Evaluation and Program Planning 79 (2020) 101750

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Developing a conceptual evaluation framework for gender equality interventions in research and innovation

Evanthia Kalpazidou Schmidt*, Ebbe Krogh Graversen

Department of Political Science, Aarhus University, Denmark

ARTICLE INFO

Keywords:
Innovative conceptual evaluation framework
Gender equality interventions and effects
R&I
Complexity
Non-linear approach
Gender sensitivity
Theory-based evaluation
Instrumentation

ABSTRACT

In this article, we discuss the development of a conceptual evaluation framework to design and assess gender equality interventions and their effects in research and innovation. The conceptual framework presented herewith embraces the complexity, gender-sensitive and theory-based evaluation approaches ensuring that design and evaluation of gender equality interventions consider the complex systems that constitute the context in which the interventions operate. The evaluation framework offers a non-linear concept, where the notion of contribution - not attribution - to achieve impact is central to the integration of team, organizational and system factors in policy design and evaluation. The paper opens the "black box" to address the question of how and why a policy intervention works and in which context and discusses a systematic process on how to approach the interwoven linkages between input, implementation and effects in gender equality interventions in research and innovation, accounting for context sensitivity and methodological pluralism. The evaluation framework may serve as reference for researchers, evaluators, policymakers and other stakeholders in designing and assessing gender equality interventions, and in further developing their evidence, and theoretical and methodological base.

1. Introduction

The issue of the underrepresentation of women in research and innovation (R&I) in decision-making and high status positions is well studied. Different policies and actions have been implemented to address the problem, but information on how effective these interventions have been is scarce (Kalev, Dobbin, & Kelly, 2006; Timmers, Willemsen, & Tijdens, 2010). Timmers et al. (2010, p. 722) point out that "the effect of gender equality policies has hardly been evaluated", while Kalpazidou Schmidt and Cacace (2017) call attention to the lack of evidence and the oversimplification of approaches in impact assessment of policy interventions. Systematic analyses of the interwoven linkages between gender equality (GE) policies and interventions, and outcomes and impacts in R&I are still lacking.

Thus, there is little known about the efficacy of the interventions, while the knowledge on the impact of GE interventions on R&I is even more restricted. This is mainly due to a range of issues that are closely related. First, the lack of understanding of the dynamics linked to such interventions in increasingly complex environments. Second, the lack of theory, and a starting point grounded solely in methodology (Chen, 2012). Third, the measurement problems, including lack of available data and of a wide-ranging adequate sets of instruments, which often lead to the use of simple impact indicators, i.e. mainly of quantitative character. Fourth, the timing issues that hinder timely data collection, robust analyses and conclusions. These are though only some of the issues linked to effect assessment of GE interventions.

However, in the recent decades, policy makers' expectations to evaluations and evidence-based policy have been growing simultaneously with an increased interest for impact assessment (Reale, Nedeva, Thomas, & Primeri, 2014). Concurrently, state-of-the-art research shows that linear casual relations between interventions and impacts are challenging to establish, due to the complex contexts involving a multitude of variables of uncertain character at multiple levels (Halpern, 2014; Kalpazidou Schmidt & Cacace, 2017). Hence, as Reale et al. (2017, p 298) point out "…speaking of 'attributable change' poses some problems, such as finding adequate tools and methods to measure impact, the time lag between the effect produced and the research activities that are supposed to have generated it, as well as the problem of disentangling the extent to which the research results were the sole or most significant causes of the effect produced".

The conceptual evaluation framework presented in this paper is an attempt to respond to the challenges that policy makers, scholars and evaluators face in complex, dynamic contexts that constantly adapt to changes. The framework adopts a holistic view with particular

* Corresponding author.
E-mail addresses: eks@ps.au.dk (E. Kalpazidou Schmidt), ekg@ps.au.dk (E.K. Graversen).

https://doi.org/10.1016/j.evalprogplan.2019.101750
Received 28 February 2019; Received in revised form 11 October 2019; Accepted 8 November 2019
Available online 12 November 2019
0149-7189/

sensitivity to complex and non-linear perspectives, which take into account the constantly emerging needs created by team, organizational and system dynamics. Grounded in the theory of change approach, the framework is set to make explicit assumptions on how change can occur in complex contexts. It was developed and tested in the frame of the Horizon 2020 project EFFORTI (Evaluation Framework for Promoting Gender Equality in Research and Innovation), which aimed at analyzing and modelling the impact of GE interventions on R&I and on establishing more responsible research and innovation (RRI)[1] systems. The main objective of the project was to provide an evaluation framework and a toolbox with instruments that are both sophisticated and practical for the design and evaluation of GE interventions in R&I. One of the main objectives of the EFFORTI project was to make a contribution to achieve the GE European Research Area (ERA) goals (European Commission, 2013) by offering a theoretical frame and instruments on how to design GE interventions and map their effects in R&I at different levels, i.e. team level (team structure, productivity, etc.), organizational/institutional level (research environment, recruitment capacity, resources, etc.), and system/policy level (regulations, funding, etc.) (Kalpazidou Schmidt et al., 2018).

2. Developing an innovative conceptual evaluation framework - moving beyond the state of the art in design and evaluation of GE interventions in R&I

2.1. The EFFORTI intervention logic

The point of departure for the EFFORTI intervention logic has been the GE objectives defined in the ERA. GE forms a crosscutting issue within the scope of the Horizon 2020 R&I funding program of the European Union, which promotes three objectives for fostering GE in the ERA: (1) increase the number of women in R&I, (2) increase the number of women in leadership positions, and (3) integrate the gender dimension in research content and curricula (European Commission, 2013). The *first* objective, more women in R&I, promotes full use of the human capital, which is expected to contribute to higher research performance. The *second* objective, more women in leadership positions, aims at a competitive global R&I economy by promoting the involvement of female scientists in leadership and decision-making positions. The *third* objective, integration of the gender dimension in research content and curricula, implies considering biological and growing social and cultural characteristics of both women and men throughout the research process to benefit society through the production of innovative research and technology advancement. The third ERA goal further requires that the gender dimension is integrated into teaching and curricula to address the needs of women and girls portraying of both genders in non-stereotypical ways, and making science and technology more attractive to all genders (European Commission, 2014; Schiebinger & Schraudner, 2011).

The EFFORTI intervention logic focused hence on providing concepts and tools for analyzing how GE interventions may contribute to the achievement of the three ERA objectives presented above. To start with the intervention logic operated with a simple ideal type effect theory of an intervention based on the concept of effects in terms of inputs, outputs, outcomes and impact, which facilitated the articulation of expected effects of interventions but which at the same time implied a simplification of a complex process, as no linear link between intervention and impact is possible to establish straightforwardly. Thus, establishing causal links between interventions and their effects posed a range of theoretical and methodological challenges (Cartwright & Hardie, 2012; Dahler-Larsen, 2012; Larsen & Lassen, 2001). EFFORTI addressed these challenges through the adoption of the complexity and gender-sensitive approach and the theory-based evaluation approach, as discussed in the following sections.

2.2. Complexity approach

Policies to address the underrepresentation of women in R&I have been initiated at different levels reflecting diverse approaches on how to address the issue, i.e. at individual, team, organizational (cultural and structural), or system level, most of them focusing on only one particular level. According to the *individual perspective*, having a gender-centered approach ascribes underrepresentation of women in R&I to differences between women and men as to psychological characteristics, socialization, and personal choices (Fagenson, 1990). GE interventions in R&I that explicitly target the *team* level are rare. However, in recent years the team unit has been recognized as an important arena for gender diversity, i.e. the proportion of women and men in a team and the associated fixed biological attributes of women and men, namely socialization processes and stereotypes that can be transmitted in different settings, i.e. family, educational or organizational contexts (Callerstig & Müller, 2016). The *cultural perspective* suggests that the lack of women in higher level positions is due to cultural and historical organizational conditions, but also due to cultural factors in society as a whole (Fagenson, 1990). Finally, according to the *structural perspective*, the obstacles to women's advancement involve organizational hierarchies, and formal and informal rules.

In contrast to the above-mentioned perspectives, Kalpazidou Schmidt and Cacace (2017) & 2019) propose a holistic perspective in the design and evaluation of GE interventions with integrated strategies based on the *complexity approach*. The key notion of complexity is that GE interventions (themselves being complex) are embedded in the complex, multilayered systems that they operate in. Such systems involve multiple variables interacting in non-linear ways to produce outcomes and impacts. Rogers (2008, p. 29) points out that there are important challenges linked to the evaluation of complex interventions and poses the question whether the models we use should "address the complexity of life". What characterizes the complexity paradigm is "emergent causality: multiple interacting influences account for a particular outcome but none can be said to have a fixed 'effect size'" (Greenhalgh & Papoutsi, 2018, Table 1). Moreover, complex systems respond to changes in the environment and adapt to new circumstances (Halpern, 2014). Design, implementation and evaluation of interventions need to take into consideration emerging issues with a focus on local dynamics and address them in a constantly changing context (Rogers, 2008). In complex contexts, impacts cannot be deterministically attributed to a particular intervention (Kalpazidou Schmidt & Cacace, 2017). Kalpazidou Schmidt and Cacace (2017, p. 103) further argue that "linear models and monodimensional approaches are not sufficient in effectively assessing the actual impact of GE programs or in adequately designing them in the first place" and point to the lack of studies attempting to consider the complexity of issues in GE interventions in R&I. As underlined in the literature, the ability of interventions to foster the right conditions for change is central in complex interventions (Reale et al., 2014), and impact assessment has to account for whether adequate conditions to achieve expected impact are in place (Kalpazidou Schmidt & Cacace, 2017).

Acknowledging the insufficiency of linear models to capture the complex linkages between inputs, outputs, outcomes, and impacts, the EFFORTI has adopted an approach which implies that rather than attributing notions of outcomes and impacts to interventions, these concepts are dealt with by means of evaluative approaches that pursue

[1] Responsible Research and Innovation (RRI) is an approach that anticipates and assesses potential implications and societal expectations with regard to research and innovation, with the aim to foster the design of inclusive and sustainable research and innovation. RRI implies that societal actors (researchers, citizens, policy makers, business, third sector organisations, etc.) work together during the whole research and innovation process in order to better align both the process and its outcomes with the values, needs and expectations of society.

intervention contributions to achieve impact (Kalpazidou Schmidt & Cacace, 2019).

2.3. Gender-sensitive approach

For a long time, the link between gender and evaluation has not been the point of attention in the literature despite an increasing demand for what is coined as gender-sensitive evaluation (Espinosa, 2013). Thus, the intersection between gender and evaluation has traditionally been studied by scholars coming from the evaluation field and to a limited degree by gender scholars. Accordingly, the majority of the publications comes from the general evaluation literature while the share of gender-sensitive contributions is limited (Espinosa, 2013, Seigart & Brisolara, 2002). As regards the field of policy studies, some phases in the policy cycle gained attention, such as policy formulation, agenda setting and mainstreaming implementation but the evaluation field as such has not attracted the gender scholars and hence the policy literature also stems from the evaluation field (Bustelo, 2017; Podems, 2010). However, there is a seminal contribution (Seigart & Brisolara, 2002) and a growing interest about evaluation among gender scholars has been noticed (Podems, 2010, Espinosa, 2013, Hay, Sudarshan, & Mendez, 2012; Brisolara, Seigart, & SenGupta, 2014; Bustelo, 2017). While gender-sensitive evaluation literature remains scarce, many international associations and organisations have active groups aiming at integrating the gender perspective into evaluation (i.e. American Evaluation Association, European Evaluation Society, German Evaluation Society) or have produced manuals based on practical work within mainly international development evaluation (cf. UN, UNWOMEN). There is though a lack of theoretical work on gender-sensitive evaluation despite the development of key ideas on evaluating gender mainstreaming and gender impact (Bacchi & Eveline, 2010; Bustelo, 2017; Squires, 2007).

What is gender-sensitive evaluation or evaluation from a gender perspective? According to Espinosa (2013: 174), "gender-sensitive evaluation seeks to discover the structural causes of inequality between women and men in the context of intervention". Gender-sensitive or feminist evaluation sees inequality as systemic and structural, and evaluation as a political activity (Bustelo, 2011; Seigart & Brisolara, 2002). According to Bustelo (2017) feminist policy studies and evaluation studies should be combined to help understand the structural aspects and how to evaluate policies from a gender perspective. How policy studies and the evaluation stage can contribute to each other to develop theories and practices from a gender perspective needs thus further elaboration and "the implications of what it means to do an evaluation from a policy perspective still remain to be explored thoughtfully" (Bustelo, 2017: 85). Bustelo (2017) discusses four crucial foci to the evaluation from a gender perspective by including the reflexivity element: the political nature of evaluation (evaluation as gendered), its public interest link (achieving social justice), the significance of involving the stakeholders (considering context and active participation) and the methodological diversity needed. Reflexive evaluation from a gender perspective applies a reflective and critical viewpoint during the whole process, questioning the neutrality of the evaluated policies and interventions. The above mentioned foci have implications as to methodological approaches in gender-sensitive evaluations in terms of considering: a) the need for gender, evaluation, context and sector expertise, b) the application of gender analysis and use of sex-disaggregated data, c) the use of self-reflection, context, stakeholders participation (also as to actions) and empowerment, d) the use of mixed methods, e) the communication of results (Bustelo, 2017).

The approach adopted in developing the herewith presented conceptual evaluation framework is consistent with the gender-sensitive evaluation approach. It moves beyond simple, mono-dimensional standpoints and adheres to the non-linearity of the process focusing on the structural and systemic factors, the context and stakeholder participation, also as regards their suggestions to empower-oriented actions, the use of gender-sensitive indicators (identifying and explaining manifestations of inequality invisible in traditional indicators) and a combination of methodological triangulation and quantitative and qualitative methods, and the sharing of evaluation results to raise awareness and contribute to organizational and political change. The self-reflexivity element and critical perspective throughout the whole process from design to evaluation are central to the conceptual evaluation framework and have guided the process of developing the herewith presented framework.

In conclusion, the complex character of inequalities suggest complex interventions to address structural and emerging issues. Complex interventions need to be evaluated in a holistic way and on the basis of a gender-sensitive and reflexive perspective with the involvement of all stakeholders and the use of methodological pluralism.

2.4. Theory-based evaluation approach

Theory-based evaluation is an approach where the assessed variables are selected according to a theory that formulates implicit or explicit assumptions about the interventions and the features expected to be important to achieving impact (Chen, 2012; Fitz-Gibbon & Morris, 1996). The two main components in theory-based evaluations are (i) the design of an intervention theory and the theory of change of a particular intervention, and (ii) the empirical investigation of the intervention theory.

The theory-driven approach used in developing the present framework is in contrast to "black box" or impact-oriented evaluation models that are restricted to the question, whether an intervention has achieved its intended impact or not (Döring & Bortz, 2016, p. 998). Theory-driven evaluations are also different from method-driven, atheoretical evaluations, where the design of an evaluation is usually determined by a specific method (Chen, 2012). Central in impact-oriented evaluation is a proof of causality: an intervention is only effective, if it is possible to verify that the observed positive effects would not be present without the intervention (Döring and Bortz, 2016). In contrast, theory-driven approaches are used to evaluate complex societal change interventions (Fulbright-Anderson, Kubisch, & Connell, 1998) and seek to map and explain the linkages between interventions, context, outcomes and impact by testing logic models (Douglas, Gray, & van Teijlingen, 2010). Theory-based evaluation models go thus beyond causality in an attempt to reconstruct impact mechanisms in a detailed way, involving all the stakeholders. Interventions are rooted in explicit assumptions made by stakeholders that allow us to understand what is done in the "black box" and why, what works or does not work and, in case of missing the intended effects, what needs adjustment, i.e. through context specific learning (Chen, 1990; Pawson & Tilley, 1997). Theory-driven evaluations focus hence on the questions (i) in which way and (ii) under which conditions an intervention causes the intended and unintended effects observed (Döring & Bortz, 2016). Such evaluations explore "not only whether the intervention works, but also how, for whom and in which context" (Van Belle, Marchal, Dubourg, & Kegels, 2010). Implicit assumptions of the stakeholders involved in the design and implementation about the links between interventions and their impacts are made explicit to highlight the mechanisms producing change through theories of change (Van Belle, Marchal, Dubourg, & Kegels, 2010). Lipsey and Pollard (1989) identify different mechanisms to make theory of change more explicit by involving policy designers, implementation teams, and the target group. Involvement of the stakeholders and implementation teams in developing the theory of change makes explicit the different kinds of stakeholders' assumptions, for example what implementers think compared to designers of interventions (Cole, 1999). This information is "…essential for stakeholders to improve their existing or future programs" (Chen, 2012, p. 17).

Theory of change works as "a systematic and cumulative study of the links between activities, effects and contexts of the initiative" (Connell & Kubisch, 1998, p. 16). The theory of change approach is

utilized in evaluations to make implicit theory explicit and hence to identify indicators of change that provide evidence of what works and why, in which contexts (Barnes, Matka, & Sullivan, 2003). It is a reflective process where key assumptions linked to a particular intervention are made explicit. Using evidence to identify, verify or challenge these assumptions and map the linkages between input, throughput, output, outcome, impact and context is part of developing a theory of change (Vogel, 2012).

Summing up, context plays a decisive role; paying explicit attention to context may guide us in our choice of evaluation approach on the one hand, and support us in understanding the plethora of contextual issues that affect a particular evaluation, on the other (Rog, 2012). Most evaluations operate in multiple contexts with numerous layers and dimensions (Greene, 2005). Rog (2012) discusses five areas that are characteristic to context in design and evaluation. These contexts comprise "the context of the problem or phenomenon being addressed, the context of the intervention being examined, the broader environment or setting in which the intervention is being studied, the parameters of the evaluation itself and the broader decision-making context" (Rog, 2012, p. 27). Dimensions of these contexts to consider in evaluations include "physical, organizational, social, cultural, tradition, political, and historical" aspects (Rog, 2012, p. 27). In our work in developing the framework, the key question guiding the process has been "which evaluation approach provides the highest quality evidence, in which contexts and for whom?" (Mark, 2001), with emphasis on the policy, organizational and team context, the needs of the stakeholders, the rigor of the practice (cf. Rog, 2012), and the adoption of a gender-sensitive perspective. Grounded in the complexity, nonlinear, gender-sensitive and theory-based evaluation approaches discussed above, a systematic process was initiated to develop the presented conceptual evaluation framework.

3. Methodological steps in the development of the conceptual evaluation framework

In this section, we present and discuss the methodological steps undertaken in the development of the conceptual evaluation framework. A number of frameworks and toolkit-like approaches are already available that address similar evaluation questions as the one discussed above, but these are limited to R&I and do not address GE in particular (Rhomberg, Steindl, & Weber, 2006; Fahrenkrog, Polt, Rojo, Tübke, & Zinöcker, 2002; Miles & Cunningham, 2005; White, 2009). Obviously, our work builds upon this knowledge but goes further in developing an innovative conceptual evaluation framework for GE in R&I. Drawing on a comprehensive desk research, a systematic step-by-step process was initiated to develop the conceptual framework presented herewith involving:

- A literature review and systematic mapping of the state of the art with an extensive collection of empirical studies and GE intervention evaluation studies.
- A mapping of contextual system conditions and country evaluation cultures.
- Collecting smart practices according to specific defined criteria.
- Developing a typology of interventions from the literature.
- Developing the initial conceptual evaluation framework.
- Identifying and developing indicators (quantitative and qualitative for the three ERA strategies and at different levels of intervention, i.e. micro, meso, macro).
- Initiating a validation process based on 19 case studies in six European countries, the theory of change approach and impact stories.
- Developing the final conceptual evaluation framework

Fig. 1 provides an illustration of the process of developing the framework. The eight-step procedure proved useful as it allowed for the set-up of a systematic, non-linear process towards the final conceptual framework, involving feedback and self-reinforcing loops to previous steps. In the following sections, we discuss the different steps in the process and their contribution to constructing the final framework.

3.1. State of the art and literature review

A state of the art mapping and systematic literature review of evaluations of GE interventions in R&I was the starting point of the work on developing the framework. Particular emphasis was given to mapping existing evaluation concepts of GE interventions and instruments, concepts for the measurement of research and innovation outputs and outcomes, approaches to impact assessment in R&I. The review also focused on previous projects undertaken within the EU 7th Framework Programme for Research and Technological Development and in the frame of Horizon 2020 (such as GARCIA, GENERA, Gender-NET, INTEGER, PRAGES, STAGES, etc.) to identify GE interventions, evaluation concepts, methodologies and instruments, and toolkits aiming at the study and promotion of GE in the ERA (e.g. Gender-NET IGAR tool, 2009; EIGE GEAR tool, 2016; GARCIA tool, 2015).

Moreover, drawing on already existing concepts and instrumentation in GE and innovation research (i.e. the Innovation Indicator, 2005; the European Innovation Scoreboard, 2016; the RIO Observatory, the OECD Science Technology & Industry Scoreboard, 2015; the OECD Science Technology & Industry Outlook, 2014), but also on recent studies on how to enlarge the conventional set of R&I indicators by considering innovative approaches, such as RRI (Ravn, Nielsen, & Mejlgaard, 2015; European Commission, 2015), a comprehensive desk research has been carried out as a basis for the collection of a preliminary list of concepts and relevant instruments to be scrutinized and further developed. As discussed above, the complexity and gender-sensitive literature, and in particular the self-reflexivity and critical perspectives, have guided this process.

3.2. Mapping country context and evaluation culture

The role of context in shaping evaluations is widely acknowledged (Lewin, 1943; Patton, 2008; Rog, 2012; Weiss, 1973). GE policies do not take place in a vacuum but are situated in specific national contexts, comprising legal regulations and policies that are formed by socio-cultural factors (Schiffbänker, 2009). Interventions are hence embedded in different contexts with regard to socio-economic and political systems, and organizational settings, and thus unfold and function in a constant interaction with their contexts (Arnold, 2004; Edler et al., 2010; Streicher, 2017). In order to map the national context, seven country notes have been produced within the EFFORTI project, one for each country participating in the project (Austria, Denmark, France, Germany, Hungary, Spain and Sweden) with the aim to identify the most significant conditions influencing GE interventions and their evaluation in R&I. Contextual conditions, such as the structure and performance of the R&I systems, GE in the labor market, GE and welfare policies, and the governance of GE in R&I, have been described among others (Reidl et al., 2018).

Besides the contextual conditions, the effects of GE policies in R&I depend upon the quality of the intervention design and its implementation. Quality can be improved through the monitoring and evaluation of the intervention, as this allows policy learning and feedback to the design and implementation process (Biegelbauer, 2013). The advancement of intervention design and impact assessment at system level is therefore closely linked to the degree of development of the evaluation culture in a particular country, which also influences interventions and their evaluation at the meso level. For that reason, key questions explored in the frame of the EFFORTI project comprised systemic and country-wide evaluation issues, such as: (i) are there explicit rules and legislation on evaluation of GE in R&I in place? (ii) are evaluations considered as part of a broader system to ensure

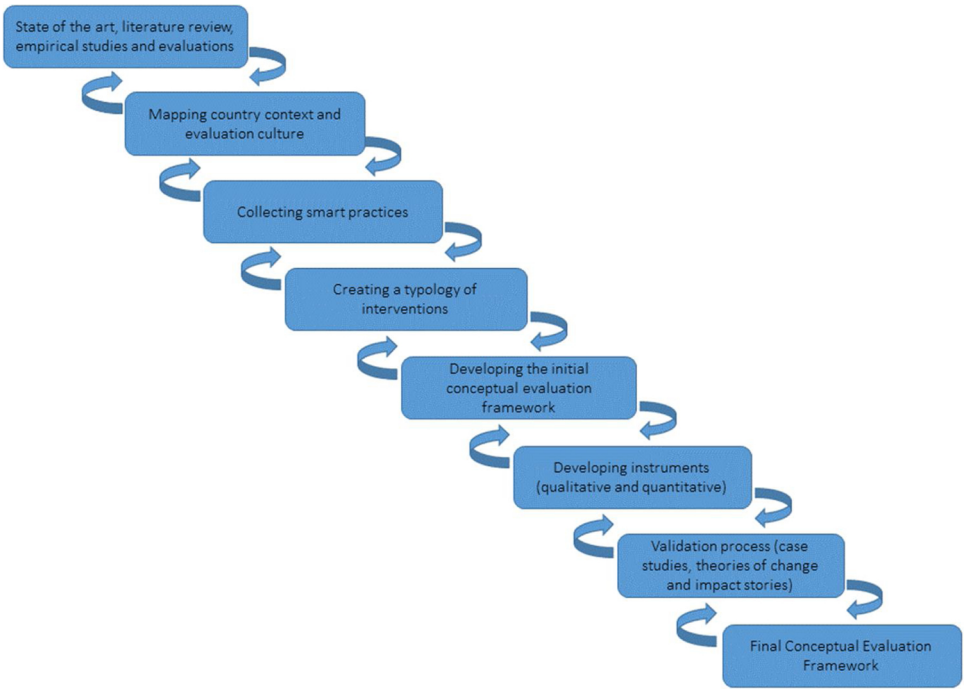

Fig. 1. A systematic process to develop the EFFORTI conceptual evaluation framework.

accountability? (iii) has institutionalization of evaluation taken place? (iv) is evaluation the exception rather than the rule? and (v) how are evaluation outcomes utilized in R&I? To address these questions, the country reports mapped evaluation cultures in all the countries involved in the EFFORTI project (Reidl et al., 2018). Not surprisingly, the mapping of evaluation cultures revealed that in countries where a stronger evaluation cultures exists (such as Austria, Denmark, Germany and Sweden), GE interventions were more comprehensively evaluated than in countries with weaker evaluation traditions such as Spain and Hungary.

3.3. Collecting smart practices

The tool of collecting smart practices of evaluations of GE interventions in R&I was employed to support the work on both the framework conceptualization and instrumentation. The collection of existing knowledge and practices of initiating and evaluating GE in R&I interventions took place through a review of relevant research publications and a range of evaluations of GE and/or R&I interventions. The review also considered progress and results of previous projects undertaken within the EU funding programs (i.e. GARCIA, GENERA, Gender-NET, INTEGER, PRAGES, STAGES, etc.) mapping concepts, methodologies and instruments.

The identification of smart practices was based on assessments of whether they were relevant, effective and efficient in the context they operated in as to their quality of both intervention and evaluation (Kalpazidou Schmidt et al., 2018). The collected smart practice examples evaluated interventions of different nature and length: some constituted large national interventions with a long-term perspective, while others were of a more limited character. Some of the smart practices aimed, for example, at evaluating the policies to mitigate gender bias and promote an inclusive culture that values all staff (as in the Athena SWAN program)[2], or mapping the visibility of female accomplishments in science (as in Laura Bassi Centres of Expertise)[3], or the participation and advancement of women in academic science and engineering careers (as in ADVANCE IT)[4], or the gender integration in university leadership (as in the AKKA program)[5] . Hence, the collected smart practices covered a wide range of GE interventions in R&I, and contributed therefore with examples of the most common GE evaluation types described in the literature.

The collected smart practices demonstrate understandings of and experiences with approaches, methodologies or techniques that aimed at studying the link between GE intervention designs and their outcome and impact, implemented by different actors in diverse contexts. Thus, the collected smart practices were relevant, effective and efficient in the context they were embedded in. However, lessons learned may be used in other contexts. Evaluation literature understands smart practices as practices that allow us to learn from others and produce considerable synergies, facilitating successful, innovative and effective evaluation practices that provide orientations for the development of new innovative tools. A smart practice is an evaluation that takes contextual and systemic factors into consideration when assessing an intervention. Finally, a smart practice demonstrates a reliable and consistent evaluation of positive or negative effects of interventions. The selection of the smart practices was based on the criteria of (i) the quality of the implemented interventions, and (ii) the impact of the interventions.

The quality of the interventions was assessed based on the parameters of relevance (the adequacy of the initiatives included in the intervention to the situation of the team/organization/system in which they are conducted and/or the wider social, cultural and economic contexts of reference for R&I and RRI in the countries involved); effectiveness (the capacity to implement the intervention according to stated objectives and attain the objectives outlined in the design of the

[2] For more on the Athena SWAN, see https://www.ecu.ac.uk/equality-charters/athena-swan/, Fehmidah et al. (2013); Kalpazidou Schmidt, Ovseiko, Henderson, & Kiparoglou, 2019; Ovseiko, Chapple, Edmunds, and Ziebland (2017) and Additional file 1.

[3] See https://rio.jrc.ec.europa.eu/en/library/laura-bassi-centres-expertise, Heckl and Dörflinger (2014) and Additional file 1.

[4] For more on the ADVANCE program see https://www.nsf.gov/crssprgm/advance/, ADVANCE IT Laursen, Austin, Soto, and Martinez, (2015)) and Additional file 1.

[5] See Lund University (2010); Lövkrona et al. (2006); Lövkrona et al., 2008; Lövkrona & Widén, 2012 and Additional file 1.

activities); efficiency (ability to make the best use of available resources, complying with the timeframes and procedures contemplated for expenses in the context of good managerial capacity); and sustainability of the interventions (capacity of the measure to continue to produce effects even after the end of the intervention).

Impact can be used as an overall term containing assessments of performance, effectiveness, efficiency, output, outcome, along with medium and long-term effects or as a narrower concept having a limited scope (Kalpazidou Schmidt & Cacace, 2017; Hansen & Jørgensen, 1995). Impacts of interventions can have effects at different points in time (short-, mid-, and long-term), vary as to range (at the direct level of participants or at the indirect level beyond the targeted objects) (Miles & Cunningham, 2005), and be intended and/or unintended[6] . Finally, there are different types of impacts, i.e. scientific, economic, socio-cultural, environmental, i.e. (European Commission, 2005; European Commission, 2009; Horvat, 2011), or constitutive effects[7] (Dahler-Larsen, 2012; Dahler-Larsen, 2014; Dahler-Larsen & Krogstrup, 2001).

The impact of the smart practice interventions is assessed in relation to its subjective and objective dimensions. Subjective impact addresses the satisfaction of the targeted beneficiaries of the intervention (as well as the capacity to promote consensus among other stakeholders involved in the intervention). Subjective impact indicators could include, for example, job satisfaction or improved work climate. Objective impact refers to the effects obtained in terms of real change due to the intervention, which may be expressed in numerical/quantitative terms (such as an increase in the proportion of women in senior and decision-making positions), but it may also be of a cultural, structural or policy character, expressed in qualitative terms. The latter involves change in policies or work procedures of institutions implementing the intervention or a change in the organization's ability to generate innovation that reflects societal needs (Kalpazidou Schmidt & Cacace, 2017).

Having the above-mentioned criteria as a point of departure and mapping the existing practices as regards context, objectives, methodology, instrumentation, and impact, the smart practice evaluations were identified and indicators for the three levels (micro, meso, macro) were developed that fulfilled the following conditions: As to the intervention itself, the smart practice examples (i) met the needs and priorities of the program initiators, (ii) achieved their initial goal (effectiveness) with the allocated resources, (iii) had a demonstrable impact, (iv) were sustainable, i.e. the results were maintained after the intervention was completed, and (v) had learning potential for other researchers, organizations and policymakers. As to the evaluation, the smart practices (i) built on a variety of evaluation concepts of GE interventions and instruments, (ii) framed a variety of different concepts for the measurement of R&I effects, (iii) represented a variety of impact assessment approaches, (iv) were exercised within diverse R&I frameworks, and (v) were characterized by various theoretical or methodological approaches.

[6] The distinction between intended and unintended effects presupposes that identifiable policy intentions have been clearly carved out beforehand, which is not always the case, especially at the level of policy-making (Dahler-Larsen, 2014).

[7] The notion of constitutive effects is analytically derived from a meta-level of abstraction and refers to the way in which the very acts of measuring, evaluating and using specific indicators affect practice. Evaluative practices may contribute to a reconfiguring of interpretive frames and worldviews. More specifically, constitutive effects emerge in different domains related to: (1) the content of what is measured, e.g. as a risk of a skewed notion of what is "central," (2) the way time is dealt with, since impact assessment naturally must identify a specific timeframe, and (3) social relations and identities in that the mechanism of labelling – and often comparing – units of analysis such as teams, institutions or individuals may contribute to the social (re-)construction of these units. See also Quantitative analysis as narrative (Stone, 2016) for more on the constitutive mechanism of all quantification.

3.4. Creating a typology of interventions

In advance of the development of the evaluation framework, concepts, indicators and methods useful for assessment of GE interventions in R&I, a mapping of existing types of GE interventions was carried out and presented as a typology to build the subsequent work of the case studies. An initial typology, developed by Kalpazidou Schmidt and Cacace (2017), assessing more than 120 GE programs in research organizations worldwide, was extended based on the outcome of a dedicated workshop, to include further interventions that also promoted the third GE ERA objective, i.e. inclusion of gender in research content and curricula. This typology was then related to the fields of action promoted by the GENERA[8] project. The produced typology (Table 1) was then used in connection with the case studies and helped our cross-case analysis in three ways. Firstly, by facilitating the grouping of interventions with similar objectives. Secondly, the types of interventions were linked to the GE ERA priorities. For example, recruitment has been linked to more women in R&I, advancement was linked to an increased gender balance in decision-making and leadership. Finally, as part of the quality assurance process the intervention typology supported the selection and construction of the large gross stock of instruments to be used in evaluations of GE interventions in R&I. Despite the fact that the typology constitutes an extensive gathering of different interventions, all intervention types and formats are hardly covered in the table as new types of interventions are constantly being developed. The typology was intended as a starting point for a dynamic update along and beyond the process.

Based on the intervention typology, impact stories (marked with bold in Table 1) were developed for a broad spectrum of the different intervention types in order to provide examples of the mechanisms regarding intervention intentions and offer a common framework for understanding the 19 case studies that were carried out in the six European countries involved in EFFORTI. These analyses served as a testing ground for the verification and further development of the framework in the final phases of the process.

3.5. Developing an initial conceptual evaluation framework

Based on the knowledge and data gathered through the first steps in the process, an initial conceptual evaluation framework was created to be refined through the case studies and stakeholders consultations. The illustrative tree presented below (Fig. 2) is a visualization of the initial conceptual evaluation framework and was produced to illustrate the complexity of the concept that was aimed to be tested and validated during the following steps. The three ERA GE strategies stand as the vision and foundation of the tree and constitute the platform for all interventions. The contextual aspects of the initial framework are illustrated by the means of weather icons and refer to the varying structural and cultural features pertaining to any specific intervention. Contextual elements may pertain to country, type of research system, sector type (i.e. public or private), position of the organization in the R&I system, type of GE intervention, and previous experiences with policy interventions of similar or dissimilar types. The fruits of the tree, harvested through theory-driven evaluation, represent the positive outcome of the intended gender-equal R&I system and involve outcome and impact of interventions such as innovation, patents, publications, funding, knowledge dissemination, science communication, research-based teaching, and societal impact, among others. A bird is pictured symbolizing the different viewpoints or perspectives available in the conceptual framework for a variety of target and stakeholder groups with different interests (policymakers, funding agencies, NGOs, the business sector, research organizations, evaluators, etc.). In addition, the initial conceptual framework enabled the development of a set of instruments

[8] See GENERA https://genera-project.com/.

Table 1
Overview of the GE intervention typology (bold = intervention covered by an impact story).
Source: Developed form Kalpazidou Schmidt & Cacace (2017).

Type of intervention	Intervention format
1. Policies	1. Mainstreaming actions
	2. Gender Equality/Action Plan
	3. Gender budgeting
2. Non-discrimination	4. Gender-sensitive practices for the attribution of tasks
	5. Gender-sensitive study and working conditions (e.g. alternative study plans for pregnancy during laboratory work period)
	6. Gender-sensitive HR management
	7. Guidelines regarding gender specifics
	8. Definition of targets regarding gender balance in decision-making positions
3. Composition & Integration	9. Definition of targets regarding gender balance in research groups
	10. Institution of quotas
	11. Mentoring programs
4. Advancement	12. Gender-sensitive practices for assessment
	13. Introduction of chairs and positions reserved to women
	14. Support to career development (counselling)
	15. Empowerment schemes
	16. Campaigns for inspiring women for MINT[1] subjects
5. Recruitment	**17. Monitoring appointments, promotions, or attributions of tasks**
6. Monitoring	18. Revision of internal policies regarding promotions
7. Deconstructing Excellence	19. Revision of internal policies regarding staff appointments
	20. Training courses (different targets)
8. Gender Awareness & Bias	21. Implementation of gender-sensitive leadership and personnel development
9. Leadership Accountability	**22. Targeting funding practices to improve women's access to research funding**
10. Funding	**23. (Targeted) funding to improve the integration of gender dimension in research**
	24. Targeted funding practices to encourage research organizations to promote gender equality measures
	25. Special funding for women researchers
	26. Gendered user involvement
11. Research	27. Inclusion and monitoring the integration of the gender dimension and impact
	28. Dissemination of information material
12. Knowledge	29. Revision of teaching curricula and texts
	30. Introduction of single-sex degree and specialization courses
	31. Provision of Gender and Women Studies or modules
	32. Integrating the Gender Dimension in Tertiary Education
13. Visibility	**33. Networking**
	34. Activities to make women (and their research) visible (e.g. introduction of awards reserved for women)
	35. Role models
	36. Support in period of absence for family needs
14. Care & Family Life	**37. Schemes for women returners**
	38. Care services and facilities (for children, the elderly and others)
	39. Support to mobility, including spouse relocation schemes
15. Work-Life Balance	**40. Introduction of flexible working hours**

[1] Mathematics, Information technology, Natural sciences and Technology.

to measure the effects of GE interventions in R&I. The branching of the tree illustrates thus the instrumentation proposed, which corresponds to a categorization of the offered tools, i.e. the division of the instrumentation into categories, each with dimensions and concrete indicators, embodied by the green color and the leaves of the tree. In line with the gender-sensitive evaluation and theory of change approaches, which honor methodological pluralism (Bustelo, 2017; Pawson & Tilley, 1997), the developed indicators are both of quantitative and qualitative character.

3.6. From concepts to instrumentation and indicators

As discussed above, the initial framework guided the collection and development of relevant instrumentation while the case studies and the related impact stories were used to verify, refine and further develop the evaluation framework. The indicators cover all three levels of policy interventions, namely micro (centering on individuals or teams), meso (focusing on organizational issues such as institutional rules, incentives, organizational structures, cultures and processes), and macro (referring to rules, incentives, structures, and processes at regional, national or supranational level). In practice, the distinction between micro, meso and macro levels may not be entirely clear-cut, since the levels are interrelated and many indicators can be employed at more than one level.

Indicators aiming at assessment at all stages (inputs, throughputs, outputs, outcomes and impact) of an intervention are built-in. The indicators are obviously not mutually excluding. From a dense list of approximately 692 indicators, five distinct categories have been identified and presented in a scheme to support future evaluations of the link between GE interventions and R&I effects. The five main categories identified are: (i) Personnel, (ii) Working conditions, (iii) Professional capabilities, (iv) Structural features, and (v) Research and Innovation, and Responsible Research and Innovation (Kalpazidou Schmidt et al., 2018).

As the indicator scheme with the full instrumentation list was rather detailed, a synthesis table was constructed to provide an overview of the main categories included in the initial evaluation framework, each divided in dimensions, subdimensions and indicators with a decreasing level of detail (see Table 2, for a more elaborated presentation of the indicators see Kalpazidou Schmidt et al., 2018). Thus, 21 dimensions and 46 meaningful subdimensions have been documented based on the best practices, stakeholder consultations and literature suggestions. It is, however, evident that the herewith presented list serves as a schematic overview of possible instrumentation for use in future evaluations of GE interventions in dynamic R&I contexts. This implies that the scheme cannot be exhaustive or static. Consequently, future evaluations of GE interventions in R&I may well, depending on the context of the intervention, the setting, and the context for decision making (Thurston, Smith, Genskow, Stalker Prokopy, & Hargrove, 2012), revise the scheme and include additional instruments.

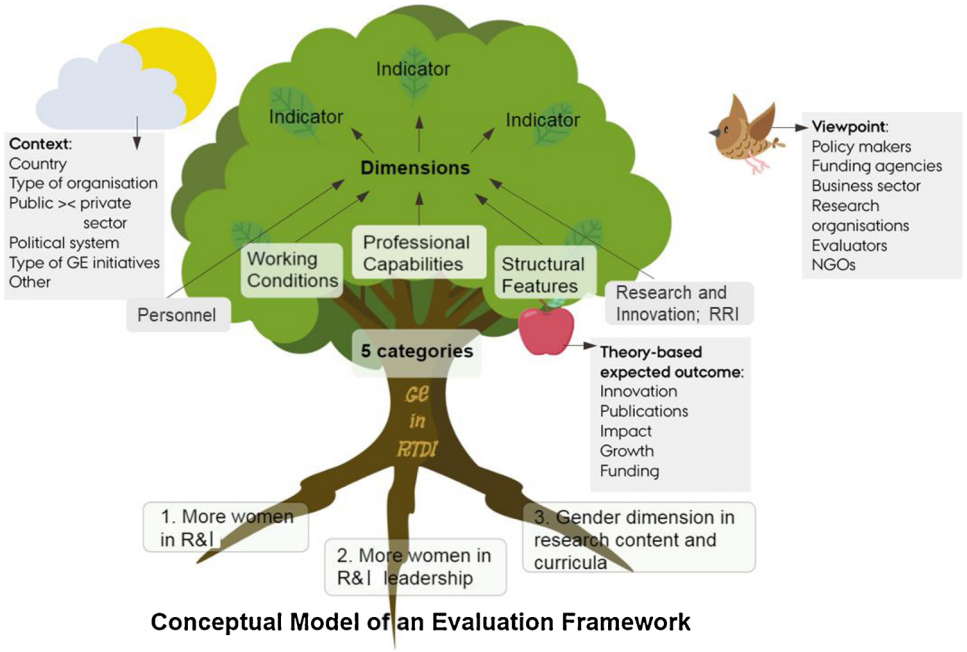

Fig. 2. An illustration of the initial conceptual evaluation framework.
Source: Kalpazidou Schmidt et al. (2018).

3.7. The validation process

The seventh step in the systematic process of developing the conceptual evaluation framework involved a validation process, which was grounded in the theory of change approach, the case studies and the related impact stories. The theory of change has been used to identify possible GE and R&I effects, to examine how change has happened and under which conditions (Mayne & Johnson, 2015; Rog, 2012). For each case study, a theory of change was thus developed which was based on three main axes: a) concept analysis, b) implementation analysis, and c) effect assessment. For the a) concept analysis, we conducted semi-structured interviews with programme managers and owners as well as content analysis of programme documents and evaluations already available, monitoring reports or reviews. The main objective of this step was to identify how an intervention was set up and expected to achieve its objectives and results. Based on the collected data an impact story for each intervention was developed and tested in order to test the theory of change and its underlying functioning. For the b) implementation analysis, we deployed a combination of quantitative and qualitative methods comprising semi-structured interviews with practitioners, desk research as well as a social network analysis. In this step of evaluation we focused on the question whether initiatives have been implemented and managed effectively and efficiently to achieve their objectives. For the c) effect assessment we identified outputs, outcomes and impacts in relation to GE and R&I for the different types of interventions. Distinct reports for all case studies were produced, containing the findings for the different instruments and levels of analysis and providing a first assessment of the effects, efficiency and impacts of each intervention. The reports were then made available to the programme managers and owners of the respective interventions in order to receive their feedback and validate the findings.

The process of developing a theory of change for each case study has involved all types of stakeholders, i.e. program managers, policy makers and beneficiaries (between 6 and 12 stakeholders for each case study were involved in the process). As mentioned above, stakeholders were requested to validate that configurations developed accurately explained expected effects while an impact story was created for each case study[9] . The impact stories have been constructed as ideal type logic models to articulate outputs, outcomes and impacts of GE interventions. The impact stories clarify hence (i) how the elements of input, output, outcome, and impact interact with each other in a specific context, (ii) through which indicators the expected effects can be verified, and (iii) which positive as well as negative, unintended effects have to be taken into account, and how they can be fostered or, respectively, avoided. The impact stories set up the basis for further refinement of the study of the linkages between inputs, output, outcomes and impacts. Thus, in terms of methodology, the impact stories are not only central to the conceptual evaluation framework, they also constitute the foundation for the verification of the effects of the case studies. The impact stories were thus instrumental in pointing out which of the set of elements that was compiled during the previous steps in the process would be included in the final framework (Palmén et al., 2018).

3.8. The final conceptual evaluation framework

Following the completion of the evaluations of the selected case studies, a critical reflexion process on two levels was hence initiated: a) at the methodological level: the experiences gained during the execution of the evaluations and while applying the initial evaluation framework were documented throughout the work done during the validation and described in the case study reports (Palmén et al., 2018). Detailed research and methodological notes were kept and used to analyse the feasibility, reliability and adequacy of the proposed conceptual evaluation framework. A comparative analysis of these was carried out, comprising a reflection on issues of attribution and contribution, lack of available information, data and instruments, time-lags, and outcomes and impact as context dependent. The conceptual evaluation framework was thus validated through three ways: the theory-based evaluation/ theory of change approach, the case study work that fed into the impact story work, and the validation of more than 251 out of 692 indicators; b) at the content level: based on the evaluation results, a comparative approach between all the evaluated

[9] See Additional file 2 for an example of an impact story.

Table 2
EFFORTI GE and R&I instrumentation by category, dimension and subdimension.
Source: Kalpazidou Schmidt et al. (2018).

Category	Dimension	Subdimension
1. Personnel	**1.1 Positions**	1.1.1 Increased number of women in academic and other R&I positions
		1.1.2 Increased number of women in decision-making positions
	1.2 Recruitment capacity	1.2.1 Improved recruitment of talented women
2. Working Conditions	**2.1 Work-life balance**	2.1.1 Improved compatibility of family and career
	2.2 Job satisfaction	2.2.1 Appropriate respect/recognition for (academic/scientific/leadership) work
		2.2.2 Positive individual job rating
		2.2.3 Overall work climate
		2.2.4 Allocation of workload
	2.3 Competitiveness/promotion and career	2.3.1 Transparent, non-biased and flexible promotion/tenure criteria
		2.3.2 Strengthened confidence for promotion and responsible positions
		2.3.3 Improved support to advance research career
	2.4 Workplace	2.4.1 Equal workspace/facilities allocation
3. Professional Capabilities	**3.1 Leadership**	3.1.1 Increased confidence and ability of leadership roles
	3.2 Professional achievements	3.2.1 Increased professional development of work skills (for career success)
		3.2.2 Improvement of network building and use
	3.3 Awareness of/commitment to GE	3.3.1 Increased gender awareness
	3.4 Funding to promote GE in terms of female careers	3.4.1 Increased funding to promote GE
4. Structural Features	**4.1 GE challenges/barriers**	4.1.1 Decrease of GE barriers
	4.2 Organizational/cultural change with regard to GE	4.2.1 Organizational/cultural change with regard to GE
	4.3 Preferential treatment	4.3.1 Equal treatment
	4.4 Funding for structural transformation	4.4.1 Increased funding to achieve structural transformation
5. R&I / RRI	**5.1 Research outputs and impacts**	5.1.1 Scientific outputs
		5.1.2 Networks
		5.1.3 Training/Human capital
		5.1.4 Strengthened R&I capacities/excellence
		5.1.5 Research priorities and outcomes in terms of GE
	5.2 Innovation outputs and impacts (incl. technological impacts)	5.2.1 Conventional input-output based innovation measures
		5.2.2 Diffusion of innovation in products, services, processes
		5.2.3 Knowledge about sex and gender incorporated into engineering innovation processes
	5.3 Economic outputs and impacts (incl. entrepreneurships)	5.3.1 Economic impacts
		5.3.2 Entrepreneurship
		5.3.3 Strengthened framework conditions for R&I
		5.3.4 Jobs, growth & competitiveness of participants (incl. small and medium enterprises (SMEs))
	5.4 Gender-sensitive research	5.4.1 Achieved gender equality in research process
		5.4.2 Research quality: integration of a gender dimension/perspective in research and content, in research projects, patents, and agreements
		5.4.3 Contributions to strengthening gender-sensitive research
	5.5 Responsible Research and Innovation (RRI)	5.5.1 Gender equality
		5.5.2 Ethics
		5.5.3 Public engagement
		5.5.4 Science education
		5.5.5 Open access
		5.5.6 RRI/governance
	5.6 Societal challenges	5.6.1 Research priorities & outcomes in terms of GE
		5.6.2 R&I indicators assessing synergies between R&I and nature, society and economy
	5.7 Societal and environmental impacts	5.7.1 Societal impacts
		5.7.2 Environmental impacts

case studies was applied to study whether interventions have been implemented effectively and display intended results. In addition, generic lessons learnt based on comparing all evaluation case studies were taken into account in developing the final framework (Palmén et al., 2018).

The different steps of creating the conceptual evaluation framework and identifying the tools for designing and assessing GE interventions in R&I resulted in the design of the final framework (see Fig. 3), which constitutes the entirety of the experiences and knowledge gathered during the process. The non-linear approach used implied a reciprocal relationship between the eight steps in the process, involving interactive dynamic procedures rather than standalone steps as illustrated in Fig. 3. Feedback loops to previous steps with "self-reinforcing" loops were employed, which together with the stakeholder consultations and the 17 workshops organized in different European countries, provided valuable knowledge that has been utilised to validate and refine the initial framework (see Fig. 2). As seen in Fig. 3, the evaluation framework considers both GE effects and R&I effects because the basic working assumption has been that GE interventions in R&I have effects that go beyond GE itself by affecting the quality and quantity of R&I. Effects are therefore presented both in relation to GE and to R&I and positioned in parallel boxes in the framework illustration as interventions effects regarding GE are expected to have direct or indirect R&I effects. Positive GE effects start a self-reinforcing process in R&I (not least as regards the third GE ERA objective, integration of the gender dimension in research). Both types of effects may need time to be manifested. On the other hand, the complexity of both context and interventions, but also design and implementation issues, may create unintended GE effects or cause evaporation of GE and thus of R&I effects.

Evidently, the eight step dynamic process presented above is not finalized with this work but is intended to function as a stepping-stone to further develop the framework, adding new knowledge, theories, practices and instrumentation to address emerging issues, providing

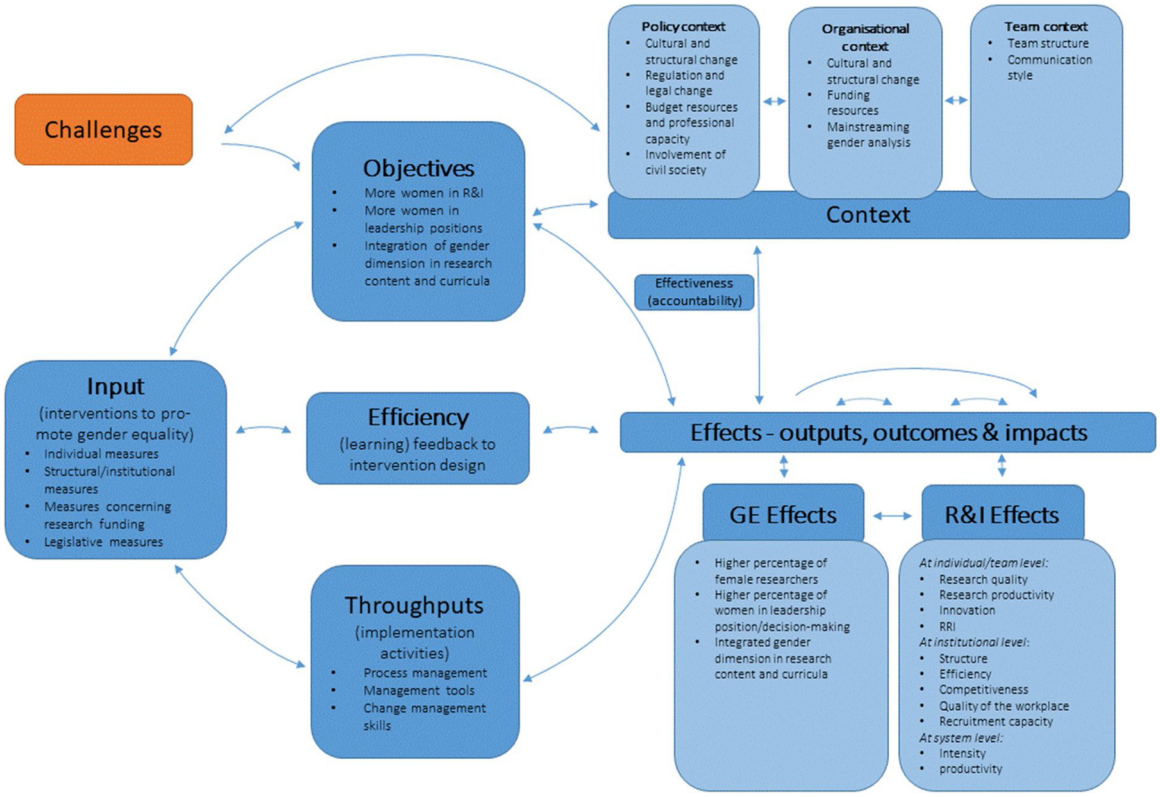

Fig. 3. A conceptual evaluation framework for gender equality interventions in R&I.
Note: In each box, a non-exhaustive number of content examples are listed. Dependent on the intervention type and context (cf. Table 1), other content and instruments (cf. Table 2) may be relevant to include and employ.

learning opportunities, as illustrated by the two-way "interacting" arrows, linking the different elements in Fig. 3. The structural and dynamic character of the inequalities in R&I have recent years led to ever-complex interventions, which require increasingly complex evaluation approaches and call for self-reflectivity and a critical standpoint, acknowledging biases and limitations of evaluations and honoring multiple perspectives (Patton, 2008, Bustelo, 2017).

4. Discussion: Implications and constraints of the adopted approach

Overall, with the above presented conceptual evaluation framework, we opened the "black box" of the relationship between GE interventions and R&I effects and provided approaches and tools to explore the black hole of context, pointing out aspects important for the design and evaluation of interventions (cf. Rog, 2012). The main aim has been to contribute to the understanding of "why, where, and for whom programs work or fail to work" (Rog, 2012, p. 35) and identify the mechanisms and contextual factors that may define impact. GE interventions are part of policy systems, and context is subject to change due to intended and unintended effects of interventions but also due to other activities taking place outside the interventions (Barnes et al., 2003). In addition, as LaFrance, Nichols and Kirkhart (2012, p. 72) underline, "considerations of both evaluation context and setting profoundly influence method choice and implementation. The importance of putting context ahead of method choice cannot be overstated". Complexity, pointing out the deficiency of the linear model, defines the study of the linkages between input, output, outcome, and impact. To address these issues the following points have been considered in the framework:

First, the conceptual framework embraces a holistic perspective to widen the considered impact areas beyond merely traditional quantitative techniques to include qualitative approaches as well because advances in structural and cultural changes, as the ones the complexity standpoint pursues, are seldom quantifiable and need instrumentation that enables tracking behavior and attitude change (Moser, 2007). This is made to include in the analysis theoretically sound inferences on the link between additional background factors and the dynamics of GE interventions, and the ability of interventions to produce effects.

Second, our approach adopts the notion of 'conditions for impact' (Reale et al., 2014) and takes a probabilistic stance in developing the evaluation framework and relevant instrumentation, in order to enable assessing of the middle and long-term effects of GE interventions in R&I in a more substantive and less deterministic way (Kalpazidou Schmidt & Cacace, 2017).

Third, in the development of the framework, the stakeholders have been involved from the beginning of the process in order to learn from local experiences and support utilizing contextually sound and gender-sensitive evaluations (Espinosa, 2013, Bustelo, 2017). Diverse stakeholders' voices (through three national workshops organized in every partner country and three international workshops) have been important in identifying local issues and the most effective practices in responding to them (in particular in connection with the case studies and in developing the impact stories).

Fourth, the above discussed contextual conditions and choices as to evaluation design and approaches may characterize other fields, different from the field of GE in R&I. Thus, the conceptual framework developed in this study may be an inspirational source for policy makers and scholars working with complex interventions within complex contexts in other study fields.

Still, the conceptual evaluation approach presented in this paper faces a range of challenges. Although it is based on existing evidence and new theoretical and empirical trends in evaluation, and is further developed based on the complexity and gender-sensitive approaches, it reflects the corresponding limitations and constraints in the evaluation field. In the following, we comment on the limitations of the approach

proposed herewith and how to mitigate them.

A crucial challenge in complex GE interventions is the issue of establishing *attribution*, i.e. attribute outcome and impact to a particular intervention. As Reale et al. (2014, p. 37) underline 'studies of impact assume that the relationship between the "impactor" and "impacted" is fairly direct and by necessity ignore the fact that in reality the social space is "noisy" and there are many intervening factors/variables'. In a similar vein Martin (2011, p. 250) states that impact is often 'indirect, partial, opaque and long-term'. As discussed above, our holistic perspective, exploring the conditions for impact, and probabilistic stance, including contribution, not attribution analysis may help mitigate this challenge.

Another key challenge is linked to the character and constitution of impact, which may vary according to the particularities of the *sector* or *field* (Martin, 2011). Institutional and national contexts are decisive in affecting impact. "This means that the same strategies and solutions may generate different impacts dependent on context, since the key variables in place, and their mutual relationships, are rarely the same" (Kalpazidou Schmidt & Cacace, 2017, p. 110). The number and diversity of stakeholder groups from different contextual settings are a critical part of the evaluation process in addressing this particular challenge. Stakeholders are not a homogeneous group but have different interests, values, assumptions and perspectives. Thus, involving them from the beginning of the evaluation process is crucial for intervention design, implementation, outcome, impact, and further policy making (Thurston, Smith, Genskow, Prokopy, & Hargrove, 2012).

Moreover, assessing impact may suffer from *lack of data and indicators*. Scholars have pointed out the need for more sophisticated frameworks and methodological diversity and suggest to go beyond traditional impact indicators and identify less tangible impacts (Bell, Shaw, & Boaz, 2011; Molas-Gallart & Tang, 2011). Our approach suggests that data collection and analysis are conducted through rigorous procedures, and based on systematic literature studies and a probabilistic stance, moving away from traditional quantitative measures, to carry out more sophisticated analyses. Another crucial question is the adequacy of the instruments and whether evaluators measure the correct object. Involving stakeholders (i.e. program managers, policy makers, evaluation experts) from the beginning and throughout the evaluation process may help address this issue (cf. Bustelo, 2017).

Another key challenge involves the measurement of the *scale and intensity* of impact, which may display great variation. Moreover, as mentioned above, not all impacts generated from particular interventions are *intended*, anticipated or desirable. The presented conceptual evaluation framework suggests a broadening of the approach in assessing impact beyond the mere traditional, including the multiple processes that interventions are embedded in. A theory-based evaluation and an intervention logic based on explicit assumptions may support the understanding of the complex dynamics and linkages between inputs and effects.

An issue often discussed in the literature is the time lag, i.e. the *time* span between a particular intervention and the assessment of impact, and the long-term perspective of impact (Reale et al., 2017). This challenge is thus linked to timing of impact assessment. If the assessment is carried out almost immediately or too late after the finalization of the intervention, stakeholders may not link the achieved effects to the intervention itself (Bell et al., 2011). Moreover, data may need collection over a longer period so that rigorous and robust impact assessments can be realized.

In addition, as pointed out by Kalpazidou Schmidt and Cacace (2017, p. 111) "assessing impact of societal interventions in general is a greater challenge, as there often are limited quantitative data available, and a *lack of consensus* on what data to collect, since there is a variety of stakeholders that try to promote their interests" (cf. Spaapen & Van Drooge, 2011). A theory-based evaluation model may guide the data collection process and help address this challenge.

Finally, the *gender sensitiveness* of our approach, at the same time as it addresses a series of critical issues, pays attention to a range of additional potential challenges, such as the risk that funders do not acknowledge the political character of evaluation (the context in which evaluations operate is politicized and evaluators perspectives imply particular political standpoints) (Seigart & Brisolara, 2002); the lack of political will to make use of evaluations; the limited institutional capacity on gender and gender-sensitive evaluation design and implementation (which may lead to resistance) (Espinosa, 2013); the misleading interchange between gender and women (may lead to counting heads of women instead of addressing structural gender issues); evaporation of the gender dimension during the implementation (Moser, 2005); lack of paying attention to findings not only useful to funders but also to the practitioners implementing the intervention (Hay, 2012).

Other risks involve implementation of evaluations in incomplete and technocratic modes (Lombardo, Meier, & Verloo, 2013) that lead to rigid exercises with criteria, methods and indicators that are pre-established without considering a reflexive and critical stance (Bustelo, 2017). In addition, gender-sensitive evaluation is perceived in general as a difficult "task of experts" as changes are considered being too complex and requiring a combination of quantitative and qualitative methodologies (Espinosa, 2013). Moreover, as mentioned earlier, the lack of adequate instrumentation, in particular of qualitative character, or the use of mono-dimensional approaches, may make analyses of effects difficult and cause resistance to evaluations (Espinosa, 2013, Hunt & Brouwers, 2003).

4.1. A reflection on the process of developing the framework

As described above, developing a conceptual evaluation framework is a challenging endeavor. Thus, the process resulting in the framework has been a comprehensive task and an exercise that involved a great amount of literature studies and consultations with a large number of experts, policy makers, practitioners and other stakeholders from many European countries. Nevertheless, the process worked out well with its many feedback supporting loops based on lessons learned that fed into the previous stages and reinforced the subsequent steps (Fig. 2) involving all the relevant stakeholders. In parallel with the reinforcing loops process, a large effort was deployed to map the evaluation landscape and traditions in a number of European countries, identify best practice examples, criteria and indicators for the assessment of GE interventions in R&I, conduct case studies and develop theories of change and impact stories.

While mapping the country context was carried out with the support of the literature and the country correspondents, experts in their country conditions, the mapping of the national evaluation culture proved to be a difficult exercise (with the exception of the Nordic countries, which have a long and well-established evaluation tradition) as the available sources were scarce. However, involving evaluators from the countries participating in the EFFORTI project supported data gathering and facilitated analyses of evaluation culture or pointed out the lack of it. Collecting smart practices was another crucial issue but as described above, the distinct criteria used facilitated the process of identification of the most important smart practice exercises at present. Based on the literature, the smart practices and the work done in European projects (such as PRAGES and GENERA), the typology of interventions was constructed, which supported our work in developing recognizable to stakeholders impact stories for a number of the types of interventions gathered. As to the instrumentation work developed within the evaluation framework, while we identified a plethora of quantitative measures, the challenge constituted in developing the qualitative instruments. In this process, the case studies, the stakeholders consulted, and the European projects mentioned earlier, which had developed some qualitative measures, inspired and assisted our work.

In the validation phase, we encountered a range of problems in the

case study work due to the fact that in some cases it was difficult to i) find the contact details of the beneficiaries of the interventions and ii) to get them to talk about the direct or indirect effects in connection with the conducted interviews because many of the beneficiaries were not aware of the effects or could not see the direct impact of the program they participated in and how the program promoted their career, in particular if the program was implemented some time ago. While this was the case, the great majority of the stakeholders involved in the case study exercise were eager to talk about the interventions and provided valuable information to feed back to the process of developing the framework. Important factors that were identified as enabling and hindering implementation, and thus effecting impact of the interventions, were shed light on, and how these contributed to the understanding of why and how interventions worked or failed to work were highlighted. Focus was thus in particular on challenges and risks to achieve impact while instruments were further developed based on the case study work.

It has been of great importance for the development of the framework to define gender-sensitive criteria, questions and instruments, to ensure the participation of all stakeholders in the process (men and women to assure diversity), to use sufficient time and resources in developing an in-depth analytical frame, paying attention to the non-linearity of the process. The herewith presented framework has been discussed among the European stakeholders through consultations, the national workshops organized in different European countries, the international conferences organized in the frame of the project as well as other international conferences (ESA; ESOF; DeGEval; Gender Summit; Gender, Work and Organisation; STEM Gender Equality Congress, STS and ISSI conferences, among others), where the framework was presented and discussed.

To sum up, while we have proposed a wide-ranging, well-developed conceptual evaluation framework for capturing the complexity of interventions and their effects in complex systems, it is evident that the framework needs to be tailored to each particular intervention to adjust to local conditions, be designed with gender- and context-sensitivity, and consider the manifold challenges that are related to assessing effects of GE interventions in R&I. Looking for linkages between interventions and observed effects can be based on theoretically founded assumptions and intervention logic models, where team, organizational and system conditions are taken into account. Hence, in accordance with the presented framework, theory may guide the empirical and methodological choices in evaluating complex policy interventions.

5. Concluding remarks

Wrapping up, the presented conceptual evaluation framework underpin the understanding of GE in R&I as a multidimensional issue, and of GE policies as complex processes of which the design, implementation, output, outcomes and impact depend on the interaction of a multiplicity of variables in dynamic contexts.

The presented framework embraces the complexity, gender-sensitivity and theory-based evaluation approaches ensuring that design and evaluation of GE interventions consider the complex contextual factors. The framework offers a non-linear concept, where the notion of contribution - instead of attribution - to achieve impact is central to the integration of team, organizational and system context factors in policy design and evaluation. The framework opens the “black box” to address the question of how and why a policy intervention works and in which context and presents a systematic process on how to evaluate the linkages between input, throughput, output, outcome and impact in GE interventions in R&I, accounting for gender sensitivity and methodological pluralism.

The conceptual evaluation framework may serve as reference for researchers, policymakers and R&I stakeholders in general, supporting them both in designing and assessing GE interventions, and in further developing their evidence, and theoretical and methodological base. R&I policymaking ought to be grounded on high-quality gender-sensitive processes that are informed by cutting-edge theories and the highest-quality data.

Funding

This work was supported by the European Union’s Horizon 2020 Research and Innovation programme award EFFORTI under grant agreement No. 710470.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the consortium of the EFFORTI project from which this paper emerged and the anonymous reviewers for their insightful comments that improved the article.

References

Arnold, E. (2004). Policymaking. Evaluating research and innovation policy: A systems world needs systems evaluations. *Research Evaluation, 13*(1), 3–17.

Bacchi, C., & Eveline, J. (2010). *Mainstreaming politics: Gendering practices.* South Australia: The University of Adelaide Press.

Barnes, M., Matka, E., & Sullivan, H. (2003). Evidence, understanding and complexity evaluation in non-linear systems. *Evaluation, 9*(3), 265–284.

Bell, S., Shaw, B., & Boaz, A. (2011). Real-world approaches to assessing the impact of environmental research on policy. *Research Evaluation, 20*(3), 227–237.

Biegelbauer, P. (2013). *Wie lernt die Politik? Lernen aus Erfahrung in Politik und Verwaltung.* Wiesbaden: Springer Verlag.

Brisolara, S., Seigart, D., & SenGupta, S. (2014). *Feminist evaluation and research. Theory and practice.* New York: Guilford Press.

Bustelo, M. (2017). Evaluation from a gender+ perspective as a key element for (Re) gendering the policymaking process. *Journal of Women, Politics & Policy, 38*(1), 84–101.

Bustelo, M. (2011). *Last but not least: Gender-sensitive evaluation as a forgotten piece of the policymaking process.* Working paper.

Callerstig, A., & Müller, J. (2016). *Gender research in GEDII conceptual framework. Project deliverable D1.1. Gender diversity impact*.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: A practical guide to doing it better.* Oxford: Oxford University Press.

Chen, H. T. (2012). Evaluation von Programmen und Projekten für eine demokratische Kultur. In R. Strobl, O. Lobermeier, & W. Heitmeyer (Eds.). *Evaluation von Programmen und Projekten für eine demokratische Kultur*. Fachmedien Wiesbaden: Springer.

Chen, H. T. (1990). *Theory-driven evaluations.* Thousand Oak, CA: Sage.

Cole, G. (1999). Advancing the development and application of theory-based evaluation in the practice of public health. *American Journal of Evaluation, 20*, 453–470.

Connell, J. P., & Kubisch, A. C. (1998). Applying a theory of change approach to the evaluation of comprehensive Community initiatives: Progress, prospects and problems. In K. Fulbright-Anderson, A. C. Kubisch, & J. P. Connell (Eds.). *New approaches to Community initiatives: Volume 2 – Theory, measurement and analysis*. Washington, DC: Aspen Institute.

Dahler-Larsen, P. (2012). *The evaluation society*. Stanford: Stanford University Press.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators. *Public Management Review, 16*(7), 969–986.

Dahler-Larsen, P., & Krogstrup, H. K. (2001). Evalueringers konstitutive virkninger. In P. Dahler-Larsen, & H. K. Krogstrup (Eds.). *Tendenser i evaluering* (pp. 232–245). Odense: Syddansk Universitetsforlag.

Douglas, F. C., Gray, D. A., & van Teijlingen, E. R. (2010). Using a realist approach to evaluate smoking cessation interventions targeting pregnant women and young people. *BMC Health Services Research, 10*, 49. https://doi.org/10.1186/1472-6963-10-49.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Berlin: Springer.

Edler, J., Cunningham, P., Gok, A., Rigby, J., Amanatidou, E., Garefi, I., et al. (2010). *Understanding evaluation of innovation policy in Europe. INNO-appraisal final report. Manchester*. Retrieved October 11, 2019 https://repository.fteval.at/168/1/2010_INNO-Appraisal_Final_Report.pdf.

EIGE GEAR tool (2016). *Gender equality in Academia and research. Gender equality in Academia and research - GEAR tool.* Retrieved January 27, 2019 https://eige.europa.eu/gender-mainstreaming/toolkits/gear.
Espinosa, J. F. (2013). Towards a gender sensitive evaluation? Practices and challenges in international development evaluation. *Evaluation, 19*(2), 171–182.
European Commission (2005). *Assessing the social and environmental impacts of European research.* Luxembourg: Office for Official Publications of the European Communities.
European Commission (2009). *Impact assessment guidelines. SEC(2009)92.* Luxembourg: Publications Office of the European Union.
European Commission (2013). *Gendered innovations: How gender analysis contributes to research.* Luxembourg: Publications Office of the European Union.
European Commission (2014). *Gender in EU-funded research and innovation: Toolkit.* Luxembourg: Publications Office of the European Union.
European Commission (2015). *Indicators for promoting and monitoring responsible research and innovation. Report from the expert group on policy indicators for responsible research and innovation*Luxembourg: Publications Office of the European Union.
European Innovation Scoreboard (2016). https://publications.europa.eu/en/publication-detail/-/publication/6e1bc53d-de12-11e6-ad7c-01aa75ed71a1/language-en/format-PDF/source-31234102. Retrieved February 22, 2019.
Fagenson, E. A. (1990). At the heart of women in management research: Theoretical and methodological approaches and their biases. *Journal of Business Ethics, 9*(4), 267–274.
Fahrenkrog, G., Polt, W., Rojo, J., Tübke, A., & Zinöcker, K. (2002). *RTD evaluation toolbox - assessing the socio-economic impact of RTD policies.* Brussels: European Commission.
Fehmidah, M., Mason, C., McDermott, H., Morris, J., Bagilhole, B., & Nevill, M. (2013). *Advancing women's careers in science, technology, engineering, mathematics and medicine: Evaluating the effectiveness and impact of the Athena SWAN Charter.* Loughborough University.
Fitz-Gibbon, C. T., & Morris, L. L. (1996). Theory-based evaluation. *Evaluation Practice, 17*(2), 177–184.
Fulbright-Anderson, A., Kubisch, A., & Connell, J. (1998). *New approaches to evaluating community initiatives.* Washington, DC: Aspen Institute.
GARCIA (2019). *Gendering the academy and research: Combating career instability and asymmetries.* Retrieved February 22 http://garciaproject.eu/.
GARCIA tool (2015). *Toolkit for Integrating Gender- Sensitive Approach into Research and Teaching.* Retrieved November 19, 2019 http://garciaproject.eu/wp-content/uploads/2015/12/GARCIA_working_paper_6.pdf.
GENERA (2019). *Gender equality network in the european research area.* Retrieved January 27 https://genera-project.com/.
Gender-NET IGAR tool (2009). *Promoting gender equality in research institutions and integration of the gender dimension in research content.* Retrieved January 27, 2019 http://www.gender-net.eu/.
Greene, J. (2005). Context. In S. Mathison (Ed.). *Encyclopedia of evaluation.*. Thousand Oaks: Sage.
Greenhalgh, T., & Papoutsi, C. (2018). Studying complexity in health services research: Desperately seeking an overdue paradigm shift. *BMC Medicine, 16*(1), 95.
Hansen, H. F., & Jørgensen, B. H. (1995). *Styring af forskning: Kan forskningsindikatorer anvendes?* Gylling: Samfundslitteratur.
Halpern, D. F. (2014). It's complicated – In fact, it's complex: Explaining the gender gap in academic achievement in science and mathematics. *Psychological Science in the Public Interest, 15*, 72–74 Association for Psychological Science.
Hay, K. (2012). Engendering policies and programmes through feminist evaluation: Opportunities and insights. *Indian Journal of Gender Studies, 19*(2), 321–340.
Hay, K., Sudarshan, R. M., & Mendez, E. (2012). Why a special issue on evaluating gender and equity. *Indian Journal of Gender Studies, 19*(2), 179–186.
Heckl, E., & Dörflinger, A. (2014). *Begleitende Evaluierung Der Impulsaktion "Laura Bassi centres of expertise".* Vienna: KMU Forschung Austria. Retrieved February 22, 2019 http://www.w-fforte.at/fileadmin/Redaktion/Daten/Downloadbereich/Endbericht_Zwischenevaluierung_LBC.pdf.
Horvat, M. (2011). *Reflections on evaluating FP7 activities. Planning the future knowledge base for evaluation.* Brussels: Conf-Shop.
Hunt, J., & Brouwers, R. (2003). *Review of gender and evaluation. Final report for DAC network on development evaluation*DAC Series Review—OECD. Retrieved July 1, 2019 http://www.oecd.org/derec/dacnetwork/35342762.pdf.
Innovation Indicator (2005). http://www.oecd.org/science/inno/2367614.pdf.
INTEGER. Institutional Transformation for Effecting Gender Equality in Research. http://www.integer-tools-for-action.eu/en. Retrieved February 22, 2019.
Kalev, A., Dobbin, F., & Kelly, E. (2006). Best practices or best guesses? Assessing the efficacy of corporative affirmative action and diversity policies. *American Sociological Review, 71*, 589–617.
Kalpazidou Schmidt, E., & Cacace, M. (2019). Setting up a dynamic framework to activate gender equality structural transformation in research organizations. *Science & Public Policy, 46*(1), 159–169.
Kalpazidou Schmidt, E., & Cacace, M. (2017). Addressing gender inequality in science: The multifaceted challenge of assessing impact. *Research Evaluation, 26*(2), 1–13.
Kalpazidou Schmidt, E., Bührer, S., Schraudner, M., Reidl, S., Müller, J., Palmen, R., et al. (2018). *A conceptual evaluation framework for promoting gender equality in research and innovation, toolbox I - a synthesis report*European project EFFORTIhttps://www.efforti.eu/sites/default/files/2019-10/EFFORTI%20D3.3%20Conceptual%20Evaluation%20Framework%20Final.pdf.
Kalpazidou Schmidt, E., Ovseiko, P. V., Henderson, L. R., & Kiparoglou, V. (2019). Understanding the Athena SWAN award scheme for gender equality as a complex social intervention in a complex system: Analysis of Silver award action plans in a comparative European perspective. *bioRxiv*. https://www.biorxiv.org/content/10.1101/555482v2.
LaFrance, J., Nichols, R., & Kirkhart, K. E. (2012). Culture writes the script: On the centrality of context in indigenous evaluation. *New Directions for Evaluation, 135*, 59–74.
Larsen, F., & Lassen, M. (2001). Evalueringen set i et policy-perspektiv. In P. Dahler-Larsen, & H. K. Krogstrup (Eds.). *Tendenser i evaluering.* Odense: Syddansk Universitetsforlag.
Laursen, S. L., Austin, A. E., Soto, M., & Martinez, D. (2015). Strategic institutional change to support advancement of women scientists in the academy: Initial lessons from a study of advance IT projects. In M. A. Holmes, S. O'Connell, & K. Dutt (Eds.). *Women in the geosciences: Practical, positive practices toward parity* (pp. 39–50). Hoboken, NJ: John Wiley & Sons.
Lewin, K. (1943). Defining the "field at a given time". *Psychological Review, 50*, 292–310.
Lipsey, M., & Pollard, J. (1989). Driving toward theory in program evaluation: More models to choose from. *Evaluation and Program Planning, 12*, 317–328.
Lombardo, E., Meier, P., & Verloo, M. (2013). *Policymaking. Oxford Handbook on Gender and Politics*. Oxford, UK: Oxford University Press.
Lund University (2010). *AKKA III – Ett genusintegrerat ledarutvecklingsprogram.* Intenal document.
Lövkrona, I., & Widén, K. (2012). *AKKA Vitbok - Ledarskapsprogrammet AKKA vid Lunds universitet år 2004-2011: erfarenheter från ett genusintegrerat ledarskapsprogram, AKKA styrgrupp.* Lund University. Retrieved February 22, 2019 https://www.yumpu.com/sv/document/view/31948666/akka-vitbok-lunds-universitet.
Lövkrona, I., Agrell, C., & Widén, K. (2006). *AKKA - AKademiska Kvinnors Ansvar: Ledarutvecklingsprogram för kvinnor vid Lunds Universitet.* Lund University. Retrieved February 22 2019 http://eige.europa.eu/sites/default/files/akka_i_rapport2006_0.pdf.
Lövkrona, I., Agrell, C., & Widén, K. (2008). *Ledarutvecklingsprogram för kvinnor vid Lunds Universitet 2006-2007: Erfarenheter av ett genusintegrerat ledarskapsprogram.* Lund University. Retrieved February 22, 2019 http://eige.europa.eu/sites/default/files/akka_ii_rapport2008_0.pdf.
Mark, M. M. (2001). Evaluation´s future: Furor, futile, or fertile? *American Journal of Evaluation, 22*, 457–479.
Martin, B. (2011). The Research Excellence Framework and the 'impact agenda: Are we creating a Frankenstein monster? *Research Evaluation, 20*(3), 247–254.
Mayne, J., & Johnson, N. (2015). Using theories of change in the agriculture for nutrition and health CGIAR research program. *Evaluation, 21*(4), 407–428.
Miles, I., & Cunningham, P. (2005). *SMART innovation: A practical guide to evaluating innovation programmes. A study for DG Enterprise and Industry*. October.
Molas-Gallart, J., & Tang, P. (2011). Tracing 'Productive interactions' to identify social impacts: An example from the social sciences. *Research Evaluation, 20*(3), 219–226.
Moser, A. (2007). *Gender and indicators.* Brighton: Institute of Development studies.
Moser, C. (2005). Has gender mainstreaming failed? *International Feminist Journal of Politics, 7*(4), 576–590.
OECD Science (2014). *Technology and Industry outlook.* Retrieved February 22, 2019 https://www.oecd-ilibrary.org/science-and-technology/oecd-science-technology-and-industry-outlook-2014_sti_outlook-2014-en.
OECD Science (2015). *Technology and Industry scoreborad.* Retrieved February 22, 2019 https://www.oecd-ilibrary.org/science-and-technology/oecd-science-technology-and-industry-scoreboard-2015_sti_scoreboard-2015-en.
Ovseiko, P. V., Chapple, A., Edmunds, L. D., & Ziebland, S. (2017). Advancing gender equality through the Athena SWAN Charter for Women in Science: an exploratory study of women's and men's perceptions. *Health Research Policy and Systems, 15*(1), 12.
Palmén, R., Reidl, S., Bührer, S., Kalpazidou Schmidt, E., Holzinger, F., Groó, D., et al. (2018). *Synthesis report: EFFORTI: Deliverable 4.2*. Retrieved October 11, 2019 https://www.efforti.eu/sites/default/files/2019-05/D4.2_EFFORTI_Comparative_Report_28032018_FINAL.pdf.
Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed). Thousand Oaks, CA: Sage 9.
Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London: Sage.
Podems, D. (2010). Feminist evaluation and gender approaches: There's a difference? *Journal of MultiDisciplinary Evaluation, 6*(14), Retrieved July 2, 2019 http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/199/291.
PRAGES (2019). *Practicing gender equality in science.* Retrieved January 27 http://www.pragesdatabase.eu/.
Ravn, T., Nielsen, M. W., & Mejlgaard, N. (2015). *Metrics and indicators of responsible research and innovation. Progress report D3.2. Monitoring the evolution and benefits of responsible research and innovation (MoRRI)*. Retrieved October 10, 2019 https://www.rri-tools.eu/documents/10184/47609/MORRI-D3.2/aa871252-6b2c-42ae-a8d8-a8c442d1d557.
Reale, E., Avramov, D., Canhial, K., Donovan, C., Flecha, R., Holm, P., 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.
Reale, E., Nedeva, M., Thomas, D., & Primeri, E. (2014). Evaluation through impact: A different viewpoint. *Fteval Journal,* 39.
Reidl, S., Holzinger, F., Streicher, J., Beranek, S., Unger, M., & Hafellner, S. (2018). *Comparative background report: EFFORTI – deliverable 2.3*. Retrieved October 11, 2019 https://www.efforti.eu/sites/default/files/2019-10/EFFORTI%20D2.3%20Comparative%20Background%20Report%20Context%20Analysis%20Final.pdf.
Rog, D. J. (2012). When backround becomes foreground: Toward context-sensitive evaluation practice. *New Directions for Evaluation, 135*, 25–40.
Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex

aspects of interventions. *Evaluation, 14*(1), 29–48.
Rhomberg, W., Steindl, C., & Weber, M. (2006). *Neue Entwicklungen im Bereich der Wirkungsanalyse und - abschätzung FTI-politischer Maßnahmen, Endbericht*.
Schiebinger, L., & Schraudner, M. (2011). Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering. *Interdisciplinary Science Reviews, 36*(2), 154–167.
Schiffbänker, H. (2009). Vereinbarkeit als Angelpunkt der Karrieregestaltung in wissensintensiven Berufsfeldern: Befunde aus Kreativwirtschaft und außeruniversitärer Forschung. In M. Steiner (Ed.). *Ein anderer Blickwinkel: Ergebnisse aus der angewandten Genderforschung* (pp. 62–87). Graz: Leykam.
Seigart, D., & Brisolara, S. (2002). Feminist evaluation: Explorations and experiences. *New Directions for Evaluation, 2002*(96), 1–114.
Spaapen, J., & Van Drooge, L. (2011). Introducing "productive interactions" in social impact assessment'. *Research Evaluation, 20*(3), 211–218.
Squires, J. (2007). *The new politics of gender equality*. New York, NY: Palgrave MacMillan.
STAGES (2019). *Structural transformation to achieve gender equality in science.* Retrieved January 27 http://stages.
Stone, D. (2016). Quantitative analysis as narrative. In M. Bevir, & R. A. Rhodes (Eds.). *Routledge handbook of interpretive political science* (pp. 151–170). London: Routledge.
Streicher, J. (2017). *Evaluations, acts and institutions. The case of research, technology and innovation policy in Austria*PhD thesis. Vienna University of Economics and Business.
Timmers, T. M., Willemsen, T. M., & Tijdens, K. G. (2010). Gender diversity policies in universities: A multi-perspective framework of policy measures. *Higher Education, 59*, 719–735.
Thurston, L. P., Smith, C. A., Genskow, K., Prokopy, L. S., & Hargrove, W. L. (2012). The social context of water improvement evaluation. *New Directions for Evaluation, 135*, 41–58.
Van Belle, S. B., Marchal, B., Dubourg, D., & Kegels, G. (2010). How to develop a theory-driven evaluation design? Lessons learned from an adolescent sexual and reproductive health programme in West Africa. *BMC Public Health, 10*, 741.
Vogel, I. (2012). *Review of the use of 'Theory of change' in international development.* UK: DFID.
Weiss, C. H. (1973). Where politics and evaluation research meet. *Evaluation,* 37–45.
White, G. (2009). *Pushing the boundaries of impact evaluation. Report on knowledge development possibilities*SQW Consulting. Retrieved October 11, 2019 http://www.sqw.co.uk/files/2813/8712/1121/63.pdf.

Evanthia Kalpazidou Schmidt is Associate Professor and Research Director at the Department of Political Science, Aarhus University. She specializes in science and innovation policy; Responsible Research and Innovation; gender in knowledge production and research organizations; European gender policies and strategies; impact of knowledge governance and policy interventions; research evaluation; European research policy and governance. She has been involved in a number of projects funded by the EU and has been frequently engaged by the European Commission as expert in the evaluations of FP6, FP7 and Horizon 2020 project proposals. She is the European Commission appointed Danish expert member of the European RTD Evaluation Network and has been an expert member of the Horizon 2020 Advisory Group for Gender.

Ebbe Krogh Graversen is Associate Professor and Head of Research at the Department of Political Science, Aarhus University. His research is on science and innovation policy, impact and effect measuring, incentive structures, knowledge relations and innovation capacities, indicator construction and development, impacts of gender equality policies, and impacts of innovation policies and knowledge capacity building in the Higher Education field. He has been involved in national and European projects on research and innovation policy, impact evaluations, indicator development and identification of societal impact of R&D and innovation capacities.