AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the post-print version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Kalpazidou Schmidt, E., & Cacace, M. (2017). Addressing gender inequality in science: the multifaceted challenge of assessing impact. *Research Evaluation*, 26(2), 102-114.

<https://doi.org/10.1093/reseval/rvx003>

Publication metadata

Title:	Addressing gender inequality in science: the multifaceted challenge of assessing impact
Author(s):	Evanthia Kalpazidou Schmidt, Marina Cacace
Journal:	<i>Research Evaluation</i> , 26(2), 102-114
DOI/Link:	https://doi.org/10.1093/reseval/rvx003
Document version:	Accepted manuscript (post-print)

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Addressing gender inequality in science: the multifaceted challenge of assessing impact

Kalpazidou Schmidt, E. & Cacace, M.

Abstract

The analysis of the reasons behind the persistent underrepresentation of women in senior positions in science is well-developed. In contrast, the assessment of the impact of policies addressing the problem suffers from a lack of evidence and an oversimplification of approaches. Based on the assessment of 125 programs for gender equality implemented in research organisations in Europe, North America and Australia, we argue that holistic approaches and multidimensional frames of reference are needed for impact assessment, also to improve program design and policy. Our analysis shows that the problem of gender inequality is rooted in so many and interrelated factors that program impact assessment has to be multidimensional and complex. Having a conceptual approach grounded in the notion of complexity as a point of departure, the paper presents an innovative impact assessment tool, pointing to effective ways to assess the impact of gender equality programs.

Key words: impact assessment, gender equality programs, research organisations, Europe, North America, Australia

1. Single-issue approach versus complex frames of reference for gender equality programs

Gender inequality in scientific careers is a persistent problem, entailing women's underrepresentation in senior and decision-making positions (Shen, 2013). Despite the fact that the situation appears more favorable for the youngest generations of female academics, the gender gap is still disproportionately high compared with the growth in the share of female students and PhD students, which contradicts the hypothesis that women will automatically "catch up" to male colleagues (European Commission, 2016). Proactive policies are hence essential to decrease the gender gap (Caprile et al., 2012). A scientific debate is ongoing over the most salient reasons behind women's underrepresentation in science, which has important implications in terms of impact assessment and policy. In the current debate, we distinguish between two main approaches.

The first approach invites to identify the most critical and urgent factor leading to inequality and primarily focus on it to avoid misplaced efforts. Ceci, Ginther, Kahn and Williams (2014) and Ceci and Williams (2011), for instance, argue that nowadays, there is one dominant cause for women's underrepresentation. According to this perspective, it is a waste of resources addressing problems that no longer exist, such as sex discrimination in grant allocation, manuscript reviewing and hiring in tenure positions. Resources should be used on the real source of underrepresentation, that is, to make institutions "responsive to differing biological realities of the sexes" (such as family life and childbirth). Garforth and Kerr (2009) and LeFeuvre (2014) also argue for the need of focusing on what the current problem really is. For LeFeuvre, the real problem to address is modern sexism (cf. Swim et al., 1995) and indirect institutional discrimination, based on

stereotypical beliefs about women's deficiencies or wrong priority settings. Instead, academic institutions adopt a combination of measures of different kinds, which sends mixed messages to policy makers about the real cause of the problem. A variant of this approach claims that many factors do concur to unequal career outcomes in organizations, but some of them are not effectively addressed by existing measures (for example, networking and mentoring have no significant impact on women's isolation as regards professional networks). It is therefore better to focus on the most impactful policies (see Kalev, Dobbin and Kelly, 2006). However, the results of Kalev, Dobbin and Kelly suggest that many variables intervene in mediating the impact of the different programs, some related to the environment (i.e. responsibility and structure of the organization), some to the specific features of the implemented programs (i.e. target, length, format, etc.).

This leads us to *the second approach*, which is more holistic and advocates for integrated solutions. This approach is based on the notion that focusing on a single aspect of gender inequality is not sufficient to design and assess impact of gender equality policies and establish gender-balanced conditions, as a multiplicity of interrelated factors of different nature is involved in the process (Glass and Minnotte, 2010). This approach is supported by results of a study investigating the efficacy of gender equality policy measures in the university system in the Netherlands (Timmers, Willemsen and Tijdens, 2010), which reveals a statistically significant correlation between the number of policy measures implemented and the universities' ability to counter inequalities. In addition, a study by Bleijenbergh, Benschop and Vennix (2008) reveals that gender inequality dynamics are strongly interconnected with other organisational problems and that analysing and addressing them separately is ineffective. According to Cullen, Junge and Ramsden (2008:127) there is "a substantial body of evidence that the complex combination

of structural, cultural, institutional and economic factors that create barriers for women in SET require a correspondingly integrated and sophisticated strategic and operational response”.

1.1. Moving beyond simple approaches in impact assessment

In the literature, we notice a consensus among scholars that impact assessments are gaining importance in particular in connection with policy making (Reale et al., 2014; Donovan and Butler, 2007). More specifically, the interest for the societal and economic impacts of interventions has led to increased efforts to develop improved assessment procedures (Reale et al., 2014; Meagher et al., 2008). Most approaches to societal impact assessments focus on simple, linear models (consult Rogers 2008) while other scholars, pointing out the complexity of programs, conclude that it might not be fruitful to address complex variables and interventions, hence questioning whether they can be worked out (Stufflebeam, 2004). Other scholars again welcome holistic approaches but find complex programs “too amorphous to deliver, and too difficult to meaningfully evaluate” (Pinnegar, 2006:4). Rogers (2008:29) argues that there are significant challenges in relation to evaluation of “interventions with complex aspects” but puts emphasis on the question of whether we can represent reality as a simple causal model or should the “models we use address the complexity of life”. Yet, there are researchers (Barnes et al. 2003; Davies, 2005; Pawson, 2006; Sanderson, 2000; Funnell, 2000; Stame, 2004) that incorporate concepts of complexity in their models but address different aspects of it in their studies, depending on how they define and conceptualize complexity.

In this paper, we argue that – even if looking for the dominant and/or more current cause of inequality may be perceived as more efficient – simple, linear

models and mono-dimensional approaches are not sufficient in effectively assessing the actual impact of gender equality programs or in adequately designing them in the first place. Framing the problem in the complexity perspective, considering multiple components, agents and strands would instead be beneficial to both aims (Glouberman, 2001; Glouberman and Zimmerman, 2002).

According to this perspective, it is maintained that a large set of factors of different nature, scale and depth interact in producing greater or smaller impacts, depending on the social, cultural, normative, economic and organizational setting of the concerned institutions and their subunits. In a similar line, Halpern (2014), responding to Ceci, Ginther, Kahn and Williams (2014: 73), argues that gender outcomes in science “depend on a multitude of variables that combine in complex, non-linear ways”. According to Halpern (2014: 74) “complex systems are adaptive”, so that the “dominant” or “current” cause may shift from time to time and in different settings, but is in continuous interaction with an array of factors.

Glouberman and Zimmerman (2002) distinguish between the concept of ‘complicated’, which has many components, and the concept of ‘complex’ that is characterized by uncertainty and emergence. The complexity perspective stresses the need to avoid linear input-output measurements of programs when addressing complex problems such as gender equality in science. Such measurements may risk only assessing the mixed effects of multiple variables, whereas the role of the implemented actions in producing impact remains unclear (Glouberman and Zimmerman, 2002; Blamey and Mackenzie, 2007). Complex programs are characterised by recursive causality (with reinforcing loops), disproportionate relationships (where at critical points, a small change can lead to a big difference) and emergent outcomes (Rogers, 2008). The notion of impact itself is affected by the complex frame. For this kind of complex programs, indeed, given the

complexity of the variables involved and their emergent character, the relation between an intervention and its impact presents “the greatest challenge for evaluation”, “because the path to success is so variable” that it cannot be securely articulated in advance (Rogers, 2008:31). In complex frames, effects cannot be deterministically linked to the program.

It is useful to recall here that in the social sciences, inherently dealing with complex phenomena, the notion of causation is hardly deterministic in general. Gerring (2005:167) identified the minimal, or core understanding of causation for the social sciences as follows: “a cause raises the probability of an event occurring”. The impact of programs insisting on a complex social phenomenon like gender inequality therefore needs to be assessed, considering how the program is increasing the probability that change occurs in those social areas and factors which are theoretically connected with gender inequality, this increased probability being in fact part of the impact itself. In this perspective, the production of impact is connected to a program’s ability to foster the right “conditions for change” (Reale et al., 2014:40), and impact assessment of complex interventions can be rather intended, related to the creation of adequate “conditions for impact”.

Not many studies are available which attempt to account for the complexity of the issues involved in complex interventions, relying on significant empirical data of implemented gender equality programs. As mentioned above, even studies inspired by holistic approaches tend to use simple impact indicators, typically the number of women full professors.

2. The study

Taking into consideration the complex nature of the issue, we carried out a studyⁱ aimed at testing an innovative assessment tool which had the complexity perspective as a point of reference. The methodological approach adopted was mainly qualitative, comprising a semi-structured questionnaire sent to key informants at the promoting organizations as well as an analysis, conducted using a grid, of documents and data of different kind provided by the programⁱⁱ promoters or collected through websites and otherwise. It has to be pointed out that the aim was not to compare similar programs or initiatives exploring which were the best in general, but to analyze different programs or initiatives and map what kind of results they yielded on different areas of equality. The aim was hence to construct and test a sensitive and flexible evaluation tool, grasping impacts on different change areas in multiple settings and initiatives. This has its point of departure in the fact that the notion of impact in this study is substantively connected to the creation of the right conditions for change, according to the theoretical perspective presented above. The objective was thus not to make a ranking, but to explore the potentialities (in probabilistic terms) of different types of interventions and operational models to trigger change processes in a specific environment. More important, the types of initiatives and programs needed to be diverse, because our approach entailed that different areas of intervention were included, given the complexity of the problem of gender inequality in science.

As for the selection process of the programs to be assessed, a mailing list of potential program promoters was built based on a preliminary analysis of documents and of relevant websites. A review of well-known organizations, associations and networks was conducted as a first step, such as the European Platform of Women Scientists (EPWS); the European Association for the Study of Science and Technology (EASST); the European Association of Science Editors

(EASE); the European Association for the Promotion of Science and Technology (EuroScience); the European Science Foundation (ESF) and the European Association of Research Managers and Administrators (EARMA); GeNet – ESRC’s Gender Equality Network; the European Union Women Inventors and Innovators Network (Euwiin). We then identified numerous other networksⁱⁱⁱ. To widen the target, a snowballiv sampling technique was subsequently applied, as this technique is particularly fruitful for populations that are not well delimited or enumerated (PRAGES 2009). The mailing list contained 1,112 entries from different organisations worldwide. Programs from both public and private organisations were included. Four main categories of organisations promoting gender equality were identified. Table 1 illustrates the breakdown by type of the 1,112 organizations included in the mailing list and the 125 organizations which responded to the questionnaires.

Table 1. Promoting organizations by type: organizations in the mailing list (N = 1,112) and respondent organizations (N = 125)

Organization type	Organizations in the mailing list %	Respondent organizations %
University/research institutes (public and private universities and research centers)	53	32
Non-for-profit bodies (networks, associations, private non-profit groups i.e. NGOs or foundations, professional associations)	22	50
Public bodies (public agencies and institutions)	13	13
Private for-profit institutions, organizations and companies	9	5
N.A.	3	/
Total	100.0	100.0

Table 2. Respondents by position/role in the program

Position/role	figures	%
Top leadership position in the organization (president, director, chair, CEO)	29	23
Management position in the organization (policy manager for careers and employment; Deputy director for Human Resources, etc.)	40	32
Program manager (PI, professor, officer, etc.)	36	29
Program staff	11	9
N.A.	9	7
Total	125	100.0

All the responsible persons of the entries in the constructed mailing list were sent an electronic questionnaire comprising 63 questions, divided into three sections: general *contextual* information on the organization where the program was carried out; information related to program *implementation*; information on the *impact* of the program. 125 questionnaires were returned about as many programs. The respondents included scholars, decision-makers, top leaders or managers of the organization, and respondents who were qualified to give feedback due to their particular role in the studied intervention (table 2). In 7% of the cases, the role of the respondent in the program was not specified, but they did respond to the questionnaire, which implies that they were involved or had enough information and data about the program.

As anticipated, 125 gender equality programs are included in the study, implemented in Europe (66 programs), North America (33 programs) and Australia

(26 programs) during the last three decades (with 3 out of 4 starting after 2000).

The analyzed programs mostly had a permanent^v character (69%), while some of them were concluded (18%) and a few were ongoing but had not a permanent character (13%) when the assessment was performed. Among the later, only 5 were initiated within the last 3 years.

Given the fact that the universe of reference (i.e. total number of existing programs) was unknown, the aim was *not to have a representative sample* of programs but to identify a panel of programs to analyze and assess from as many European countries as possible (18 from the EU are actually represented in the study as well as Switzerland), as well as from Australia, Canada and the USA. Table 3 illustrates the respondent organizations by country and type of promoter, according to the same categories used for the mailing list entries.

Table 3. Respondent organizations by country and type of promoter of the program/intervention (figures)

	Country	University/ research institutes	Non-for- profit body	Public body	Private for- profit body	Total
1	Australia	12	8	4	2	26
2	Austria	2	1	1		4
3	Belgium		1	1		2
4	Canada	2	9		3	14
5	Czech Rep.	1				1
6	Denmark	4	1			5
7	Estonia		1			1
8	EU wide		1			1
9	Finland	2	1			3
10	France	1	3			4
11	Germany	7	2	1		10
12	Greece			1		1
13	Italy	4	2	1		7
14	Malta			1		1
15	Netherlands	1				1
16	Norway			1		1

17	Slovakia	1				1
18	Slovenia		1	1		2
19	Spain	5	2	3	1	11
20	Sweden			1		1
21	Switzerland	2				2
22	UK	4	3			7
23	USA	15	4			19
	Total	63	40	16	6	125

2.1. Impact assessment strategies

Gender equality programs in research institutions were studied and their impact analyzed as complex processes. This was accomplished according to *four strategies*. The *first* was conducting a **preliminary quality assessment** of the implementation process, so to exclude programs where severe management problems could have too heavily interfered with the production of impacts. The *second* was to **widen the areas where impact could be recorded**, going beyond (but still including) the narrow set of outcomes traditionally considered (such as numerical targets and achievements as to the increase of women in senior positions). Based on a literature review^{vi} and the preliminary document analysis connected to the setup of the mailing list of the organizations promoting gender equality, **three impact areas** have been identified.

Area 1, women-friendly environment, includes programs' impacts on research institutions as professional environments, making them more friendly and enabling for women's advancement and improved working life; *Area 2, gender-aware science*, comprises programs' impacts on the very image and understanding of science as a male domain and on the inclusion of a gender dimension in research priorities, design and teaching; *Area 3, women's leadership*, consists of programs' impacts on women's presence in different kinds of scientific leadership positions

(not just senior academic positions, but also scientific management, communication, innovation, etc.). Nine specific objectives (see Table 4) and 61 lines of actions^{vii} have been identified. The programs have therefore been rated for their ability of producing impacts on their selected targets.

Table 4. Strategic areas and specific objectives pursued by gender equality programs implemented in Europe, North America and Australia

Strategic areas	Objectives
1. Women-friendly environment	1.1. Promoting change in the culture and formal/informal behaviors of scientific work environments 1.2. Promoting work-life balance 1.3. Supporting early-stage career-development
2. Gender-aware Science	2.1. Overcoming stereotypes of women and science, which lead to gendering scientific fields, sub-fields and tasks 2.2. Including the gender dimension in the very process of research and innovation design
3. Women's leadership of science	3.1. Supporting women's leadership in research practice 3.2. Strengthening women's leadership in research management 3.3. Strengthening women's role in science communication 3.4. Supporting women's role in innovation processes

The *third* strategy to deal with the challenge of complexity consisted in considering, besides actual impacts, also **potential impacts** connected to the creation of conditions conducive to change in the concerned organizations, thus taking into account – in a developmental perspective – impacts which were likely to derive from the implemented measures in a longer time-frame. This probabilistic assessment was performed through a set of indicators pointing to the activation of

internal change processes (e.g. the successful involvement of internal and external actors, the modification of relevant rules, the creation of internal groups of actors aimed at pursuing change), as internal processes are likely to produce additional impacts with time (see also Reale et al., 2014). The *fourth* strategy to manage complexity in impact assessment was accounting for and integrating the influence of **local conditions**. This was done by pointing out the contextual elements which affected each program's ability to produce specific impacts, thus avoiding generalizing conclusions about the effectiveness of specific measures (for example statements such as “mentoring works”, “mentoring does not work”). An assessment of the transferability of each program was performed to this aim. Transferability is a widely used concept in the management-oriented benchmarking literature, aimed at identifying the conditions for the positive adoption of a managerial practice in different settings (Von Eiff, 2015). The *transferability assessment* in this study was based on the analysis of both procedural and structural elements of the initiatives. This was done to understand – on the one hand – which elements have been critical to success and – on the other – if these are more or less linked to the specific characteristics of a given setting, which would be difficult to reproduce elsewhere.

2.2. The assessment process

Following the above-identified strategies, the assessment of the programs was performed in three phases. The **first phase** comprised the evaluation of the quality of the implementation process (first strategy) considering its inputs and outcomes (Sartorius, 1991). The **quality dimension** of the programs was assessed based on the four parameters of *relevance*, *effectiveness*, *efficiency* and *sustainability*. Indicators were developed for each parameter (see Table 5) and related algorithms were defined and calculated, leading to the assignment of four analytic quality

index values for each program (poor, limited, medium, good, excellent), one for each quality parameter. A synthetic quality index value was then calculated based on the analytic values. Only programs overcoming a basic quality threshold were admitted to the second evaluation phase (i.e. 109 out of a total of 125 programs).

Table 5. Quality assessment indicators

Relevance

Relevance has been defined referring to two parameters: a) provision of activities intended to increase promoters' knowledge of the reality and needs of national, local and institutional research settings in which the programme would operate; b) programme capacity of adaptation to its specific setting.

Answers to questions and items contained in the questionnaire and in the grid pertaining to relevance have been evaluated, a Relevance Index calculated (0-10) and a label (Poor, Limited, Medium, Good or Excellent) assigned based on the scaling system results.

Effectiveness

Effectiveness has been defined according to two parameters: a) achievement of the identified objectives; b) performance of the activities envisaged.

Answers to questions and items contained in the questionnaire and in the grid pertaining to effectiveness have been evaluated, an Effectiveness Index calculated (0-20) and a label (Poor, Limited, Medium, Good or Excellent) assigned based on the scaling system results.

Efficiency

Efficiency has been defined according to four parameters: a) extension of the activities during implementation (increase or decrease in planned activities, increase or decrease in beneficiaries, expansion of the programme over time, etc.); b) availability of resources to the extent envisaged; c) financial reporting capacity; d) managerial capacity.

Answers to questions and items contained in the questionnaire and in the grid pertaining to efficiency have been evaluated, an Efficiency Index calculated (0-10) and a label (Poor, Limited, Medium, Good or Excellent) assigned based on the scaling system results.

Sustainability

Sustainability has been defined according to two main parameters and two subsidiary ones. Main dimensions: a) institutionalisation of the initiative (adoption of the programme, including financial aspects, by the sponsoring institution); b) degree of men's involvement in the promotion of the project. Subsidiary dimensions: c) diversification of the financing sources; d) identification of further economic resources.

Answers to questions and items contained in the questionnaire and in the grid pertaining to sustainability have been evaluated, a Sustainability Index calculated (0-10) and a label (Poor, Limited, Medium, Good or Excellent) assigned based on the scaling system results.

Note: Quality has been assessed according to four dimensions: Relevance, Effectiveness, Efficiency, Sustainability. For each of these dimensions, Analytical Quality Indices were produced. Finally, a Synthetic Quality Index was assigned on the basis of the results achieved by the programmes in each dimension. For both Synthetic and Analytic Quality Indices, a five-point scaling system has then been applied, and performances in the different dimensions have been identified as Poor, Limited, Medium, Good or Excellent. At least one medium Synthetic Quality Index has been considered necessary for the inclusion in the database.

In the **second evaluation phase**, the *impact* of the programs was in focus. First of all, the strategic scope of the programs was analyzed, which helped identify the specific set of problems addressed depending on local conditions and the specific objectives pursued by each institution. The actions implemented were thus connected to the three general impact areas previously mentioned (second strategy, see Table 1), and their impact assessed through the application of a set of quantitative and qualitative indicators explicitly developed for the different impact areas and objectives identified.

The impact dimension of the programs was assessed through both **objective and subjective impact** indicators. To the aim of the study, these have been defined highlighting their probabilistic understanding (third strategy). In this perspective, *objective impact* included the effects of the program on the promoting organisation in terms of change affecting the three impact areas identified, but it also included the creation of conditions pointing to the activation of further change processes. *Subjective impact* instead referred to the satisfaction of beneficiaries and the program's ability to promote consensus among the key actors involved, as a prerequisite for change to take root. Based on these definitions, indicators were developed (overall, 28 indicators for objective impact, and 8 for subjective impact; see Table 6), as well as algorithms to calculate analytic impact index values for

each program (still in the poor, limited, medium, good and excellent range). A synthetic impact index value was then calculated based on the analytic values.

Table 6. Impact indicators

Objective impact indicators

1. Program renewed or institutionalized
2. Ability to reach the intended direct beneficiaries
3. Ability to reach the intended indirect beneficiaries
4. New organizational approaches adopted as standard
5. Modification of customary behaviors
6. Adoption of gender-sensitive language
7. Setting up of associations or networks devoted to gender issues within the organization
8. The organization joining external associations or networks devoted to gender issues
9. Individual researchers joining external associations or networks devoted to gender issues
10. Emergence or intensification of conflicts within the organization
11. Emergence or intensification of conflicts between the organization and other connected bodies
12. Creation or intensification of relations with different kinds of relevant associations
13. Formalization and dissemination of the experimented initiative
14. Start up of new initiatives for gender equality in the organization
15. Start up of new initiatives for gender equality in partner organizations
16. Start up of new initiatives for gender equality in other bodies, linked to the organization and/or on the higher levels of the same organization
17. Invitation of the promoters to public initiatives to illustrate the program
18. Promotion of laws and political initiatives based on proposals stemming from the programs
19. Adoption of the program in local or national government administrations
20. Unforeseen positive or negative effects regarding aspects which do not pertain to gender issues
21. Increase in the authority and visibility of women researchers
22. Increase in the presence of women in high-profile decision-making positions in research management
23. Increase in the presence of women in key roles in scientific communication

24. Increase in the presence of women in decision-making roles regarding the evaluation of scientific excellence
25. Increase in the presence of women in key roles regarding the management of innovation processes (patenting, relation with industries, etc.)
26. Increase in the presence of women in important positions in governance processes aimed at linking S&T with objectives of collective interest (relations with political institutions, etc.)
27. Inclusion of sex and gender variables in relevant research and teaching
28. Mitigation of horizontal segregation in research areas and task allocation

Subjective impact indicators

1. Collaboration of external players in the implementation of the activities
2. Internal or external opposition to the program
3. Internal or external support beyond the promoting groups
4. Promoters' level of satisfaction with the success of the program
5. Beneficiaries' level of satisfaction with the results of the program
6. Involved men's level of satisfaction with the program
7. Positive reports about the program from relevant stakeholders
8. Recognition received at local, national or international level

In the third and last phase, the weight of the specific enabling and hindering factors was considered which, at each institution, concurred in determining the impact of the implemented measures, so to assess their **transferability potential** (fourth strategy). This was aimed at identifying under which conditions a similar program would produce comparable results in a different research organization^{viii}. Enabling and hindering factors have been divided into structural and processual. Structural factors pertain to the cultural, social and normative features of the organization, and the environment it is located, as mentioned by the promoters for their effects on project implementation (for instance, a long record of gender equality programs at the university, or the existence of financial rules hindering provision of childcare services). Process factors point instead to those implementation strategies mentioned by the promoters as having been instrumental

to deal with structural circumstances to the advantage of the program (for instance, early men's inclusion in the interventions or the adoption of permanent evaluation procedures). The identification of the enabling and hindering factors was thus based on the data, provided by the people involved in the interventions, on both their own organization and the implementation process, which they often were leading.

Besides enablers and obstacles, two other dimensions were also considered, for their importance to the aim of transferability, i.e. the attitude to disclosing and disseminating information on the implementation of the programs, and the possible replication of the programs themselves (for instance, if they were part of a nation-wide scheme, or had been adopted as good practices by other organizations in the same network). The dimensions considered in this analysis are presented in Table 7.

Table 7. Transferability dimensions

<p><i>Information disclosure</i></p> <p>This dimension assesses the aptitude of promoters to provide detailed information about the programme, its aims and implementation methods, as well as information on obstacles found, facilitating factors or possible forms of replication. Account was taken of the transparency level of the promoters' website and the quality and quantity of online resources provided, as well as the thoroughness with which promoters completed the semi-structured questionnaire.</p>
<p><i>Occurred replication</i></p> <p>This dimension focuses on whether the programme had already been transferred to another context, either through planning or spontaneously, on the assumption that a programme that has been replicated has already shown some aptitude for replicability. Indications are provided on the setting where the programme was adopted, to better assess this aptitude. It also indicated whether a programme was itself a replication of a scheme implemented elsewhere.</p>
<p><i>Enablers: structural factors</i></p> <p>This dimension focuses on structural factors (for example, social and cultural settings, the institutional and regulatory environment, resource availability, etc.) which facilitated the implementation of the programme or made it possible. This is an important aspect for those interested in transferring the programme, since it enables them to assess in advance whether these structural factors exist in the environment the programme will be transferred to (for</p>

example, it would be difficult to transfer a programme that depended on a specific legislative incentive to a setting where there is no such legislation).

Enablers: process factors

This dimension concerns process factors which are important or necessary to acknowledge for successful transfers, such as the existence of a motivated leadership, administrative support from the promoter, the presence of monitoring mechanisms, the inclusion of men besides women in the implementation of the programme, and specific institutional, organisational, communicative, managerial and relational measures or solutions which proved effective.

Obstacles

This dimension considers the obstacles that delayed or complicated the implementation of the programme. Of course, the obstacles are of different types, relating to both structure and process.

The transferability dimension did not entail the assignment of a specific index value to individual programs, because of the intrinsically relativistic character of the transferability concept^{ix}: a program can in fact be judged as more or less transferable depending on the features of the institution willing to adopt it. A narrative evaluation covering the transferability dimensions considered above was instead added to the other profiles of each program, providing additional information on the circumstances and factors connected to its outcomes.

The assessment model developed and tested for the study, as described above, is summarized in Figure 1.

Figure 1. An illustration of the impact assessment model

2.3. Credibility and validation

As Patton (1999) points out, credibility of a study can be demonstrated through rigorous data collection and analysis, and the expertise of researchers in methodology. In our study, data collection and analysis have been conducted through rigorous procedures and based on guidance provided in the literature. These procedures were developed to better respond to the challenge of addressing the complexity of the programs in a credible way, linking literature with data collection. This allowed us to incorporate all the variables we identified as significant in the literature in order to evaluate (Stame, 2004) and explain specific dimensions of phenomena in the studied programs (Reimers and Johnston, 2008), thus enabling us to bound our research object theoretically. The coding process of the collected information for each program was theory-driven, conducted through the identified indicators (Reimers and Johnston, 2008). Cross-checking of the coding process was performed iteratively by the two research teams directly

involved in the assessment task (Guba, 1981; Morse et al., 2002). The results were then recorded in an access database input mask, set up to calculate analytic indexes (for each quality and impact dimension) and synthetic indexes (for overall quality and impact) following the algorithms.

In addition, three validation and credibility steps have been initiated to test the instrumentation of the study. The first entailed the testing of the evaluation tool by two scientific organizations involved in the study. The second step comprised the cross-check of the results of the application of the evaluation algorithms to the analyzed programs, and involved all the eight research teams participating in the study. Attention was made that programs from a given country were assigned – when possible – to partners from that same country or geographical area, taking into consideration the specific contextual factors. Divergent evaluations were then discussed and modifications made accordingly. The third step entailed sending to all promoting groups the synthetic description of their program to be commented. Also in these cases, modifications were made based on received feedback (which occurred in 24 out of 109 cases).

3. Results

Before reporting the results of the assessment in terms of quality and impact, it may be useful to indicate the most recurrent types of actions which were included in the analyzed programs. Considering that most were integrated programs, they in fact comprised multiple components. Table 8 provides a distribution of the intervention types in the studied programs.

Table 8. Distribution of intervention types in the studied programs

Type of intervention	figures	%
Networking	81	74
Dissemination of information material	64	59
Mentoring programs	61	56
Training courses (different targets)	49	45
Empowerment schemes	37	34
Mainstreaming actions	33	30
Gender-sensitive practices for assessment	22	20
Introduction of awards reserved for women	22	20
Monitoring appointments, promotions or attribution of tasks	22	20
Revision of internal policies regarding promotions	19	17
Revision of internal policies regarding staff appointments	18	17
Support in period of absence for family needs	18	17
Gender-sensitive practices for the attribution of tasks	15	14
Schemes for women-returners	14	13
Targeting funding practices to improve women's access to research funding	14	13
Care services (for children, the elderly and other)	12	11
Definition of targets regarding gender balance in decision-making positions	12	11
Support to mobility, including spouse-relocation schemes	10	9
Definition of targets regarding gender balance in research groups	9	8
Introduction of chairs and positions reserved to women	9	8
Revision of teaching curricula and texts	9	8
Support to career development (counselling)	8	7
Institution of quotas	5	5
Introduction of single-sex degree and specialisation courses	4	4

Three types of actions (networking, collection and dissemination of gendered data and information, as well as mentoring) were present in over 50% of the studied programs. Networking, collection and dissemination of gendered data and information were often used as a basis for developing more specific and structured interventions. In part, this was also true for mentoring, which is a very flexible type of intervention, in terms of use, characteristics and objectives, and can easily be linked to other actions (Kalpazidou Schmidt and Faber, 2016). Three other types of

actions were also widespread, and were found in between 30 and 50% of programs. Among these, training courses and empowerment schemes, were oriented at promoting equality by acting directly in support of individual women. They could be implemented through a wide and varied set of measures, depending on the specific content of the program.

The remaining types of interventions – used in fewer than 30% of the programs – were highly specific, had a high degree of relevance, but were less flexible. Of this last group – actions with a medium-low or low level of diffusion – the most common (present in between 15 to 30% of the programs) were mechanisms to monitor progress and the attribution of responsibilities as concerns gender equality in the organization. Other common types of interventions were the adoption of gender-sensitive practices to evaluate staff performance, the revision of internal policies regarding promotions and staff appointments, targeted funding practices, and the adoption of gender-sensitive teaching or research practices. The introduction of quotas and the establishment of specific targets to increase the share of women in senior positions were the least widespread actions (included in less than 15% of the programs).

Turning to the overall quality of the studied programs, we can conclude that the quality was high, also because none of the programs admitted to the second phase of evaluation was below the medium level in the synthetic quality index. Nonetheless, more than one third of the programs displayed excellent quality (42 in 109), and almost half reached a good quality level.

As for the geographic area, the highest levels of quality were found in European and North American programs (in both cases 4 in 10 programs achieved levels of excellence). However, when the “excellent” and “good” categories are summed, North American programs show a slightly higher level of quality than the

European. In Australian programs, the level of quality appeared relatively lower: only 2 in 10 programs reached a level of excellence, while about 2 in 10 were of medium quality (twice as many than in North American programs). Overall, a comparison of the results on the four different quality parameters showed that a weak point of the programs was efficiency, while relevance and sustainability represented strong points.

Yet, the picture is generally more positive compared to the dimension of impact. Only 9 programs out of 109 reached an excellent impact level (compared to the 42 programs that reached the same level in terms of quality), while 40 were assessed as having a good impact level (compared to 51 programs that reached the same level in terms of quality). In addition, more than 1 program out of 10 (15 in total) demonstrated an impact below the medium level.

These results highlight the complexity of the task and the difficulties for gender equality programs to achieve significant impacts even when they meet high-level quality standards in terms of management and implementation. Indeed, there seems to be a *non-linear relationship between project quality and impact* (see Figure 2) as promoters and those managing the program – no matter how experienced – are unable to fully control the many different and interconnected variables mediating and resulting in certain impacts.

(P=poor; L=limited; M=medium; G=good; E=excellent)

Figure 2. Impact and quality of programs (figures)

However, results vary as concerns objective and subjective impact, with the latter which revealed easier to achieve (Figure 3). This pattern has several explanations: more time is needed to produce objective than subjective impacts; the factors that come into play in determining objective impacts are more complex and more difficult to control by the promoters compared to those affecting subjective impacts; the assessment of subjective impact is more conditioned by the opinion expressed by the promoting teams than the assessment of objective impacts.

(P=poor; L=limited; M=medium; G=good; E=excellent)

Figure 3. Objective and subjective impact of programs (figures)

Figure 4. Excellent and good impact distribution on the three strategic areas (figures)

In regard to geographical distribution, European programs, compared to the others, seemed to have a greater capacity to impact the gender dimension of research,

teaching, and design (i.e. the second area of the impact assessment strategies; see Table 4), while they seemed to have less objective impact on the creation of an enabling environment for women (the first impact area). In this domain North American and Australian programs tended to be more focused. No clear differences were noticed as concerns the third impact area, related to promoting women in decision-making positions.

An analysis was carried out considering the impacts recorded in the three impact areas by all the programs as a whole. In this perspective, it is interesting to notice that 114 impacts (assessed as good or excellent) were overall achieved by 71 programs (65% of total). It was possible thus to observe how the three impact areas have been affected to different extents by the implemented programs (Fig. 4). This analysis revealed how the impact area dealing with what has been called “fixing the science” (Schiebinger, 2008), i.e. gender-aware science and integration of the gender aspect in research, was the less practiced, its relation with gender equality still insufficiently explored. It can be underlined, in this perspective, that a high proportion of programs (more than 1 in 3) did not have any impact on this area. In most cases, this was not because of a failing in the implementation of the programs, but because of an explicit initial decision not to commit on this issue.

The final results of the study have been collected – in a narrative and anonymous form – in a publicly available database^x aimed at supporting the efforts of promoters of gender equality-oriented programs for women in science. In the database, the results of the quality, impact and transferability assessments were reported for each program, providing the necessary framework to assess the probability that similar impacts could be produced in different organizational settings.

4. Discussion

This paper is based on a study where an assessment tool was developed and tested on programs oriented at addressing gender inequality in scientific organizations and academia, interpreted as a composite phenomenon deriving from a multiplicity of factors which interact in complex frameworks according to varied patterns. This approach had as a point of departure a conceptual framework grounded in Glouberman and Zimmerman's (2002) complicated (multiple components) and complex (uncertain, emergent) notions in program interventions. The concept of complexity underpinning the proposed method, and defining its possible fields of application, comprises: the high number of variables involved and their uncertain character; the deficiency of linear causal relations between the actions and the structural modifications observed in the domain where the actions have been implemented; the interaction between the context, i.e. local dynamics (in this case, connected to the single unit and research organization, or to the local context, depending on the chosen perspective) and widespread social phenomena (in this case, gender inequality and programs to address them).

In order to handle these factors and identify the most effective ways to assess the impact of gender equality programs four methodological choices were made as to:

- conducting a preliminary analysis of the quality of the programs, excluding those which displayed very poor quality levels;
- adopting a more holistic approach, widening the considered impact areas beyond women's access to senior and decision-making positions. This was made in order to include in the analysis, in the light of the most recent literature, theoretically sound inferences on the relationship between

additional variables and phenomena with the social dynamics of gender inequality in science;

- adopting the notion of “conditions for impact” (see also Reale et al., 2014) and developing relevant indicators, thus taking a probabilistic stance as concerns the relations between actions and their middle and long-term impacts, which makes impact analysis of programs addressing complex phenomena less deterministic and more substantive;
- using the concept of transferability, and developing suitable indicators, in order to manage the influence of the local context and to evaluate to which extent the success of a given action is linked to specific circumstances or can be reproduced in other contexts.

The above described conditions also occur in sectors different from gender equality in science, requiring corresponding decisions as to research design and methodology. For this reason, the methodological approach of this study can be inspirational for assessments of programs dealing with equally complex and widespread phenomena like gender inequality dynamics. Nonetheless, the conceptual and methodological choices presented herewith face a range of challenges. As Martin (2011:250) states, impact is often “indirect, partial, opaque and long-term”. In the following we discuss the limitations of the complex approach proposed and how to mitigate them.

4.1.Limitations in assessing impact of complex interventions

A number of challenges were identified in the effort to assess the impact of gender equality programs. *First*, is the key challenge of establishing attribution, i.e. how to measure and attribute the impact to a particular intervention since impacts can

have multiple causes, which makes it problematic to identify causal relations between the program and the generated impact. From the domain of research policy impact, research (Reale et al., 2014:37) reveals that “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”. This is a significant challenge in particular in assessing complex interventions, as those addressing gender equality (cf. Rogers 2008). Bell et al. (2011) in their work within environmental research suggest as a solution to explore the context, broadening the area of confounding factors, shaping outcomes and effects. This was also the strategy adopted in this study (fourth strategy) where contextual issues and a broadening of the studied factors, in combination with a qualitative approach, helped mitigate the attribution challenges.

Second, assessing impact often suffers from lack of available information, data and indicators. Molas-Gallart and Tang (2011) point out the need for advanced models and multiple methods to address the issue. Ryan and Garrett (2004) emphasize the importance of institutionalizing impact assessment procedures, recording outputs and effects, as an answer to the problem. Bell et al. (2011) suggest to carefully considering appropriate data collection methods, going beyond traditional impact indicators and identifying less tangible impacts, such as contextual changes, which also was the case in this study. In addition, as described in the credibility and validation section, data collection and analysis have been conducted through rigorous procedures and based on literature studies to construct more sophisticated indicators.

Third, the time-lag, i.e. the time span between the intervention and the assessment of particular impacts and how to verify the persistence of impacts, are other challenges. Moreover, the timing and when it is appropriate to conduct impact

assessment, has to be taken into consideration. If the assessment is carried out during, too soon or too late after completion of the program assessed, this might affect the outcome as key stakeholders may not link the effects to the intervention (Bell et al., 2011). As the majority of the programs included in our study (almost 70%) had a permanent character, this was most likely not a threat to the outcome of the study as a whole.

Fourth, the scale and intensity of impact can vary. Besides this, not all generated impacts are expected, desirable or good. Reale et al. (2014) distinguish between four types of impact: intended, unintended, expected and unexpected. Most impact assessments focus on the intended and expected or intended but unexpected impacts. By broadening the assessment to capture impact beyond the traditional ways, including the varied processes the programs set in motion (sometimes resulting in resistance and backlash), all four types of impact have been recorded in our study, a result which might better inform policy and design.

Fifth, even though there are many parallels to the mechanisms identified in other domains, the character and constitution of impact may vary according to the particularities of the field (Martin, 2011). Castaño et al. (2010), reveal the decisive role played by institutional contexts in shaping gender balance in academia in terms of intensity and features. 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. This is one of the limitations of our study, despite our effort to include different types of organizations, taking into consideration the specific context and addressing the transferability potential of the programs.

Finally, 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 (Spaapen and Van Drooge, 2011). Data collection, based on guidance provided in the literature (as practiced in the current study) might help address this challenge.

Despite the above-mentioned limitations, the approach discussed herewith looks promising as it points to the need for a method that considers a holistic approach and addresses the challenges as to the three dimensions of quality, impact and transferability, thus improving the possibility of capturing the complexity of the programs. We hereby propose a well-developed framework for impact assessment of gender equality programs (and not a ready to use tool) that needs to be modified and tailored to better adapt to other contexts.

5. Conclusions

Wrapping up, the results of the herewith presented impact assessment support the interpretation of gender inequality in science as a multi-dimensional issue, and of gender equality policies as complex processes, the outcomes of which depend on the interaction of a multiplicity of variables in dynamic contexts. The results support the notion of widening and integrating the strategic scope of gender equality programs in research organizations, as well as their careful designing, leading to the adoption of multidimensional, complexity frames of reference for impact assessment.

Based on the results of a comprehensive study, this article presents an innovative impact assessment tool and argues that holistic approaches in design, evaluation and policy making are crucial to ensure progress towards gender equality in

science. The methodology and outcome of the study could serve as reference for researchers, policy makers and research agents attempting to assess impact of gender equality interventions, further developing their evidence and science base. Recent studies reveal that there is no evidence of natural decline of gender inequality over time in Europe (European Commission, 2013; Caprile et al., 2012). Science policy-making ought to be grounded on high-quality processes that are informed by the highest quality data. This study provided an important piece of evidence as a basis for the launch of the “structural change” strategy^{xi} of the European Commission (2011) and initiatives among member and associate states. The strategy aimed to implement systemic and comprehensive tailored action plans at universities and research institutions addressing structural conditions and the many and interconnected layers of the gender inequality issue.

Acknowledgement

The study was conducted in the frame of the PRAGES project, funded by the FP7 of the European Union and co-funded by the Italian Government (Department for Equal Opportunities). The authors are grateful to the research organizations participating in the study and to their project partners: Research School of Social Science, University of Queensland (Australia); Danish Centre for Studies in Research and Research Policy, Aarhus University (Denmark); European University Institute (EU); TETALAP (Hungary); ASDO (Italy); Department of Sociology and Social Research, University of Milano-Bicocca (Italy); Department of Sociology, University of Cambridge (UK); Centre for Equality and Diversity at Work, Manchester Business School (UK); Center for Gender in Organizations, Simmons College School of Management (USA). Finally, the authors gratefully

acknowledge the contribution of Daniela Falcinelli and her team at the Center for the Study of Women and Gender Differences (renamed GENDERS), University of Milan, in creating the list of program promoters.

ⁱ The study was carried out in the frame of the PRAGES (Practicing Gender Equality in Science) project.

ⁱⁱ We use the concept program for easy reference. A program can include different types of initiatives. In this study, we considered a program either as a single initiative (when the program consisted of just that initiative for example mentoring, which could anyway incorporate multiple actions i.e. networking for mentees, workshops for mentors) or a coordinated set of initiatives (when these were promoted under a unitary implementation framework, as often was the case in this study). Both single and integrated initiatives were analysed. Obviously, the more integrated tended to generate impacts on more areas (impacts have hence been described in the PRAGES database as the emergent result of integration).

ⁱⁱⁱ The most frequently encountered were ADVANCE, a U.S.-based network that aims to increase the participation and advancement of women in academic science and engineering; The ATHENA SWAN scheme in the UK, promoting good employment practice for women working in science, engineering and technology; the Women Members Network of the Royal Society of Chemistry, based in the UK and committed to providing services relevant to the scientific and professional needs of women in chemical sciences; SET-Routes, a pan-European network of women scientists; and AWIS, a U.S.-based network promoting women in science.

^{iv} To construct our sample, we used the snowball sampling technique, reviewing the websites of the most significant actors in order to pinpoint further promoters of programs. We then reviewed these new promoters, proceeding in the same manner until our material became saturated, i.e. once we found no further new contacts in the websites, or were sure that we had collected most of the main organisations operating in a particular country/region. In order to avoid a common disadvantage of this sampling technique, i.e. failing to pinpoint smaller organisations, we started the snowball chain from not one but from several different national associations/organizations or networks.

^v A program was considered to be permanent when it was institutionalised and had a recurring and continuing, diachronic character. Examples of such programs were training courses for research leaders or mentoring programs provided every year or long-lasting, institutionalized family-friendly policies.

^{vi} For literature on informal networks and the existence of “hidden quotas” for women’s and minority groups’ presence in high-level positions see Allmendinger and Podsiadlowski, 2001; Gupta, Kemelgor, Fuchs and Etzkowitz, 2005; for work-life balance issues see Stafford, 1996; Zimmer, 2003; Meulders, 2003; European Commission, 2004; Matysiak and Vignoli, 2008; Martin-Garcia, 2009; for literature on pay gap see Brookman, Haines, Koenig, McLean and Westernwick, 2004; Ivey, 2005; European Commission, 2007 and 2009; for the “leaky pipeline” phenomenon see Etzkowitz, Kemelgor, Neuschatz and Uzzi, 1994; Fielding and Glover, 1999; Gürer and Camp T., 2002; Blickenstaff, 2005, Jensen, 2005; Etzkowitz and Gupta, 2006; Sonnert, Fox and Adkins, 2007; for access to resources for research and early-stage career development see Lawler, 2001; Ellemers, van de Heuvel, de Gilder, Maas and Bonvini, 2004; for evaluation of scientific merit see Wennerås and Wold, 1997; Valian, 1998; Foschi, 2000; Sheridan, 2006.

^{vii} For a detailed description of the different lines of action, see www.pragesdatabase.eu

^{viii} A description of the different dimensions of Quality, Impact and Transferability and their enablers and obstacles in the assessment of two programs, one from Europe and another from the USA, is provided in the supplementary material.

^{ix} For a discussion of the concepts of relative and absolute transferability see Fernandez et al., 2001, Messina et al., 2008.

^x The basic features of each program, along with a narrative account of its results, in terms of quality and impact assessment as well as transferability, are described in the PRAGES database, available at www.pragesdatabase.eu.

^{xi} The European Commission adopted a new strategy moving away from an emphasis on “fixing the women” to “fixing the system” and “fixing the knowledge” approach.

References

Allmendinger, J. and Podsiadlowski A. (2001), ‘Segregation in Organisationen und Arbeitsgruppen’ (Segregation in Organizations and Work Groups), in Heintz, B. (ed) *Geschlechtersoziologie*. Sonderheft Nr. 41/2001 der Kölner Zeitschrift für Soziologie und Sozialpsychologie, pp. 276-307. Opladen: Westdeutscher Verlag.

Barnes, M., Matka, E. and Salivan, H. (2003), ‘Evidence, Understanding and Complexity: Evaluation in Non-Linear Systems’, *Evaluation* 9(3): 265-84.

Bell, S., Shaw, B. and Boaz, A. (2011), ‘Real-world approaches to assessing the impact of environmental research on policy’, *Research Evaluation* 20(3): 227-37.

Blamey, A. and Mackenzie, M. (2007), 'Theories of Change and Realistic Evaluation: Peas in a Pod or Apples and Oranges', *Evaluation* **13**, 439–455.

Blickenstaff, J.C. (2005), 'Women and Science Careers. Leaky Pipeline or Gender Filter?', *Gender and Education*, vol. 17, no. 4.

Bleijenbergh, I.; Benschop, Y. and Vennix, J. (2008), 'Making Gender Equality a Shared Problem in Organizations: Group Model Building as a Gender Mainstreaming Method', Paper presented at the Critical Management Studies Research Workshop, USC Los Angeles, August 7-8.

Brookman, J., Haines, E., Koenig, J., McLean, W. and Westernwick, E. (2004), 'Report on the Cambridge AWiSE, Questionnaire and Discussion Meeting on Experiences of Part-Time and Flexible Working in Science, Engineering and Technology', AWiSE, Cambridge.

Caprile, M., Addis, E., Castaño, C., Klinge, I., Larios, M., Meulders, D., Müller, J., O'Dorchai, S., Palasik, M., Plasman, R., Roivas, S., Sagebiel, F., Schiebinger, L., Vallés, N. and Vázquez-Cupeiro, S. (2012), 'Meta-Analysis of Gender and Science Research – Synthesis Report', Brussels.

Castaño, C.; Muller, J.; Gonzales, A. and Palmen, R. (2010), 'Policies Towards Gender Equity in Science and Research. Meta-Analysis of Gender and Science Research – Topic Report', Brussels.

Ceci, S. J., Ginther, D. K. Kahn, S. and Williams W. M. (2014), 'Women in Academic Science. A Changing Landscape', *Psychological Science in the Public Interest* 1–67. *Association for Psychological Science*. DOI: 10.1177/1529100614541236.

Ceci, S. J. and Williams, W. M. (2011), 'Understanding current causes of women's underrepresentation in science', *PNAS* 108, 3157–3162.

Cullen, J., Junge, K. and Ramsden, C. (2008), 'Evaluation of the UK Resource Centre for Women in Science, Engineering and Technology - Final Report', April 2008, The Tavistock Institute.

Davies, R. (2005), 'Scale, Complexity and the Representation of Theories of Change (Part II)'. *Evaluation* 10:101-21.

Donovan, C. and Butler, L. (2007), 'Testing Novel Quantitative Indicators of Research "Quality", Esteem and User Engagement: An Economics Pilot Study', *Research Evaluation*, 16(4): 231-242.

Ellemers, N., Van Den Heuvel, H., De Gilder, D., Maas, A. and Bonvini, A. (2004), 'The Underrepresentation of Women in Science. Differential Commitment or the Queen Bee Syndrome?', *British Journal of Social Psychology*, vol. 43, no 3.

Etzkowitz, H. and Gupta, N. (2006), 'Women in Science. A Fair Shake?', *Minerva*, vol. 44.

Etzkowitz, H., Kemelgor, C., Neuschatz, M. and Uzzi, B. (1994), 'Barriers to Women in Academic Science and Engineering' in Person W. Jr., Fechter I. (eds.) *Who Will do Science? Educating the Next Generation*, Johns Hopkins University Press, Baltimore.

European Commission (2016), 'She Figures, Gender in Research and Innovation', Luxembourg: Office for Official Publications of the European Communities.

European Commission (2011), 'Structural Change in Research Institutions: Enhancing Excellence, Gender Equality, and Efficiency in Research and Innovation', Luxembourg: Office for Official Publications of the European.

European Commission (2009), 'She Figures 2009. Statistics and Indicators on Gender Equality in Science', Brussels.

European Commission (2007), 'Remuneration of Researchers in the Public and Private Sectors. Final report', Brussels.

European Commission (2004), 'Gender and Excellence in the Making', Luxembourg.

Fernandez, P., McCarthy, I.P. and Rakotobe-Joel, T. (2001), 'An Evolutionary Approach to Benchmarking', *Benchmarking. An International Journal* 8 (4), 281–305.

Fielding, J. and Glover, J. (1999), 'Women and Science Graduates in Britain. The Value of Secondary Analysis of Large-Scale Data Set', *Work, Employment and Society*, vol. 13, no.2.

Foschi, M. (2000), 'Double Standards for Competence. Theory and Research', *Annual Review of Sociology*, vol. 26.

Funnell, S. (2000), 'Developing and Using a Program Theory Matrix for Program Evaluation and Performance Monitoring', in P. Rogers, T. Hacsí, A. Petrosino and T. Huebner (eds), *Program Theory in Evaluation: Challenges and Opportunities*, New Directions for Evaluation, 87, pp. 91-101. San Francisco, CA: Jossey-Bass.

Garforth, L. and Kerr, A. (2009), 'Women and Science: What's the problem', *Social Politics: International Studies in Gender, State and Society* **16**, 380–403.

Gerring, J. (2005), 'Causation. A unified framework for the social sciences', *Journal of Theoretical Politics* **17**.

Glass, C. and Minnotte, K. L. (2010), 'Recruiting and hiring women in STEM fields', *Journal of Diversity in Higher Education* **3**, 218–229.

Glouberman, S. (2001), 'Towards a New Perspective on Health Policy', CPRN Study no. H703. Canadian Policy Research Networks Inc., Ottawa.

Glouberman, S. and Zimmerman, B. (2002), 'Complicated and Complex Systems: What Would Successful Reform of Medicare Look Like?', Commission on the Future of Health Care in Canada, Discussion Paper 8. Available at [http:// www.hc-sc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf](http://www.hc-sc.gc.ca/english/pdf/romanow/pdfs/8_Glouberman_E.pdf).

Guba, E.G. (1981), 'Criteria for assessing the trustworthiness of naturalistic inquiries', *Educational Communication and Technology Journal*, 29 (2), 75-91.

Gupta, N., Kemelgor, C., Fuchs, S. and Etzkowitz, H. (2005), 'Triple Burden on Women in Science. A Cross- Cultural Analysis', *Current Science*, vol. 89, no. 8.

Gürer, D. and Camp T. (2002), 'Investigating the Incredible Shrinking Pipeline for Women in Computer Science, Final Report – NSF Project 9812016, Arlington.

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.

Ivey, E. (2005), 'Gender Differences among Contingent Faculty. A Literature Review, Final Report', AWIS, Washington.

Jensen, K.S.H. (2005), 'Women Working in Science, Engineering and Technology. Higher Education and Industry. A Literature review', IRIS (Informatics Research Institute), 6, Salford University, Manchester

Kalev, A., Dobbin, F. and 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. and Faber, S.T. (2016), 'Benefits of Peer Mentoring to Mentors, Female Mentees and Higher Education Institutions', *Mentoring & Tutoring: Partnership in Learning*, DOI: 10.1080/13611267.2016.1170560.

LeFeuvre, N. (2014), 'Women, Gender & Science: Modern Sexism in Action', Power point presentation, European Platform of Women Scientists, Paris, June 11, 2014.

Lawler, A. (2001), 'Women in Science. College Heads Pledge to Remove Barriers', *Science*, vol. 291, no. 5505.

Martin B. (2011), 'The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?', *Research Evaluation*, 20(3), 247-254.

Martín-García, T. (2009), 'Bring Men Back In. A Re-examination of the Impact of Type of Education and Educational Enrolment on First Births in Spain', *European Sociological Review*, vol. 25, no. 2

Matysiak, A. and Vignoli, D. (2008), 'Fertility and Women's Employment. A Meta-analysis', *European Journal of Population*, 24(4)

Meagher L., Lyall, C. and Nutley, S. (2008), 'Flows of knowledge, expertise and influence: a method for assessing policy and practice impacts from social science research', *Research Evaluation*, 17(3), 163-173

Messina, C., Bardouin, J., Auricchio, L. and Blanc, M.L. (2008), 'Good Practice in the Field of Regional Policy and Obstacles to the Use of the Structural Funds', Brussels: Directorate-General for Internal Policies of the European Union. Policy Department Structural and Cohesion Policy.

Meulders, D., Plasman, R., Lemièrre, S., Danis, S., O'Dorchai, S., Tojerow, I., Jepsen, M., Gangji, A., Moreno, D., Caprile, M. and Kruger, K. (2003), 'Women in industrial research – Analysis of statistical data and good practices of companies', European Commission, Luxembourg.

Molas-Gallart, J. and Tang, P. (2011), 'Tracing 'productive interactions' to identify social impacts: an example from the social sciences', *Research Evaluation*, 20(3), 219-226.

Morse, J.M., Barrett, M., Mayan M., Olson, K. and Spiers, J. (2002), 'Verification strategies for establishing reliability and validity in qualitative research', *International Journal of Qualitative Methods*, 1(2), 1-19.

Patton, M.Q. (1999), 'Enhancing the quality and credibility of qualitative analysis', *Health Services Research*, 34, 1189–1208.

Pawson, R. and Tilley, N. (1997) '*Realistic Evaluation*', London: SAGE.

Pinnegar, S. (2006), '*Are Complex Programs the Best Response to Complex Policy Issues?*', City Futures Research Centre Issues paper. Kensington, NSW: University of New South Wales.

PRAGES: Guidelines for Gender Equality Programs in Science (2009). Available, http://www.retepariopportunita.it/rete_pari_opportunita/userfiles/progetti/prages/pragesguidelines.pdf.

Reale, E., Nedeva, M., Thomas, D. and Primeri, E. (2014), 'Evaluation through impact: a different viewpoint', *Fteval journal*. Issue 39: September 2014. Vienna.

Reimers, K. and Johnston, R.B. (2008), 'The use of an explicitly theory-driven data coding method for high-level theory testing in IOIS', *ICIS Proceedings*, Association for Information Systems, 1-17.

Ryan, J.G. and Garrett, J.L. (2004), 'The impact of economic policy research: lessons on attribution and evaluation from IFPRI', in Stone, D. and S. Maxwell (eds) *Global Knowledge Networks and International Development: Bridges Across Boundaries*. Routledge: London.

Rogers, P. (2008), 'Using Programme Theory to Evaluate Complicated and Complex Aspects of Interventions', *Evaluation* **14**, 29–48.

Sanderson, I. (2000), 'Evaluation in Complex Policy Systems', *Evaluation* 6(4): 433-54.

Sartorius, R. (1991), 'The Logical Framework Approach to Project Design and Management', *Evaluation Practice* **12**(2), 139-47.

Schiebinger, L. (2008), 'Getting More Women into Science and Engineering – Knowledge Issues', in L. Schiebinger (ed) *Gendered Innovations in Science and Engineering*, pp. 1-21. Stanford University Press, Stanford.

Shen, H. (2013), 'Inequality Quantified: Mind the Gender Gap', *Nature* **495**, 23.

Sheridan, B. (2006), 'Straniere in terra straniera, *Conoscenza e Innovazione*', vol. 2

Spaapen, J. and Van Drooge, L. (2011), 'Introducing "productive interactions" in social impact assessment', *Research Evaluation* 20(3): 211-218.

Sonnert, G., Fox, M.F. and Adkins, K. (2007), 'Undergraduate Women in Science and Engineering. Effects of Faculty, Fields, and Institutions over Time', *Social Science Quarterly*, vol. 88, no. 5

Stafford, S.G. (1996), 'Finding Leadership Opportunities in an Era of Dual Careers', *BioScience*, vol. 46, no. 1.

Stame, N. (2004), 'Theory-based Evaluation and Types of Complexity', *Evaluation* 10(1): 58-76.

Stufflebeam, D.L. (2004), 'The 21st Century CIPP Model', in M.C. Alkin (ed) *Evaluation Roots*, pp. 245-66. Newbury Park, CA: SAGE.

Swim, J. K., Aikin, K. J., Hall, W. S. and Hunter, B. A. (1995), 'Sexism and Racism: Old-Fashioned and Modern Prejudices', *Journal of Personality and Social Psychology* **68**, 199–214.

Timmers, T. M., Willemsen, T. M. and Tijdens, K. G. (2010), 'Gender Diversity Policies in Universities: A Multi-Perspective Framework of Policy Measures', *Higher Education* **59**, 719–735.

Valian, V. (1998), *Why So Slow? The Advancement of Women*, MIT Press, Cambridge.

Von Eiff, W. (2015), 'International Benchmarking and Best Practice Management: In Search of Health Care and Hospital Excellence', *Advances in Health Care Management* **17**, 223–252.

Zimmer, A. (2003), 'Women in European Universities. Final Report 2000-2003 of the Research and Training Network', Available at: <http://www.women-eu.de>.

Wennerås, C. and Wold, A. (1997), 'Nepotism and Sexism in Peer-Review', *Nature*, vol. 387.