of sentiment. This exercise showed that for stage 1, Board meeting discussions focused strongly on novelty and at stage 2 far more on scientific quality. The Board discussions are of course separate from the written reviews themselves, but the reviews are the main material informing the Board discussions. While there is room here for alternative explanations, it is certainly plausible that the Board discussions were shaped by the reviews. In other words, the strong focus on novelty-aspects in Board discussions on stage 1 synopses was likely at least in part due to an emphasis on novelty in the reviews available to the Board.

Secondly, the evaluation team conducted a survey of individuals who acted as external reviewers for stage 2 full applications. One survey item asked them to self-assess which of a list of criteria they emphasised most in their judgement of the application they reviewed. Reviewers judged themselves to have emphasised criteria around scientific quality over all other aspects. This gives us confidence that the stage 2 reviews of full applications were indeed centrally about scientific quality rather than novelty.

Neither of these two method components allows us to look inside the reviewing process as directly as our analysis presented in this paper, so neither can be a substitute method capable of generating the type of findings presented in this paper. However, we note them here as an additional mechanism of triangulation underscoring the robustness of our findings.

CONCLUSION/DISCUSSION

This study provides evidence for the effectiveness of a two-stage review system in research funding processes, particularly for programmes emphasising scientific novelty. In the case of the Austrian FWF's Emerging Fields (EF) scheme, the two-stage approach – first requiring applicants to submit a short synopsis, followed by a full proposal for shortlisted candidates – has proven beneficial in multiple ways. Moreover, our study demonstrates that generative AI can contribute to a more systematic and transparent review of the evaluation process. This conclusion highlights two major insights: (1) how the initial stage of short synopses helps ensure that novelty is given a strong and meaningful focus, and (2) the potential of using generative AI to enrich and support the peer-review process, capturing nuanced reviewer priorities and sentiments while reducing the risk of overlooking critical dimensions.

One of the most salient insights from this analysis is that the emphasis on novelty – an essential ingredient of pioneering research – emerges strongly during the first stage of the EF scheme. Indeed, when reviewers assess shorter proposal documents, they often devote significant attention to how innovative or groundbreaking a proposed project might be. A condensed synopsis, typically only a few pages long, appears to be particularly conducive to focusing the reviewer's attention on the "big idea": the conceptual leap or novelty that distinguishes one proposal from another. There are several possible reasons for why this emphasis on originality is pronounced at the first stage:

- Information economy: Short synopses can limit the amount of detail applicants can provide, naturally directing reviewers' attention to the headline objectives, the main research question or hypothesis, and why the proposal is new (Koronis et al., 2021; Nomaguchi et al., 2019). The absence of exhaustive methodological details forces reviewers to focus on conceptual and theoretical novelty. Thus, they can understand whether the proposal pushes the field's boundaries and is worth pursuing further.

- Early filter for originality: The EF scheme aims to seed disruptive or unorthodox projects that might not always fit within traditional funding calls. By spotlighting novelty in the first stage, the scheme ensures that highly innovative yet underdeveloped proposals are not prematurely rejected simply because their details are still taking shape. The short stage 1 application format encourages creative thinkers, whose primary strength might be a bold vision, to apply without being caught up by voluminous instructions or a requirement for fully-formed methodologies.

- Reduced cognitive load for reviewers: Reading a three-page synopsis is far less time-intensive than reading a detailed 30-page proposal. Reviewers can more easily compare multiple synopses, identify the key ideas, and grasp the potential impact of each. This approach can be particularly valuable for busy evaluators, often leading to clearer feedback and more decisive endorsements – or rejections – based on novelty.

Following this first novelty filter, the second stage necessarily foregrounds scientific rigour. Once a proposal has passed an initial threshold for originality, it must be presented in greater detail: applicants provide a full proposal elaborating on data sources, methods, pilot results, team composition, feasibility, risk management, and ethical considerations. Thus, there is a natural shift in reviewer priorities. Our analysis shows a marked increase in attention paid to scientific quality – both in terms of word count devoted to that criterion and the overall sentiment expressed by reviewers. This shift is logical and follows the progressive structure of the most thorough academic and funding processes: first, identify which ideas are truly worth developing; second, ensure that those promising ideas can be executed successfully.

At times, novelty may come at the cost of methodological weaknesses or incomplete feasibility planning. Conversely, high scientific rigour may inadvertently deprioritise groundbreaking concepts, favouring familiar or incremental approaches. The two-stage approach helps strike a balance between these extremes. It provides a structured means to:

- Allow original thinkers to enter the funding pipeline: By focusing on originality first, this process mitigates the risk that truly novel ideas will be lost because their methods are still evolving.
- Evaluate scientific quality at an appropriate time: By deferring an extensive methodology assessment to the second stage, the approach prevents early-stage proposals from being dismissed purely on technical grounds, which might be refined later.
- Promote transparency and accountability: Clear demarcation of stage 1 vs. stage 2 criteria allows applicants to understand precisely what is being evaluated and when. This fosters more targeted writing at each step, improving the coherence and efficiency of the review process.

These strengths underscore that a two-stage system can help funders manage large cohorts of applications while ensuring that novelty receives ample attention. The EF case study illustrates the potential for such systems to be replicated in other funding contexts where multi-dimensional criteria (for instance, interdisciplinarity, societal relevance, or potential for transformative impact) must be weighed alongside more traditional scientific standards.

Our analytical approach focused on the formal dimensions the programme formally prioritised because they had more apparent reviewer consensus and variation to measure with our tools. While recognising that concepts like inter- and transdisciplinarity were part of the broader evaluative landscape, we did not directly operationalise them in the analysis. Future work could explore more targeted methods to surface how inter- and transdisciplinary framings influence peer assessment and how funders might more clearly guide reviewers in how to weigh such factors.

Relatedly, future work could also explore how the proposals' content and framing shape reviewers' assessments. Analysing both proposals and their corresponding reviews could help disentangle whether an emphasis on novelty or scientific quality reflects the reviewers' preferences or specific signals in the application texts.

A secondary but equally crucial aspect of our work involves integrating generative AI tools into the funding evaluation process. Our study demonstrates that these advanced natural language processing systems can serve as powerful complements to human peer reviewers, with multiple potential benefits:

- Systematic analysis of reviewer reports: AI can parse large volumes of text across multiple documents, detecting recurrent themes, sentiments, and prioritisations. By automating part of this analysis, AI reduces the manual workload, ensuring that commonalities and outliers are consistently surfaced and providing relevant insights at the portfolio level to funders.

- Triangulation with other methods: The generative AI approach can be combined with surveys, interviews, or panel discussions to confirm – and sometimes challenge – emerging patterns. This triangulation is valuable for verifying whether AI-generated insights align with feedback from evaluators or applicants, thereby enhancing the reliability of conclusions.

- Detection of implicit biases: Because AI systematically categorises text, it can more systematically highlight biases or inconsistencies. For instance, an AI might identify that certain criteria – such as team qualifications – are underdiscussed or overshadowed by excessive focus on feasibility. This insight allows funding organisations to refine guidelines or training for reviewers.

- Speed and scalability: As the volume of applications grows, reviewer committees struggle to maintain consistency, timeliness, and thoroughness in their evaluations. Conversely, AI can process thousands of documents in parallel, providing preliminary analyses that human reviewers can interpret and refine. This scale-up potential means that large or complex calls can be handled more efficiently without sacrificing thoroughness.

While our results highlight the promise of generative AI, there are important cautionary notes and considerations that funding agencies and evaluators must consider. One major concern is hallucinations, as large language models can sometimes produce information that is not grounded in the input data. To address this, we required the AI system to provide direct quotes from reviewers' text wherever possible, explicitly linking claims to textual evidence. However, only requesting quotes is insufficient because, in a worst-case scenario, the requested quotes could be hallucinations. Additional safeguards such as human cross-checking and carefully crafted prompts remain essential to prevent misleading or fabricated outputs. We manually verified a random sample of 20% of the answers to ensure the quotes were not hallucinations and to sense-check the results.

Another critical issue is biases inherent in model training. AI models are trained on vast amounts of publicly available text, which can inadvertently encode societal and cultural biases. While prompt engineering and domain-specific fine-tuning can partially mitigate these effects, users of AI-based tools must remain vigilant about potential skew or unfairness, particularly in sensitive contexts such as equity, diversity, and inclusion.

Overreliance on automation also poses risks, as AI should not entirely replace human judgment in the evaluation process. Peer review is inherently rooted in disciplinary expertise, contextual understanding, and intangible aspects of academic rigour that are challenging to codify in algorithms. The optimal use of AI lies in its role as an assistant, helping to identify patterns or areas that warrant deeper scrutiny while leaving final decisions to human experts. Moreover, data confidentiality is paramount when dealing with sensitive proposals and confidential reviewer statements. Ensuring the security and privacy of this data is critical, and the approach we used – processing reviews through dedicated API connections and securing stored data strictly in EU jurisdictions – should serve as a baseline best practice for funding organisations. Addressing these considerations will allow AI's responsible and effective integration into evaluation processes while maintaining trust, fairness, and accountability.

Taken together, the two primary lessons from this study are deeply intertwined. First, a multi-stage review process helps segment the evaluation of novelty from subsequent methodological and conceptual depth assessments, thereby allowing truly fresh ideas to surface and preventing premature dismissal. Second, generative AI has clear potential to streamline and enrich the evaluation of those multi-stage processes, furnishing an objective, data-driven lens that can detect patterns in reviewer feedback, highlight areas of discrepancy or consensus, and surface deeper insights into the strengths and weaknesses of each proposal.

Future developments in large language models, including improved capacity to handle domain-specific technical jargon and better safeguards against hallucinations, will likely expand their utility in research funding evaluations. Meanwhile, the ongoing conversation around how best to balance novelty and feasibility—exemplified by a two-stage approach—could continue to evolve. As funding bodies experiment with new ways of fostering ground-breaking science (including open peer review, co-creation processes, and user-driven research agendas), the careful integration of AI tools offers a viable path to ensuring thorough, data-enriched, and balanced assessments.

Ultimately, what this study underscores is the value of intentional design in the funding process. Different evaluation criteria (such as novelty, feasibility, and team capacity) require distinct review frameworks, and by supplementing peer review with structured AI insights, funding agencies can better identify, nurture, and support transformative research proposals that truly push the boundaries of knowledge. The two-stage EF mechanism is one such manifestation of this design philosophy, demonstrating that when novelty is given pride of place early on, many more original and pioneering ideas have the opportunity to flourish, subsequently facing the rigorous scientific scrutiny they require at the full-proposal stage. In tandem, generative AI ensures a level of systematic analysis that is both scalable and transparent, paving the way for an iterative cycle of continuous improvement in research funding evaluation.

BIBLIOGRAPHY

Boudreau, K., Guinan, E., Lakhani, K., & Riedl, C. (2012). The Novelty Paradox & Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations. https://doi.org/10.2139/ssrn.2184791.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2017). Evaluating Novelty: The Role of Panels in the Selection of R&D Projects. IRPN: Other Development of Innovation (Topic). https://doi.org/10.2139/ssrn.3650896.

Fleurence, R., Bian, J., Wang, X., Xu, H., Dawoud, D., Higashi, M., & Chhatwal, J. (2024). Generative AI for Health Technology Assessment: Opportunities, Challenges, and Policy Considerations. ArXiv, abs/2407.11054. https://doi.org/10.48550/arXiv.2407.11054.

Kolarz, P., Vingre, A., Machado, D., Sutinen, L., Dudenbostel, T., & Arnold, E. (2024). Accompanying process evaluation of FWF's Emerging Fields. Zenodo. https://doi.org/10.5281/zenodo.13911479.

Kolarz, P., Vingre, A., Vinnik, A., Neto, A., Vergara, C., Obando Rodrigues, C., Nielsen, K. & Sutinen, L. (2023). Review of peer review. UKRI, UK. https://www.ukri.org/publications/review-of-peer-review/.

Koronis, G., Casakin, H., & Silva, A. (2021). Crafting briefs to stimulate creativity in the design studio. Thinking Skills and Creativity, 40, 100810. https://doi.org/10.1016/J.TSC.2021.100810.

Langfeldt, L. (2006). The policy challenges of peer review: managing bias, conflict of interests and interdisciplinary assessments. Research evaluation, 15(1), 31–41.

Laudel, G. & Gläser, J. (2014). 'Beyond breakthrough research: Epistemic properties of research and their consequences for research funding.' Research Policy, 43(7): 1204–16.

Luukkonen, T. (2012). Conservatism and risk-taking in peer review: Emerging ERC practices. Research Evaluation, 21(1), 48–60

Morgan, B., Yu, L., Solomon, T., & Ziebland, S. (2020). Assessing health research grant applications: A retrospective comparative review of a one-stage versus a two-stage application assessment process. PLoS ONE, 15. https://doi.org/10.1371/journal.pone.0230118.

Nomaguchi, Y., Kawahara, T., Shoda, K., & Fujita, K. (2019). Assessing Concept Novelty Potential with Lexical and Distributional Word Similarity for Innovative Design. Proceedings of the Design Society: International Conference on Engineering Design. https://doi.org/10.1017/DSI.2019.147.

OECD (2021). „Effective policies to foster high-risk/high-reward research“, OECD Science, Technology and Industry Policy Papers, No. 112, OECD Publishing, Paris, https://doi.org/10.1787/06913b3b-en.

Su, X., Wambsganss, T., Rietsche, R., Neshaei, S., & Käser, T. (2023). Reviewriter: AI-Generated Instructions For Peer Review Writing, 57–71. https://doi.org/10.18653/v1/2023.bea-1.5.

Wong, R. (2023). Role of generative artificial intelligence in publishing. What is acceptable, what is not … The Journal of ExtraCorporeal Technology, 55, 103–104. https://doi.org/10.1051/ject/2023033.

AUTHORS

PETER KOLARZ
Research on Research Institute
Email: p.kolarz@researchonresearch.org
ORCID: 0000-0003-4443-580X

DIOGO MACHADO
Technopolis
Email: diogo.machado@technopolis-group.com
ORCID: 0000-0001-6147-876X

Chief Editors of issue #57: Michael Dinges (AIT), Elisabeth Nindl (JRC Sevilla), Isabella Wagner (fteval)
Proofreading and editorial support: Theresa Kahri (fteval)

Published by the Austrian Platform for Research and Technology Policy Evaluation (fteval)
Linke Wienzeile 246, c/ZSI, 1150 Vienna, Austria
T: +43 1 495 04 42 - 33, **E:** office@fteval.at, **W**: www.fteval.at

The fteval Journal for Research and Technology Policy Evaluation is a forum to discuss current evaluation practices in the field of research, technology and innovation policy.

AUSTRIAN PLATFORM FOR RESEARCH AND TECHNOLOGY POLICY EVALUATION MEMBERS:

Federal Ministry of Women, Science and Research; Federal Ministry for Economy, Energy and Tourism; Federal Ministry for Innovation, Mobility and Infrastructure; Austrian Academy of Sciences; ACR - Austrian Cooperative Research; FORWIT - Austrian Council for Sciences, Technology and Innovation; Austrian Institute For SME Research; AIT - Austrian Institute of Technology; AQ Austria - Agency for Quality Assurance and Accreditation Austria; AWS - Austria Wirtschaftsservice Gesellschaft mbH; CDG - Christian Doppler Research Association; convelop cooperative knowledge design GmbH; FH Campus Wien, FFG - Austrian Research Promotion Agency; FWF - Austrian Science Fund; Gesellschaft für Forschungsförderung Niederösterreich; IHS - Institute for Advanced Studies; IWI – Industriewissenschaftliches Institut; Joanneum Research Forschungsgesellschaft mbH; Klima- und Energiefonds; Ludwig Boltzmann Gesellschaft; Technopolis Group | Austria; OeAD – Agency for Education and Internationalisation; Vienna Business Agency. A service offered by the City of Vienna; WIFO - Austrian Institute of Economic Research; WWTF - Vienna Science and Technology Fund; WPZ Research GmbH; ZSI - Centre for Social Innovation

ZVR-Zahl: 937261837, **ISSN-Nr**.: 1726-6629, **DOI:**10.22163/fteval.2025.692

AUSTRIAN PLATFORM
for Research and Technology Policy Evaluation

Linke Wienzeile 246, c/o ZSI
1150 Vienna, Austria

The fteval Journal for Research and Technology Policy Evaluation is a forum to discuss current evaluation practices in the field of research, technology and innovation policy.

+43 1 495 04 42 - 33
office@fteval.at
www.fteval.at

ZVR-Zahl: 937261837
ISSN-Nr. 1726-6629
DOI: 10.22163/fteval.2025.692