Research Evaluation, 28(2), 2019, 158–168
doi: 10.1093/reseval/rvy042
Advance Access Publication Date: 23 January 2019
Article

OXFORD

How is gender being addressed in the international development evaluation literature? A meta-evaluation

Steven Lam [1,*] Warren Dodd [2] Jane Whynot [3] and Kelly Skinner [2]

[1]Department of Population Medicine, University of Guelph, 50 Stone Road East, Guelph, ON N1G 2W1, Canada, [2]School of Public Health and Health Systems, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, and [3]Institute of Feminist and Gender Studies, University of Ottawa, 75 Laurier Avenue East, Ottawa, ON K1N 6N5, Canada

*Corresponding author. Email: lams@uoguelph.ca.

Abstract

Gender equity is an increasingly discussed priority and cross-cutting theme within international development evaluation. However, it is unclear whether advances being made in evaluating the outcomes in this area are reflected in the scholarly literature. In this context, a fundamental question is: How is gender being addressed in international development evaluation? To answer this question, we conducted a meta-evaluation to identify, synthesize, and assess published evaluation studies in international development with a focus on gender. We searched the Web of Science™ Core Collection database along with nine evaluation-focused journals using variations of the terms 'program evaluation' and 'gender'. A total of 2027 studies were identified, of which 70 met *a priori* inclusion criteria. Of the reviewed evaluations, many targeted gender-specific programs and specifically women. While the number of studies that report on gender is growing, and nearly all studies included gender-disaggregated data, often only outcomes by 'women' and 'men' were considered without going further to raise larger questions of gender equity. For evaluation to further contribute to gender equity, we suggest that future peer-reviewed evaluation studies provide data on diverse groups of genders, engage with evaluation stakeholders, consider the larger socio-cultural-political context of programming, encourage the use of evaluation findings, and provide actionable recommendations.

Key words: program evaluation; gender; meta-evaluation, systematic review methodology; international development

1. Introduction

A number of global efforts have been made over the last two decades to increase the prospect of incorporating gender in evaluation. In 2005, the United Nations Evaluation Group (UNEG) developed a set of norms and standards for evaluation in the UN system and highlighted the need for human rights and gender equality to be centered in evaluation (UNEG 2005a, 2005b). Later in 2011, UNEG published a handbook to guide the integration of these dimensions throughout the evaluation process (UNEG 2011) which prompted the adoption of gender integration in evaluation across the UN system (ILO 2014; UN Women 2014; UNDP 2014a; FAO 2017). Authors and editors are also encouraged to integrate assessment of sex and gender into all manuscripts by using the Sex and Gender Equity in Research (SAGER) guidelines (Heidari et al. 2018). These efforts were motivated by concerns surrounding gender equity, as well as the potential role of evaluation in contributing to gender equity.[1]

Integrating gender in evaluation involves considering not only gender-related outcomes of programs, but also gender dimensions in the evaluation design, processes, products, and use (Evaluation Cooperation Group 2017; FAO 2017). Evaluation matters for gender equity because it allows us to assess whether a program is

meeting the needs of all people—women, men, girls, boys, and gender-diverse people (World Bank 2012; Espinosa 2013; Heider 2015). Furthermore, it plays a crucial role in demonstrating the impact of programs on various genders and builds the case for considering gender in development programs. Finally, by providing information and recommendations, evaluation could improve interventions that address the needs of diverse individuals and by extension, contribute to greater gender equity (International Labour Organization 2014).

While there is increasing emphasis on understanding the gender dimensions of evaluation, concerns have been expressed about the progress made in this regard. Research suggested that gender mainstreaming in evaluation has not been widely carried out by the international development community (African Development Bank 2012; Govinda 2012; Tirivanhu and Jansen van Rensburg 2018). For instance, an assessment of the World Bank's performance showed that considerable progress was made in addressing gender issues over the past 15 years; however, the monitoring and evaluation frameworks currently do not adequately measure or report on gender results (World Bank 2016). Where gender is being integrated into evaluations, less is known about their quality and use. Given the large sums of money invested in international development programs by national governments, bilateral and multilateral donors, and nongovernmental organizations (NGOs), there is a need for assessments of, and reporting on, gender within evaluations.

Planning to collect gender-disaggregated data is often the first step in designing and evaluating effective international development programs (UNDP 2014b), with gender-responsive analysis key to understanding the conditions that different gender groups face. Yet, closing the gender equity gap goes beyond data collection and analysis. Greater efforts are required to raise awareness and promote the use of evaluation findings relating to gender. One often neglected step in such efforts is identifying and synthesizing existing program evaluations in a systematic manner. Assessing trends and gaps in the evaluation of international development programs is important because it reveals the current state of gender integration in evaluation.

A notable effort to assess gender considerations in evaluations is a review initiated by UNEG in 2016. The review explored reporting on the evaluation performance indicator (EPI) of the UN System Wide Action Plan (UN-SWAP) (Barnes and Bishop 2016). UN-SWAP is the first accountability framework for gender mainstreaming in the UN system, with EPI serving as a tool to help UN entities integrate gender equity and human rights into evaluation efforts. The initial sample to be reviewed included 378 evaluation reports across 25 UN entities, of which 46 minimum and high scoring evaluation reports were retained and assessed for trends, challenges, and good practices in meeting EPI requirements. The final sample was categorized as either 'missing' (11%), 'approaching' (33%), 'meeting' (37%), or 'exceeding' (19%) EPI requirements. As UN entities should be 'meeting' requirements, this review highlights opportunities for improving gender mainstreaming in evaluations among UN entities. Evaluation units of UN entities, such as those of UNDP and the World Bank, have also started to evaluate UN contributions to gender equity (Independent Evaluation Office 2015; Bardasi and Garcia 2016). These examples point to increasing action placed on evaluating gender mainstreaming within the UN system; however, the responses of researchers and evaluators outside the UN system are unclear.

International scholarly research, evaluation, and dialog predominately take place in peer-reviewed publications. As such, exploring the published evaluation literature represents one avenue for understanding broader responses to mainstreaming gender into evaluations. We aim to examine whether gender is incorporated in international development evaluation by using gender reporting in scholarly literature as a proxy for the integration of gender in evaluation processes and products. Further, we explore how gender is being measured by assessing these existing evaluations against selected evaluation standards. In doing so, we address an important gap in our understanding of how gender is being addressed within this scholarship. To guide our examination, we pose the following research question: How is gender being addressed in international development evaluation?

2. Materials and methods

We conducted a meta-evaluation to assess whether and how gender is incorporated in evaluations. A meta-evaluation may refer to the evaluation of a single evaluation or the evaluation of multiple evaluations (Hedler and Gibram 2009; Scriven 2009). In this paper, meta-evaluation refers to a synthesis of a number of related evaluations, with the purpose of identifying their strengths and limitations against a set of quality standards (Stufflebeam 2001; Cooksy and Caracelli 2009; Scriven 2009; Good 2012). To identify evaluations, we searched the peer-reviewed literature using systematic review approaches involving a stepwise procedure of search, selection, extraction, and synthesis of the literature (Arksey and O'Malley 2005; The Cochrane Collaboration 2008; Levac et al. 2010).

2.1 Search strategy

We searched for peer-reviewed articles on 5 July 2018 using the citation database Web of Science™ CORE Collection. This database was selected as it is one of the most comprehensive and widely used search engines available for retrieving peer-reviewed research from all scientific areas (Hightower and Caldwell 2010). We also searched some evaluation-focused journals including: *Evaluation and Program Planning*; *American Journal of Program Evaluation*; *Evaluation Review*; *Evaluation*; *Evaluation & The Health Professions*; *Evaluation Journal of Australasia*; *African Evaluation Journal*; *Canadian Journal of Program Evaluation*; and *Research Evaluation*. The search strategy used Boolean operators to pair the keyword evaluation and its synonyms with the keyword gender and its synonyms (Table 1). No time or language restrictions were applied to the search. All citations were imported into the web-based application DistillerSR© (Evidence Partners Incorporated, Ottawa, ON, Canada) for relevance screening.

2.2. Relevance screening

The titles and abstracts of articles were screened according to *a priori* inclusion criteria (Table 2). In phase one, all peer-reviewed English language articles that had a focus on program evaluation and gender were included. In phase two, all citations deemed relevant went through the second round of screening to determine whether they were relevant to development. Programs that focused on development in a low- or middle-income country (LMIC) according to 2018 World Bank classifications were included (World Bank 2018a). Programs that focused on Indigenous communities in high-income countries were also included. We consider such programs to be 'development' as they seek to reduce inequities between Indigenous communities and non-Indigenous communities. Finally,

Table 1. Search strategy to identify peer-reviewed evaluation studies that are reporting on gender

Database or journal	Search string
Web of Science™ CORE Collection database,[a] *Evaluation and Program Planning*, *American Journal of Program Evaluation*, *Evaluation Review*, *Evaluation*, *Evaluation & The Health Professions*, *Evaluation Journal of Australasia*, *African Evaluation Journal*	('program evaluation' OR 'program monitoring' OR 'program assessment' OR 'program measurement') AND (gender OR women OR men OR masculine OR feminine OR masculinity OR femininity)
Canadian Journal of Program Evaluation, *Research Evaluation*[b]	(evaluation AND gender) (gender)

[a]It is noteworthy that our focus was on evaluation; we acknowledge that the database used may not index all development, gender, and feminist journals that have evaluation-related papers.

[b]For *Canadian Journal of Program Evaluation* and *Research Evaluation*, shortened search strings were used as the original search string yielded zero hits.

Table 2. Inclusion criteria to screen for peer-reviewed studies that are reporting on gender

Inclusion	Exclusion
Phase one	
Peer-reviewed articles	Editorials, abstracts, commentaries, book reviews
Discusses gender in some capacity (e.g. men, women, masculinity, femininity)	Does not discuss gender in any way
Reports on an evaluation of a program	Reports on program development or implementation only, or evaluation theory only
Phase two	
Focuses on development programming (e.g. programs with communities in low- and middle-income countries or with Indigenous communities in high-income countries).	Focuses on a program in a high-income country with non-Indigenous communities

in some cases, a full-text review was conducted in order to assess suitability.

2.3 Data extraction and assessment

A charting form was developed to capture count data and descriptions of the study (Levac et al. 2010). Information extracted included the year of publication, study country, funding, and last name, gender, and affiliation of the first author. We determined the gender of the first author by inspection of their name, and if unclear, we used Google to find photographs and/or biographical paragraphs (Filardo et al. 2016). Descriptions of the program extracted included program scale and gender focus. Descriptions of the evaluation extracted included evaluation focus, approach, methods, and assessment type (i.e. external vs. internal). To determine the assessment type, we searched the affiliation, acknowledgement/contribution section, and description of the program and evaluation. We also considered whether studies referred to the five evaluation criteria from the Development Assistance Committee of the Organization for the Economic Cooperation and Development (OECD-DAC), the most common evaluation criteria in international development (OECD 2010). These include relevance, effectiveness, efficiency, impact, and sustainability.

To assess the quality of evaluations, we used selected professional evaluation standards: propriety and utility (Yarbrough et al. 2011). By the nature of selecting only evaluation studies in the peer-reviewed literature for review, we assumed that there has been some level of quality control (e.g. validity, credibility). As such, we sought to apply specific standards that could be relevant to gender (Table 3). For example, utility refers to meeting the needs of evaluation users and ensuring that evaluations provide relevant and useful information. Propriety is concerned with whether the evaluation engages with those most directly affected by the program while protecting their rights. Finally, to assess the extent to which

Table 3. Quality assessment criteria applied to peer-reviewed evaluation studies

Criteria	Questions
Propriety	Is there a statement of research ethics approval?
	Does the evaluation engage with those most directly affected by the program?
Utility	Does the paper acknowledge sharing of evaluation findings?
	Does the paper acknowledge use of evaluation findings?
Gender	Does the paper acknowledge socio-cultural-political context?
	Does the paper acknowledge diverse groups of genders (e.g. ethnicity, age, class, education, location)?
	Does the paper acknowledge gender issues in the introduction?[a]
	Does the evaluation question/objective of the paper reflect gender?[a]
	Are gender sensitive indicators provided?[a]
	Are gender-disaggregated data provided?[a]
	Are unintended consequences reported?[a]
	Do different gender groups benefit equally from the program?[a]
	Does the paper suggest actions to address gender issues?[a]

[a]Questions were asked to gender nonspecific programs only and were adapted from SAGER guidelines (Heidari et al. 2018) and UN-SWAP (UNEG 2018).

gender was integrated into evaluations, we adapted questions based on existing tools that support standardization of sex and gender reporting in publications (Heidari et al. 2018). Further, we adapted questions from an international assessment criteria for gender integration in evaluation (UNEG 2018).

2.4 Synthesis

We used a narrative synthesis approach to provide an overview of the existing literature. Firstly, an overall summary of the study findings was presented including the characteristics of the article, program, and evaluation. Then, study results were coded and organized into categories. Within the categories, the codes were analyzed and summarized into themes using thematic analysis techniques (Braun and Clarke 2006).

3. Results

The initial search returned 2027 articles; after removal of duplicates and nonrelevant articles, a total of 70 scholarly evaluation articles were included (see Supplementary Data for a complete list of included documents). Each article represented a single program. Most articles were excluded because they did not explicitly mention gender or its synonyms (Figure 1).

3.1 Reporting of gender in scholarly evaluations is limited but increasing

Evaluations with a gender focus first appeared in the peer-reviewed literature in 1991 (n = 2). Publications steadily increased annually from 2011 onwards and peaked in 2017 (n = 10) (Figure 2). Based on analysis of the affiliation of lead authors, studies were authored by academic institutions (63%, n = 44), followed by NGOs (26%, n = 18), government (9%, n = 6), and consultants (3%, n = 2). Most evaluations were conducted internally (67%, n = 47), of which funding was received from an international donor (n = 24), research institute (n = 6), government (n = 6), or a combination of funding sources (n = 11). In 14 studies (n = 20%), evaluations were externally conducted, of which many did not specify whether funding was received (n = 7). Only four of these studies reported receiving funding, and three did not receive funding. In nine studies (13%), it was unclear what role the authors played in the program or evaluation. Of the gender-identified first authorships, women represented 51% (n = 36) and men represented 37% (n = 26). The gender of the first author was unclear in 11% of articles (n = 8).

Articles were categorized as either gender-specific (63%, n = 44) or non-gender specific (37%, n = 26) depending on whether articles focused on one gender or multiple genders, respectively. Of the gender-specific programs (n = 44), many targeted women only (66%, n = 29) or women and children (20%, n = 9). Limited articles focused on girls only (7%, n = 3), men only (5%, n = 2), and transgender persons (2%, n = 1). Of the gender nonspecific programs (n = 26) (Table 4), women and men (81%, n = 21), followed by girls and boys (19%, n = 5). Nearly half of all included articles (49%, n = 34) provided disaggregated data of the identities and subjectivities of participants beyond gender, including age (41%, n = 29), socio-economic status (19%; n = 13), education (17%; n = 12), ethnicity (6%; n = 4), and location (4%; n = 3).

3.2 Many gender-focused scholarly evaluations were conducted in Africa and Asia

The reviewed articles were conducted in 55 countries (Figure 3). Regionally, many studies were conducted in Africa (51%, n = 59), Asia (30%, n = 35), and North America (9%, n = 10). In

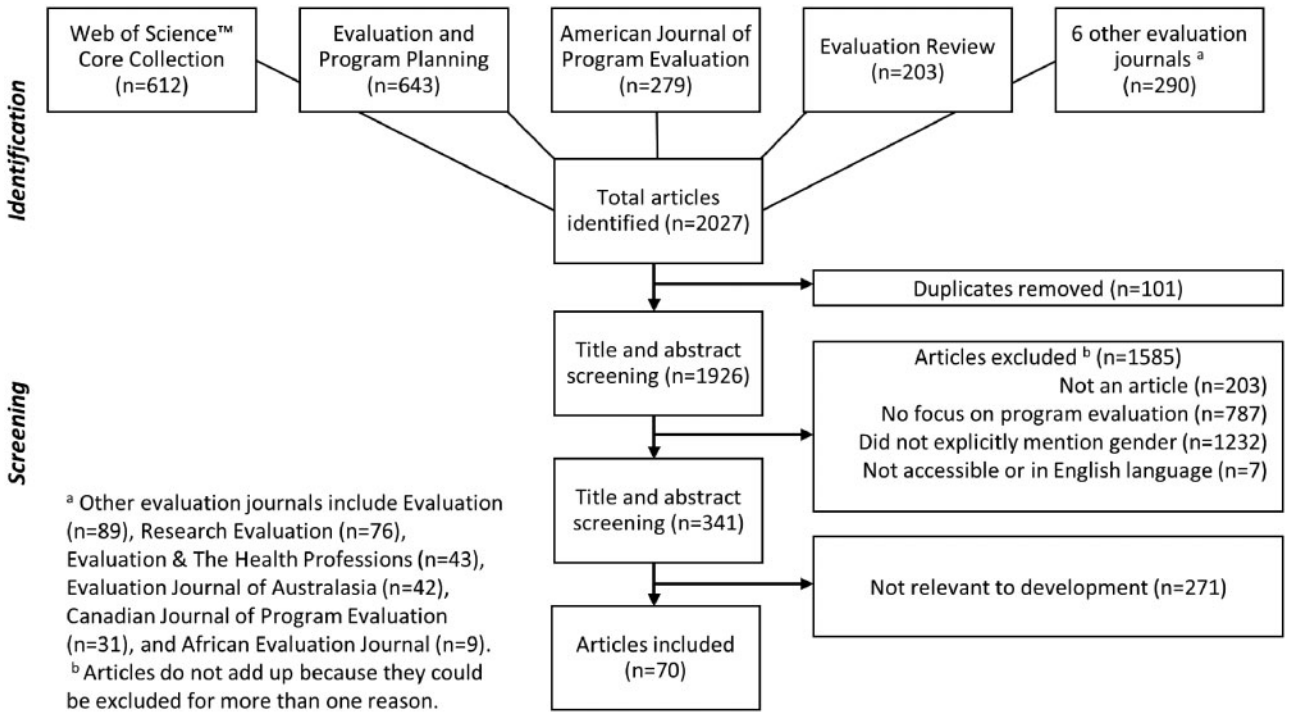

Figure 1. Flow chart of the selection of peer-reviewed evaluation studies in international development with a gender focus.

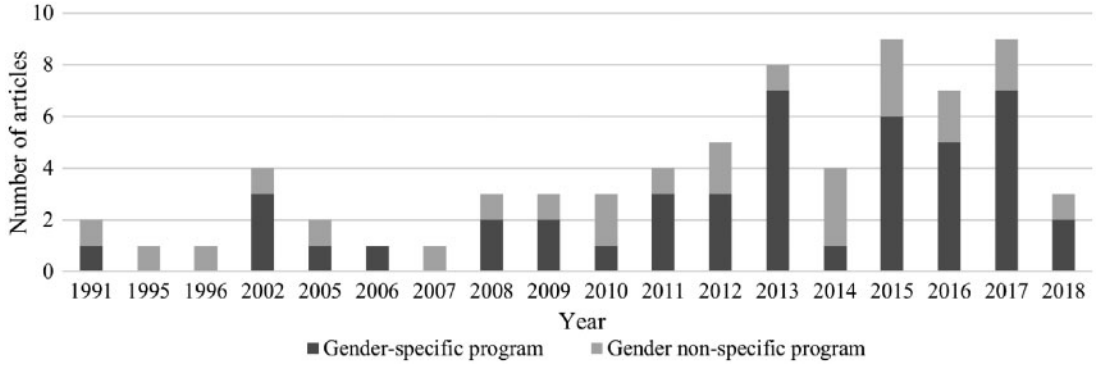

Figure 2. Number of peer-reviewed evaluation studies in international development with a gender focus over time.

comparison, research in South America (5%, n = 6), Australia (2%, n = 2), and Europe (3%, n = 4) were limited likely due to search criteria around LMICs. Several articles (9%, n = 7) focused on Indigenous communities, of which two were from high-income countries (Skinner et al. 2012; Kelley et al. 2015).

Many programs were conducted at the local or pilot scale (66%, n = 46), followed by national (20%, n = 14) and provincial/regional (14%, n = 10) levels. Evaluations focused on outcomes (46%, n = 32), impact (24%, n = 17), process (10%, n = 14), or some combination of evaluation foci (16%, n = 11). More than half of articles (54%, n = 38) did not specify an evaluation approach. Rather, many of these articles (81%, n = 31) framed the methods as the evaluation approach (e.g. qualitative evaluation, mixed-methods evaluation), followed by presenting a description of the program evaluation process or outcomes (11%, n = 4), and a qualitative or quantitative analysis of a specific aspect of the program evaluation (8%, n = 3). Where an evaluation approach was specified (46%, n = 32), they mainly included quasi-experimental (59%, n = 19) or quasi-experimental without a control group (19%, n = 6). Few articles used randomized control trials (13%, n = 4) or participatory approaches (9%, n = 3). Methods used included quantitative only (70%, n = 49), mixed quantitative and qualitative (16%, n = 11), or qualitative only (14%, n = 10). While no articles made explicit reference to the OECD-DAC criteria, most articles (79%, n = 55) referenced at least one of the five evaluation standards: effectiveness (n = 39), impact (n = 20), efficiency (n = 7), sustainability (n = 5), and relevance (n = 3). Notably many articles (n = 21) referenced more than one standard. Health was the main sector of programming (= 73%, n = 51), followed by agriculture, nutrition, and food (14%, n = 10), education (9%, n = 6), employment (3%, n = 2), and violence prevention (1%, n = 1). No articles explicitly explored the issues of environmental sustainability or climate change adaptation.

3.3 Low levels of reporting on engagement with evaluation stakeholders

In the context of propriety standards, we considered what is proper, fair, legal, right, and just in evaluations by using research ethics approval as a proxy. Although not all journals require an ethics statement, we found that 63% (n = 44) articles shared an ethics statement, and in some cases ethics approval was sought from relevant review boards and deemed exempt (e.g. Zhongdan et al., 2008; Ippolito et al., 2017). We considered to what extent evaluations engaged with stakeholders and found only two studies (n = 3%) that reported some level of engagement. For example, Limato et al. (2018) developed an approach that incorporated a loop of feedback to the beneficiaries and program implementers to directly engage them in the evaluation process. Similarly, Nandi et al. (2015: 38) reported engaging stakeholders in the evaluation to varying degrees, including: 'defining the objectives, designing questions, data collection and data analysis in the context of their aspirations and expectations'.

The literature reviewed provided insights into how evaluation findings were shared and used. In general, most articles aimed to advance knowledge or share lessons learned from the program and evaluation (90%, n = 62). Several articles (9%, n = 6) explicitly mentioned how findings were used, while two articles (3%) explicitly mentioned how findings were shared. In a collaborative NGO-academic institution partnership in Malawi, Weinhardt et al. (2014: 11) reported: 'Results are fed back to the implementers on a more real-time basis. The NGO, thus, has the opportunity to adjust future programming based on this study's findings'. Maticka-Tyndale et al. (2010: 184) shared results of an HIV prevention program in Kenya by presenting them to '[the Ministry of Education, Science & Technology], and donor agency staff and representatives following pre-program data collection and again at the end of the evaluation'.

3.4 The broader socio-cultural-political environment motivated some studies

Around 21% of articles (n = 15) discussed the socio-cultural-political environments in which gender inequities and programs were situated. In the context of a nutrition program in India, gender differences in growth and dietary intake were examined 'because boys are typically favored over girls in South Asian countries' (Avula et al. 2011: 681). Exploring an HIV prevention program in sub-Saharan Africa, the authors stated: 'Knowledge of HIV positive status is of great social consequence, and stigma and violence are still actual threats for African women' (Audureau et al. 2013: 5). In Australia, maternal and child health programs were supported since 2008 when 'the Australian Government recognized that improving maternal and child health also had the potential to play an important role in creating the foundations for improved Indigenous health and wellbeing throughout the lifetime and pledged to halve the gap in mortality rates for children under five by 2018' (McCalman et al. 2015: 2).

Table 4. Descriptions of peer-reviewed evaluation studies of international development programs with a focus on gender

Author and year	Study country	Program sector	Are gender-sensitive indicators provided?	Are gender-disaggregated data presented?	Do outcomes differ by gender?	Which gender group benefited more?	Are actions to address gender issues suggested?
Popescu and Roman (2018)	Romania	Employment training	✓	✓	✓	Women	✗
Lee et al. (2002)	Zimbabwe	HIV prevention	✓	✓	✓	Women	✓
Zhongdan et al. (2008)	China	HIV prevention	✓	✓	✓	Women	✗
Ippolito et al. (2017)	Guatemala	Health-care use	✓	✓	✓	Women	✗
Kaufman et al. (2015)	Tanzania	HIV prevention	✓	✓	✓	Not clear (qualitative findings)	✓
Othman and Nasrudin (2016)	Malaysia	Education	✓	✗	N/A	N/A	NA
Mendola and Simtowe (2015)	Malawi	Rural land development	✓	✓	✗	N/A	N/A
Berti et al. (2015)	Honduras	Maternal health	✓	✓	✓	Different outcomes for women and men	✓
Weinhardt et al. (2014)	Malawi	Food security	✓	✗	N/A	N/A	N/A
Swamy (2014)	India	Economic development	✓	✓	✓	Women	✗
Naidoo and Johnson (2013)	Namibia	HIV prevention	✓	✓	✓	Men	✗
Lambdin et al. (2012)	Mozambique	HIV prevention	✓	✗	N/A	N/A	N/A
Avula et al. (2011)	India	Nutrition	✓	✓	✓	Boys	✓
Abebaw et al. (2010)	Ethiopia	Food security	✓	✓	✓	Women	✗
Reis et al. (2010)	Brazil	Physical activity promotion	✓	✓	✓	Men	✗
Carvalho et al. (2009)	Brazil	Disease prevention	✓	✓	✓	Women	✓
D'Agnes et al. (2005)	Philippines	Food security	✓	✓	✓	Women	✓
Jegede and Okebukola (1996)	Nigeria	Education	✓	✓	✗	N/A	N/A
Anantha et al. (1995)	India	Tobacco reduction	✓	✓	✓	Women	✗
Lee and Shute (1991)	Mali	Agriculture	✓	✓	✓	Men	✗
Nandi et al. (2015)	India	Nutrition and health	✓	✓	✗	N/A	N/A
Liao (2014)	China	Education	✓	✓	✓	Girls	✗
Maticka-Tyndale et al. (2010)	Kenya	HIV prevention	✓	✓	✓	Girls	✗
Duman and Akbaş (2017)	Turkey	Education	✓	✓	✓	Girls	✗
Skinner et al. (2012)	Canada	Nutrition and health	✓	✓	✓	Boys	✗
Aker and Ksoll (2016)	Niger	Agriculture	✓	✓	✓	Women	✗

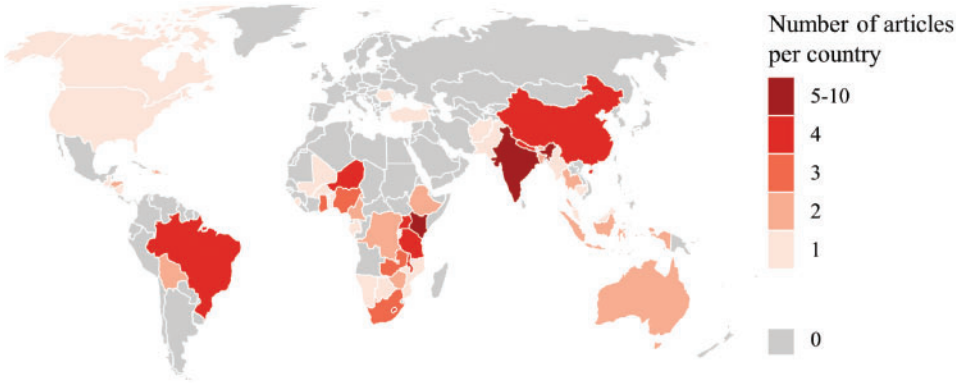

Figure 3. Geographic distribution of gender-focused evaluation studies in international development.

3.5 Gender relations in gender nonspecific programs

Among the articles categorized as gender nonspecific, more than half (56%, n = 14) considered gender in the introduction with some (28%, n = 7) reflecting gender in the evaluation question or objective of the paper. All articles explored gender-sensitive indicators by providing a breakdown of participants by gender. Gender disaggregated data on programming outcomes or experiences were also presented in nearly all articles. In two cases (3%), data on the number of participants by gender were collected but no analysis was done to determine the difference in program outcomes or experiences by gender (Lambdin et al. 2012; Othman and Nasrudin 2016). Gender considerations in data collection were elaborated in only one study where interviews and focus group discussions were conducted by 'gender-matched local fieldworkers' (Kaufman et al. 2015: 997).

In more than three quarters of gender nonspecific programs (77%, n = 20), program experiences between gender groups differed, with 10 programs benefiting women more than men. For example, an impact evaluation of a financial inclusion program in India found greater impact for women than men due to 'awareness levels and access to instruments of economic progress' (Swamy 2014: 14). For some of these studies, authors explained that there was more impact for women than men due to little emphasis on men as participants in the program (Lee et al. 2002; Zhongdan et al. 2008). In a few studies, however, programs benefited men more than women (Lee and Shute 1991; Reis et al. 2010; Naidoo and Johnson 2013). For example, an impact evaluation of an HIV prevention program in Namibia found that the program significantly reduced multiple sexual partnerships among men, but not among women (Naidoo and Johnson 2013). Among the limited studies (n = 5) that explored differences in outcomes among children, three of them reported greater benefits for girls compared to boys. Finally, no articles reported unintended program impacts on participants.

3.6 Limited studies highlight the need to address gender inequity

We identified limited studies (n = 6) directly recommending actions that address gender inequity within the programming context. Several papers highlighted the need to increase the involvement of men in the program (Lee et al. 2002; Berti et al. 2015; Kaufman et al. 2015). In the context of a maternal health intervention in Honduras, for example, Berti et al. (2015: 96) stated: 'Engaging "men as agents of positive change" moves from supporting women's health to promoting gender equity. This requires the full participation of men and women to serve the interests, survival, and well-being of all family members' (original quotations used). The authors caution, however, for the need to consider whether such actions are appropriate given the culture and context. An impact evaluation of a nutrition program in India highlighted the need to support girls: 'Given the propensity to favor male children, it is important to mitigate gender bias in food allocation and caregiving through complementary feeding counseling and to create an environment where boys and girls receive adequate and equal attention to achieve potential growth and development' (Avula et al. 2011: 684).

4. Discussion

We find that while gender is an increasing focus in both evaluation practice and published evaluation studies in international development, reporting on gender remains limited. We documented 70 gender-focused evaluation studies in LMICs or with Indigenous communities in high-income countries, representing only 21% of the 341 gender-focused evaluation studies identified through our search strategy. In contrast, then, 79% of gender-focused evaluation studies were conducted in high-income countries or did not focus on reducing disparities between Indigenous and non-Indigenous communities in high-income countries. The results of our review, particularly the inclusion of gender-disaggregated data in nearly all of the reviewed studies, show that gender is consistently considered in program evaluations. This trend is perhaps in response to the growing calls for studies to provide gender-disaggregated data (Runnels et al. 2014; Heidari et al. 2018) along with the commitment among the United Nations family of organizations and the World Bank to integrate gender in evaluation (UNEG 2011). The Gender Data Portal of the World Bank, for example, is another new international response to promote sharing of gender data and include topics ranging from health and education to jobs and political participation (World Bank 2018b).

Articles often considered the experience of women with few articles explicitly focusing on men and only one study focusing on gender-diverse indivduals (e.g. transgender persons) (Subramanian et al. 2013). The absence of program evaluations targeting transgender persons could be due to the lack trans-inclusive sex/gender measures (Reisner et al. 2015) or discrimination in countries where evaluations take place. In 2017, a new Multidimensional Sex/Gender Measure was proposed to support trans-inclusive

assessments (Bauer et al. 2017). The predominant focus on women in evaluations is a response to broader concerns of experiences of marginalization and discrimination. These developments are welcomed, however, several studies also emphasized the need to engage other genders in addition to women participants (Merrill et al. 2014; Berti et al. 2015; Kaufman et al. 2015). Furthermore, gender equity in evaluation means exploring the wider impact of development initiatives beyond impact on diverse gender groups to how it influences gender relations (Espinosa 2013). Adopting a holistic approach in gender assessments could help researchers and practitioners consider the impact of programs on broader gender relations (Schmidt and Cacace 2017).

Our review of gender-focused evaluations in the scholarly literature suggests that reporting on the engagement process with stakeholders, along with sharing and use of evaluation findings, tend to be marginal. Similarly, articles that report on recommendations and actions to address gender issues within the context of programming were limited. Indeed, evaluation has been critiqued for its lack of focus on evaluation use, which are implicitly aligned with gender approaches (Patton 2000; Podems 2010). However, given that many articles were outcome- and impact-focused, authors may have prioritized results over process and implementation in the peer-reviewed literature. Nevertheless, the sharing of findings and recommendations is a pivotal moment for learning and improvement. Engaging with program beneficiaries, implementers, and partners in the evaluation is crucial for supporting the use of results and in ownership in the process (Patton 2000; Espinosa 2013).

Nearly half of reviewed articles provided disaggregated data by diverse groups of genders, highlighting that different categories of gender, such as 'men' and 'women', do not form homogenous groups. Collecting these disaggregated data could help to better understand the nuanced impact of programs; for instance, in an agricultural development project in Niger, differential benefits were found among subpopulations of household residence and socioeconomic variables (Aker and Ksoll 2016). There is significant under-reporting of gender disaggregated data in Indigenous populations, which is a particular concern in light of the often heightened disparities among Indigenous peoples (United Nations Declaration 2008). Without additional disaggregation, of data, it is often assumed that what works for one group may work for another. Evaluations should provide feedback on how a program's various activities affect diverse groups of genders, with corrective action taken to address disparities in the distribution of benefits.

Among the articles that specified the evaluation approach, quasi-experimental and RCT approaches (91%) were commonly used. This finding could be explained by: (1) the high amount of outcome- and impact-oriented evaluations (70%); (2) health as the predominant sector of programming (73%); and (3) the focus on scholarly evaluation publications which generally prioritizes experimental approaches. Many researchers and practitioners also quantified the outcome of programs, with some using mixed methods, and few referring to program outcomes qualitatively. While quantitative evaluations are important for understanding outcomes on different gender groups, advances in gender equity are not always quantifiable. Qualitative evaluations offer useful information on barriers and opportunities for improving program processes and outcomes (Ramakrishnan et al. 2012; Osur et al. 2013; Kaufman et al. 2015). Several authors suggest that developing qualitative indicators using participatory approaches that actively involve beneficiaries and stakeholders to build ownership of the evaluative process (Lee et al. 2002; Nandi et al. 2015; Limato et al. 2018). Including a diversity of approaches, especially qualitative and participatory approaches, could help better understand how gender may differentially shape evaluation outcomes and experiences.

The OECD-DAC quality standards for evaluation encourages that 'the evaluation questions also address cross-cutting issues, such as gender, environment and human rights' (OECD 2010: 9). Of note, among the gender nonspecific programs, gender was incorporated in the evaluation question in only seven studies. However, peer-reviewed articles are not required to follow the OECD-DAC criteria, and indeed, no articles specifically referenced the criteria. Nevertheless, many articles did reference an evaluation standard, particularly effectiveness and impact. This finding highlights opportunities to bring gender equity to the forefront, starting at the creation of evaluation question. Engaging with gender in evaluation also requires contextual consideration that goes beyond the usual focus on processes and outcomes. Although evaluation reports should describe the development context according to OECD-DAC, we found that around 21% of articles (n = 15) discussed broad socio-cultural-political environments. Understanding these environments allow for programs to be designed to address not only the differences in gender outcomes but also broader factors contributing to these outcomes (Lee and Shute 1991; D'Agnes et al. 2005; Avula et al. 2011). If gender is to be engaged in a meaningful way, evaluation studies need to move beyond just presenting gender-disaggregated data, to critically examine the underlying socio-cultural-political processes that determine gender inequity and prompt gender-responsive action (United Nations Development Programme 2014b). As evaluators become more cognizant of these issues, they can better consider components of evaluation design that encourage more transformational change.

Without measuring and reporting on gender in evaluations, gender and related inequities are often overlooked. A question that may be asked is whether it is appropriate to evaluate gender processes and outcomes when the program does not have a focus on gender. For example, the main motivation of the study by Mendola and Simtowe (2015) was to improve household well-being through rural land development. While perhaps not explicitly stated, we argue all development programs that involve humans have some impact on gender, and the impact may be different for diverse gender groups, intended or unintended. One of the main contributions of this study is to serve as a reminder for researchers and practitioners to design, implement, and evaluate programs with a gender equity lens.

4.1 Limitations

We note several limitations of this approach including a focus solely on peer-reviewed articles. We acknowledge that many evaluation reports are confidential or documented outside of peer-reviewed literature (e.g. grey literature). While including grey literature would improve the breadth of insights, it would add considerable complications for quality assessment given the diversity of evaluation methodologies and reporting. Furthermore, although our purpose was to assess gender and evaluation in the scholarly context, we found that many studies (n = 26) were lead-authored by NGOs, government, and consultants, thus highlighting the diversity of perspectives captured. Secondly, articles were evaluated based on information presented in the article. We note that articles themselves may not fully explain the extent to which gender concepts were integrated into a program evaluation; as such, gender considerations may not have

been documented, or instead, overstated. Moreover, in some studies (n = 9) it was unclear what role the authors played in the program or evaluation, thus hindering a complete understanding of evaluation context. We recommend that authors make explicit their role in the evaluation. Finally, article screening and data extraction were done by one author (S.L.), presenting possible concerns over reviewer bias. To address this bias, this author discussed the reviewing strategy with coauthors and refined the strategy in the process. Nevertheless, the review is comprehensive and global in scope and provides a baseline for gender reporting in the evaluation of international development programs.

5. Conclusion

Reviewing and evaluating the published literature offers a means of understanding what we know about gender in evaluation. Here, we used meta-evaluation approaches to examine if and how gender is integrated in evaluation. Our results offer preliminary insights on how gender integration is taking place in the evaluation of international development programs. Assuming that the number of evaluations overall has remained steady, we find that gender reporting is limited but increasing, with a strong focus on programs that target women. Gender is reflected in the evaluation objectives in many articles and the findings reflect a degree of gender analysis in nearly all articles. Exploring these trends also highlights key research gaps in this area. Namely, researchers and practitioners should investigate the gendered experience across diverse groups of genders along with the need to consider the larger socio-cultural-political context in which gender is constructed and contested. Finally, evaluation could further contribute to gender equity by increasing efforts to encourage the use of evaluation findings and provide recommendations for action.

Note

1. Note that the United Nations often use the term equality to describe what some other agencies would define as equity; for consistency purposes, we used the term equity.

Supplementary data

Supplementary data is available at *Research Evaluation Journal* online.

Acknowledgments

The motivation for this work was driven by thought-provoking conversations held at the 2018 Canadian Evaluation Society conference. Thank you to the two anonymous reviewers for their insightful comments which helped to improve the paper.

Conflict of interest statement. None declared.

References

Abebaw, D., Fentie, Y., and Kassa, B. (2010) 'The Impact of a Food Security Program on Household Food Consumption in Northwestern Ethiopia: A Matching Estimator Approach', *Food Policy*, 35/4: 286–93.

African Development Bank. (2012) *Mainstreaming Gender Equality: A Road to Results or a Road to Nowhere*. Tunis, Tunisia.

Aker, J. C., and Ksoll, C. (2016) 'Can Mobile Phones Improve Agricultural Outcomes? Evidence from a Randomized Experiment in Niger', *Food Policy*, 60: 44–51.

Anantha, N. et al. (1995) 'Efficacy of an Anti-tobacco Community Education Program in India', *Cancer Causes and Control*, 6/2: 119–29.

Arksey, H., and O'Malley, L. (2005) 'Scoping Studies: Towards a Methodological Framework', *International Journal of Social Research Methodology*, 8/1: 19–32.

Audureau, E. et al. (2013) 'Scaling up Prevention of Mother-to-child HIV Transmission Programs in Sub-Saharan African Countries: A Multilevel Assessment of Site-, Program- and Country-Level Determinants of Performance', *BMC Public Health*, 13/1: 286.

Avula, R. et al. (2011) 'Enhancements to Nutrition Program in Indian Integrated Child Development Services Increased Growth and Energy Intake of Children', *Journal of Nutrition*, 141/4: 680–4.

Bardasi, E., and Garcia, G. (2016) *Integrating Gender into IEG Evaluation Work*. Washington, DC <https://www.oecd.org/dac/evaluation/Integrating-Gender-into-IEG-Evaluation-Work.pdf> accessed November 15, 2018.

Barnes, J., and Bishop, J. (2016) *Review of UN SWAP Evaluation Performance Indicator Reporting*. England: ImpactReady.

Bauer, G. R. et al. (2017) 'Transgender-Inclusive Measures of Sex/gender for Population Surveys: Mixedmethods Evaluation and Recommendations', *PLoS One*, 12/5: 1–28.

Berti, P. R. et al. (2015) 'An Adequacy Evaluation of a Maternal Health Intervention in Rural Honduras: The Impact of Engagement of Men and Empowerment of Women', *Revista Panamericana de Salud Publica*, 37/2: 90–7 8p.

Braun, V., and Clarke, V. (2006) 'Using Thematic Analysis in Psychology Using Thematic Analysis in Psychology', *Qualitative Research in Psychology*, 3/2: 77–101.

Carvalho, B. G. et al. (2009) 'Diseases of the Circulatory System before and after the Family Health Program, Londrina, Paraná', *Arquivos Brasileiros de Cardiologia*, 93/6: 597–601.

Cooksy, L. J., and Caracelli, V. J. (2009) 'Metaevaluation in Practice: Selection and Application of Criteria', *Journal of MultiDisciplinary Evaluation*, 6/11: 1–15.

D'Agnes, H. et al. (2005) 'Gender Issues within the Population-Environment Nexus in Philippine Coastal Areas', *Coastal Management*, 33/4: 447–58.

Duman, S. N., and Akbaş, O. (2017) 'Evaluation of Turkish and Mathematics Curricula According to Value-Based Evaluation Model', *Cogent Education*, 4/1. DOI: 10.1080/2331186X.2017.1291174

Espinosa, J. (2013) 'Moving towards Gender-Sensitive Evaluation? Practices and Challenges in International-Development Evaluation', *Evaluation*, 19/2: 171–182. DOI: 10.1177/1356389013485195

Evaluation Cooperation Group. (2017) *Integrating Gender into Project-level Evaluation*. Abidjan.

FAO. (2017) *Guidelines for the Assessment of Gender Mainstreaming*. Rome, Italy <http://www.fao.org/3/a-bd714e.pdf> accessed November 15, 2018.

Filardo, G. et al. (2016) 'Trends and Comparison of Female First Authorship in High Impact Medical Journals: Observational Study (1994–2014)', *BMJ*, 352: i847.

Good, B. (2012) 'Assessing the Effects of a Collaborative Research Funding Scheme: An Approach Combining Meta-Evaluation and Evaluation Synthesis', *Research Evaluation*, 21: 381–91.

Govinda, R. (2012) 'Mapping "Gender Evaluation" in South Asia', *Indian Journal of Gender Studies*, 19/2: 187–209.

Hedler, H. C., and Gibram, N. (2009) 'The Contribution of Metaevaluation to Program Evaluation: Proposition of a Model', *Journal of MultiDisciplinary Evaluation*, 6/12: 210–23.

Heidari, S., Babor, T. F., De Castro, P., Tort, S. and Curno, M. (2018) 'Sex and Gender Equity in Research: Rationale for the SAGER Guidelines and Recommended Use', *Research Integrity and Peer Review, 1/2*. DOI: 10.1186/s41073-016-0007-6.

Heider, C. (2015) 'Why Evaluation Matters for Gender Equality'. Retrieved from <https://ieg.worldbankgroup.org/blog/why-evaluation-matters-gender-equality> accessed November 15, 2018.

Hightower, C., and Caldwell, C. (2010) 'Shifting Sands: Science Researchers on Google Scholar, Web of Science, and PubMed, with Implications for

Library Collections Budgets', *Issues in Science and Technology Librarianship*, 63. DOI: 10.5062/F4V40S4J

Independent Evaluation Office. (2015) *Evaluation of UNDP Contribution to Gender Equality and Women's Empowerment*. New York.

International Labour Organization. (2014) 'Integrating Gender Equality in Monitoring and Evaluation of Projects' <http://www.ilo.org/wcmsp5/groups/public/—ed_mas/—eval/documents/publication/wcms_165986.pdf> accessed November 15, 2018.

Ippolito, M. et al. (2017) 'Expectations of Health Care Quality among Rural Maya Villagers in Sololá Department, Guatemala: A Qualitative Analysis', *International Journal for Equity in Health*, 16/1: 51.

Jegede, O. J., and Okebukola, P. A. O. (1996) 'Students' Ranking of and Opinions about the Standards of Learning in Nigerian Science Education Program', *Journal of Research in Science Teaching*, 33/6: 665–75.

Kaufman, M. R. et al. (2015) 'Protect Your Loved Ones from Fataki": Discouraging Cross-Generational Sex in Tanzania', *Qualitative Health Research*, 26/7: 994–1004.

Kelley, S., DeCourtney, C., and Thorsness, J. (2015) 'Development and Evaluation of a Support Program for Prostate Cancer Survivors in Alaska', *International Journal of Circumpolar Health*, 74: 28605.

Lambdin, B. H. et al. (2012) 'An Assessment of the Accuracy and Availability of Data in Electronic Patient Tracking Systems for Patients Receiving HIV Treatment in Central Mozambique', *BMC Health Services Research*, 12/1: 30.

Lee, R. A., and Shute, J. C. M. (1991) 'An Approach to Naturalistic Evaluation: A Study of the Social Implications of an International Development Project', *Evaluation Review*, 15/2: 254–65.

Lee, T. et al. (2002) 'Families, Orphans and Children under Stress in Zimbabwe', *Evaluation and Program Planning*, 25/4: 459–70.

Levac, D., Colquhoun, H., and O'Brien, K. K. (2010) 'Scoping Studies: Advancing the Methodology', *Implementation Science: IS*, 5: 69.

Liao, M.-Y. (2014) 'An Evaluation of an Airline Cabin Safety Education Program for Elementary School Children', *Evaluation and Program Planning*, 43: 27–37.

Limato, R. et al. (2018) 'Use of Most Significant Change (MSC) Technique to Evaluate Health Promotion Training of Maternal Community Health Workers in Cianjur District, Indonesia', *Evaluation and Program Planning*, 66: 102–10.

Maticka-Tyndale, E., Wildish, J., and Gichuru, M. (2010) 'Thirty-Month Quasi-Experimental Evaluation Follow-up of a National Primary School HIV Intervention in Kenya', *Sex Education*, 10/2: 113–30.

McCalman, J. et al. (2015) 'Empowering Families by Engaging and Relating Murri Way: A Grounded Theory Study of the Implementation of the Cape York Baby Basket Program', *BMC Pregnancy and Childbirth*, 15/1: 119. DOI: 10.1186/s12884-015-0543-y

Mendola, M., and Simtowe, F. (2015) 'The Welfare Impact of Land Redistribution: Evidence from a Quasi-Experimental Initiative in Malawi', *World Development*, 72: 53–69.

Merrill, K. G. et al. (2014) 'Linking At-Risk South African Girls to Sexual Violence and Reproductive Health Services: A Mixed-Methods Assessment of a Soccer-Based HIV-Prevention Programme and Pilot SMS Campaign', *Evaluation and Program Planning*, 70: 12–24.

Naidoo, R., and Johnson, K. (2013) 'Community-Based Conservation Reduces Sexual Risk Factors for HIV Among Men', *Globalization and Health*, 9/1: 27.

Nandi, R., Nanda, R., and Jugran, T. (2015) 'Evaluation from inside out: The Experience of Using Local Knowledge and Practices to Evaluate a Program for Adolescent Girls in India through the Lens of Gender and Equity', *Evaluation Journal of Australasia*, 15/1: 38–47.

OECD. (2010) *Quality Standards for Development Evaluation*. Paris, France.

Osur, J. et al. (2013) 'Implementation of Misoprostol for Postabortion Care in Kenya and Uganda: A Qualitative Evaluation', *Global Health Action*, 6/1: 1–11.

Othman, N., and Nasrudin, N. (2016) 'Entrepreneurship Education Programs in Malaysian Polytechnics', *Education and Training*, 58/7–8: 882–98.

Patton, M. Q. (2000) 'Utilization Focused Evaluation', *Evaluation Models*, 49: 425–438. DOI: 10.2307/447919

Podems, D. R. (2010) 'Feminist Evaluation and Gender Approaches: There's a Difference?', *Journal of MultiDisciplinary Evaluation*, 6/14: 1–17.

Popescu, M. E., and Roman, M. (2018) 'Vocational Training and Employability: Evaluation Evidence from Romania', *Evaluation and Program Planning*, 67: 38–46.

Ramakrishnan, U. et al. (2012) 'Public Health Interventions, Barriers, and Opportunities for Improving Mate. . .: EBSCOhost', *Food and Nutrition Bulletin*, 33/2: S104–37.

Reis, R. S. et al. (2010) 'Promoting Physical Activity through Community – Wide Policies and Planning: Findings from Curitiba, Brazil', *Journal of Physical Activity and Health*, 7: 137–45.

Reisner, S. L. et al. (2015) 'Counting" Transgender and Gender-Nonconforming Adults in Health Research', *TSQ: Transgender Studies Quarterly*, 2/1: 34–57.

Runnels, V. et al. (2014) 'The Challenges of Including Sex/Gender Analysis in Systematic Reviews: A Qualitative Survey', *Systematic Reviews*, 3/1: 33.

Schmidt, E. K., and Cacace, M. (2017) 'Addressing Gender Inequality in Science: The Multifaceted Challenge of Assessing Impact', *Research Evaluation*, 26/2: 102–14.

Scriven, M. (2009) 'Meta-Evaluation Revisited', *Journal of MultiDisciplinary Evaluation*, 6/11: iii–viii.

Skinner, K. et al. (2012) 'Impact of a School Snack Program on the Dietary Intake of Grade Six to Ten First Nation Students Living in a Remote Community in Northern Ontario', *Canada', Rural and Remote Health*, 12/3: 2122.

Stufflebeam, D. L. (2001) 'The Metaevaluation Imperative', *American Journal of Evaluation*, 22/2: 183–209.

Subramanian, T. et al. (2013) 'Increasing Condom Use and Declining STI Prevalence in High-risk MSM and TGs: Evaluation of a Large-Scale Prevention Program in Tamil Nadu, India', *BMC Public Health*, 13/1: 857.

Swamy, V. (2014) 'Financial Inclusion, Gender Dimension, and Economic Impact on Poor Households', *World Development*, 56: 1–15.

The Cochrane Collaboration. (2008) *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5. Chichester, England: The Cochrane Collaboration.

Tirivanhu, P., and Jansen van Rensburg, M. (2018) 'Assessing Gender Responsiveness of the Government-Wide Monitoring and Evaluation System in South Africa', *Development Southern Africa*, 35: 163–178.

UN Women. (2014) *Gender Mainstreaming in Development Programming*. New York: UN Women <http://www.unwomen.org/~/media/headquarters/attachments/sections/how we work/unsystemcoordination/gendermainstreaming-issuesbrief-en pdf.pdf> accessed November 15, 2018.

United Nations Declaration. (2008) 'United Nations Declaration on the Rights of Indigenous Peoples', *United Nations General Assembly*, Resolution 61/295: 10. DOI: 10.1093/iclqaj/24.3.577

United Nations Development Programme (UNDP). (2014a) *UNDP Gender Equality Strategy 2014–2017*. New York, NY.

United Nations Development Programme (UNDP). (2014b) *Guidance note: Gender Statistics*.

United Nations Evaluation Group (UNEG). (2005a) *Norms for Evaluation in the UN System*. New York, NY.

United Nations Evaluation Group (UNEG). (2005b) *Standards for Evaluation in the UN System*. New York, NY.

United Nations Evaluation Group (UNEG). (2011) *Integrating Human Rights and Gender Equality in Evaluation – Towards UNEG Guidance*. New York, NY <http://www.unevaluation.org/documentdownload? doc_id=980&file_id=1294> accessed November 15, 2018.

United Nations Evaluation Group (UNEG). (2018) *UN-SWAP Evaluation Performance Indicator Technical Note*. New York, NY.

Weinhardt, L. S. et al. (2014) 'Methods and Protocol of a Mixed Method Quasi-Experiment to Evaluate the Effects of a Structural Economic and Food Security Intervention on HIV Vulnerability in Rural Malawi: The SAGE4Health Study', *SpringerPlus*, 3/1: 1–13.

World Bank. (2012) *Gender Issues in Monitoring and Evaluation in Agriculture*. Washington, DC.

World Bank. (2016) *Results and Performance of the World Bank Group 2015*. Washington, DC <http://ieg.worldbankgroup.org/sites/default/files/Data/reports/rap15_fullreport.pdf> accessed November 15, 2018.

World Bank. (2018a) 'World Bank Country and Lending Groups'. <https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-90 country-and-lending-groups> accessed November 15, 2018.

World Bank. (2018b) 'Gender Data Portal' <http://datatopics.worldbank.org/gender/> accessed July 23, 2018.

Yarbrough, D. et al. (2011) *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*. Thousand Oaks, CA: Sage.

Zhongdan, C. et al. (2008) 'The 100% Condom Use Program: A Demonstration in Wuhan, China', *Evaluation and Program Planning*, 31/1: 10–21.