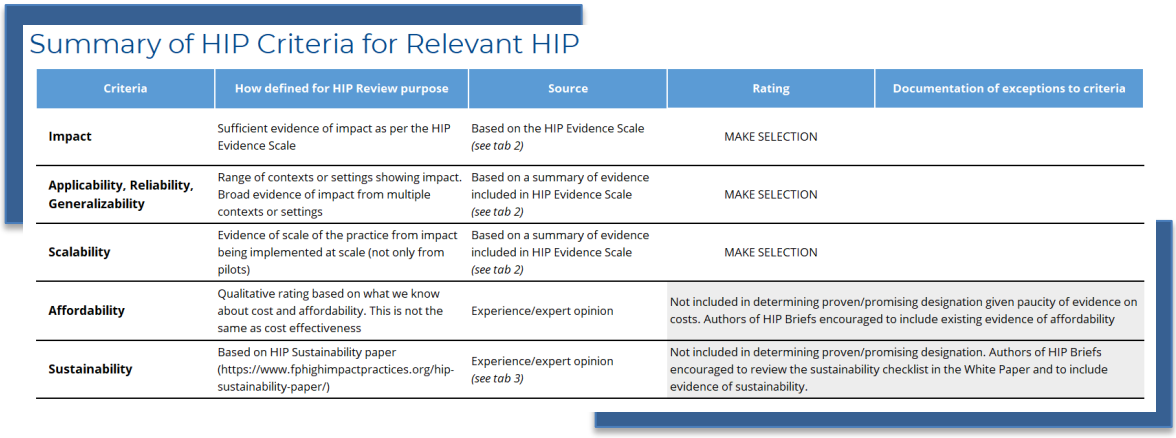
Finding balance with the importance of rigorous research and tacit learning in assessing "What works?": Experience of the High Impact Practice (HIP) Partnership



Summary of HIP Criteria for Relevant HIP

Criteria	How defined for HIP Review purpose	Source	Rating	Documentation of exceptions to criteria
Impact	Sufficient evidence of impact as per the HIP Evidence Scale	Based on the HIP Evidence Scale *(see tab 2)*	MAKE SELECTION	
Applicability, Reliability, Generalizability	Range of contexts or settings showing impact. Broad evidence of impact from multiple contexts or settings	Based on a summary of evidence included in HIP Evidence Scale *(see tab 2)*	MAKE SELECTION	
Scalability	Evidence of scale of the practice from impact being implemented at scale (not only from pilots)	Based on a summary of evidence included in HIP Evidence Scale *(see tab 2)*	MAKE SELECTION	
Affordability	Qualitative rating based on what we know about cost and affordability. This is not the same as cost effectiveness	Experience/expert opinion	Not included in determining proven/promising designation given paucity of evidence on costs. Authors of HIP Briefs encouraged to include existing evidence of affordability	
Sustainability	Based on HIP Sustainability paper (https://www.fphighimpactpractices.org/hip-sustainability-paper/)	Experience/expert opinion *(see tab 3)*	Not included in determining proven/promising designation. Authors of HIP Briefs encouraged to review the sustainability checklist in the White Paper and to include evidence of sustainability.	

Karen Hardee[1]
Michelle Weinberger[2]
Maria Carrasco[3]
Annie Preaux[4]
Saad Abdulmumin[5]
Caroline W. Kabiru[6]
Shawn Malarcher[3]

July 2024

[1]Hardee Associates
[2]Avenir Health
[3]United States Agency for International Development
[4]Tulane University
[5]Bill and Melinda Gates Foundation (formerly)
[6]African Population and Health Research Center

Acknowledgments

The authors thank the members of the Technical Advisory Group of the HIP Partnership for their contributions during development of the HIP Evidence Scale and the HIP Criteria Tool.

Suggested citation: Hardee K, Weinberger M, Carrasco M, Preaux A, Abdulmumin S, Kabiru C, Malarcher S. Finding balance with the importance of rigorous research and tacit learning in assessing "What works?": Experience of the High Impact Practice (HIP) Partnership. Washington, DC: HIPs Partnership; 2024 July.

Contents

Abstract

Since its inception in 2010, the Family Planning High Impact Practice (HIP) Partnership has sought to provide the field with family planning practices that both demonstrate impact and have the potential to be scaled in a range of country contexts and program settings. Determining the appropriateness of evidence and its strength to inform policies and programming is challenging. The partnership's Technical Advisory Group (TAG) needed a standardized way to review and assess the evidence that would center on rigor and value experiential, or tacit, learning. This paper explains the resulting HIP Evidence Scale and calibration of the criteria for determining whether a service delivery or social behavior change HIP is proven or promising. A custom-built, Excel-based HIP Criteria Tool is used to score the assessment of the five criteria on which HIPs are based (impact, applicability/reliability/generalizability to a range of settings, scalability, affordability, and sustainability). The scale and tool can accommodate a range of programmatic interventions and outcomes (centered, but not exclusively, around contraceptive use). The scale, based on the philosophy of using the best available evidence along with practitioner expertise to make decisions on programmatic interventions, is suitable for other health areas.

Key Messages

- Careful work over a decade to find an appropriate evidence framework has ensured that the HIP Evidence Scale and HIP Criteria Tool are tailored for the Family Planning HIP Partnership.
- The HIP Evidence Scale and Excel-based HIP Criteria Tool are built on a philosophy of using the best available evidence, along with practitioner expertise and tacit knowledge, to make decisions on programmatic interventions.
- Use of the scale and the tool facilitates consistent evidence vetting across service delivery and social behavior change HIPs.
- By describing both the scale and the tool, along with the process for vetting evidence and the tips for determining proven vs. promising HIPs, this paper contributes to the transparency of the HIP Partnership.
- The HIP Evidence Scale and HIP Criteria Tool can be adapted for other health areas.

Background

> **High Impact Practices (HIPs)** are a set of evidence-based family planning practices vetted by experts against specific criteria and documented in an easy-to-use format. HIPs help programs focus resources for greatest impact (https://www.fphighimpactpractices.org/).

In 2010, the United States Agency for International Development (USAID)'s family planning and reproductive health program consisted of a large cohort of new family planning technical advisors. Thus, USAID recognized that these advisors would need support to access and use learning from more than 40 years of family planning programming globally to ensure development investments were most effective. The task was daunting. USAID needed to quickly distill mountains of evidence and learning into a manageable, easy-to-understand format through a process that was credible and reduced the potential for bias.

In the process of defining USAID's approach, the team met with colleagues at the United Nations Population Fund (UNFPA) who were facing similar challenges. Upon deciding to join efforts, USAID and UNFPA approached colleagues at the Department and Reproductive Health and Research at the World Health Organization (WHO) in hopes of soliciting their support. Recognizing that WHO has its own well-established process for synthesizing and developing guidelines, UNFPA and USAID were clear that the current gap required a different approach and product than the typical WHO guidelines. Country-based colleagues needed something quickly that was nimble, responsive, and able to incorporate learning that did not lend itself to randomized control trials. Finally, the group identified the need for country-based representation and approached the International Federation of Planned Parenthood to help guide the work. Thus, the partnership was born.

A technical advisory group (TAG) was established. The group, which includes experts from donor agencies and research institutions, country-based stakeholders, and development partners, provides ongoing guidance on practices in family planning that both demonstrate impact and have potential to be scaled in a range of country contexts and program settings. From the beginning the TAG struggled with assessing the evidence. Did the practice have compelling evidence and was "proven" or was the practice "promising" with some evidence pointing to impact but with need for more evidence (Box 1)? Some practices were seen as "common sense" investments and thus lacked documented evidence from formal impact evaluation, such as social marketing. Other practices faced significant scrutiny stemming from cultural concerns, such as post-abortion family planning or professional caution regarding community-based workers providing injectable contraceptives, which required the TAG to consider

Box 1. Definitions of Proven and Promising for HIPs

Proven: Sufficient evidence exists to recommend widespread implementation provided there is careful monitoring of coverage, quality, and cost.

Promising: Good evidence exists that these interventions can lead to impact; more research is needed to fully document implementation experience and impact. These interventions should be implemented widely, provided they are carried out in a research context and evaluated for both impact and process.
https://www.fphighimpactpractices.org/hip-development/

concerns beyond stated outcomes. Thus, the TAG needed a standardized way to review and assess evidence that would value experiential learning.

By 2022, the list of HIPs, comprising the categories of enabling environment, service delivery, social behavior change, and enhancements, had grown to 25.[1] Presentation of the HIPs has evolved over the years. Current HIP Briefs comprise eight pages with standard sections: background including a theory of change, why the practice is important, evidence of impact, tips for implementation, tools and resources, and references.

Purpose

The purpose of this paper is to describe the process of developing the HIP Evidence Scale, which unfolded over a decade, and to explain its use in the HIP Criteria Tool to contribute to establishing whether a service delivery or social behavior change HIP is labeled "proven" or "promising."

Assessing Frameworks for Standards of Evidence

Determining the appropriateness of evidence and the strength of that evidence to inform policies and programming is challenging. In the case of the HIPs, while contraceptive use has been the primary outcome of interest, others re also important, ranging from decreasing unintended pregnancies and reaching diverse and underserved groups to reducing access barriers, and addressing social and cultural barriers (Figure 1). Recognizing the need to develop clearer guidance on assessing the evidence in HIP briefs, the HIP TAG held two consultations in 2013 on standards of evidence; the first at a TAG meeting and the second at a consultation on developing standards for identifying evidence-based practices in reproductive health, held in collaboration with the Department for International Development (DFID)-funded Strengthening Evidence for Programming on Unintended Pregnancy (STEP UP) Programme.[2]

Figure 1. Outcomes of Interest to the Family Planning HIP Partnership

Family Planning Outcomes	Increase contraceptive prevalence rate (CPR), modern contraceptive prevalence rate (mCPR), birth spacing; decrease unwanted pregnancies; delay marriage/sexual debut (for adolescents)	Expand method choice, quality, and coverage	Reach diverse underserved groups	Address social and cultural barriers	Reduce financial barriers

The 2013 TAG meeting recommended reviewing existing standards of evidence frameworks used for classifying proven and promising practices. The consultation with STEP UP reviewed the research designs and methodologies that can be used to generate evidence on the impact of family planning and reproductive health interventions and on their implementation, the mechanisms and structures through which such evidence is reviewed and translated in recommendations, and the implications for organizing and funding evidence generation to maximize its quality and utility.[2] Both meetings made it clear that using a strict hierarchy of evidence assessed through systematic reviews, while useful for assessing clinical interventions, was not necessarily the most appropriate for assessing the impact of programmatic interventions. As noted at the consultation on developing standards for identifying evidence-based practices in reproductive health:

> ***A systematic, transparent, and replicable process, guided by an explicit evidence framework, should be followed when developing practice recommendations from a body of evidence. The evidence framework should incorporate those domains that are of specific interest to particular decision-makers; different evidence frameworks may be appropriate for summarizing evidence to inform different types of decisions.[2(p10)]***

To inform the selection of an evidence framework for the HIPs, an analysis of the types of evidence included in five existing HIP briefs that had been designated as proven at the time (Figure 2) gave the TAG a picture of the range of evidence supporting the HIPs, from systematic reviews to quasi-experimental, nonexperimental, and qualitative studies, among others. The TAG tasked a small working group to further refine the HIP classification criteria for proven versus promising practices that TAG was then using and to clarify the evidence review process.

Figure 2. Types of Evidence in Five Early HIP Briefs, 2013

FP High Impact Practice Briefs: Summary of Types of Evidence, For Selected Briefs

	Enabling Environment				Service Delivery					
	Supply Chain Management		Health Communication		Social Marketing		Postabortion Care		Community Health Workers	
	Proven		Proven		Proven		Proven		Proven	
Type of Study	Why important	Impact	Why important	Impact	Why important	Impact	Why important	Impact	Why important	Impact
Systematic review			**X**	**X X** x		**X** x x			**X**	
Experimental (RCT)										
Quasi-experimental			**X**	**X X X X X X X X**		**X X**		**X X X** x x x	x x	**X X**
Non-experimental/ analytical										
Cohort							**X**			
Longitudinal										
Case-control				**X**						**X**
Large surveys (+ analytic reports)		x	x	**X X** x	x	**X**			**X**	
Other										
LSAT Index		x								
Cause of death audit							**X**			
Pre-post, single group								**X X** x		
Non-experimental/descriptive										
Interviews (in-depth), focus groups					**X**		**X**	**X X**		
Case studies						**X**		x x x x	x	**X** x
Non-experimental evaluation					x x x x x			x x	x	
Other (cause of death statistics)							**X**			
Other review			**X X** x	x x		x	**X X X** x x		x x x	**X** **X** x x x
Expert Meeting/technical panel									x	
Anecdotal										
Modeling (simulation)										
Other										
Wall chart					x					
Country supply chain database	x									
Project Brief/Success Story	x x									

Note: X = one reference. Bold indicates that it is published in the peer-review literature.

Bolded **X** indicates that the reference is published in the peer-review literature.

Some references are mentioned more than once in the HIP Briefs. Each reference is noted at the most twice on this table - once in the why important column and once in the impact column.

Analysis conducted by K Hardee, Annex 3 in June 2013 HIP TAG meeting report, http://www.fphighimpactpractices.org/wp-content/uploads/2017/05/hip_tag_meeting_report_june_2013.pdf

The standards for assessing evidence needed to reflect that the HIPs address a range of interventions and that HIP briefs are not intended to be systematic reviews or equivalent to WHO guidelines, which use the GRADE (Grading of Recommendations, Assessment, Development, and Evaluation) system for rating the quality of evidence for systematic reviews and clinical guidelines.[3-5] The TAG in 2013 reflected that at a maximum of eight pages, HIP briefs "were intended to provide the key audience with a sense of

the evidence base (what we know, what we don't know, and what the gaps are)."[6] And furthermore, that "one of the most valuable parts of the briefs, according to our key audience, is 'Tips,' which is based on experiential knowledge."[6]

The TAG continued to discuss classification of proven and promising at the 2014 TAG meeting. Until then, as explained in the meeting report, "assigning a HIP to these categories is determined through a process in which the authors of the HIP brief consider the strength and consistency of the body of evidence that they reviewed, and then make a recommendation to the TAG, whose members then confirm or revise the recommendation."[7] Seeking to further systematize the process of determining the 'strength and consistency' of the evidence, the TAG discussed the range of domains that can be assessed through evidence frameworks, beyond the quality and strength of evidence, such as the magnitude of benefits vs. harm, consideration of context or generalizability, procedures for implementation, feasibility, costs, sustainability, and other health benefits, among others. The TAG concluded that it needed to define the domains of interest for HIPs (criteria on which HIPs are assessed) and to develop a system for assessing the criteria.

A webinar facilitated by the Implementing Best Practice (IBP) HIP Task Team in 2015 further explored standards for identifying evidence-based practices in reproductive health, along with policymakers' views on evidence-based decision-making (see https://www.youtube.com/watch?v=0vRMxdCDRRA). Panelists noted that while randomized control trials (RCTs) are the gold standard for deciding if an intervention works, there are other non-randomized yet rigorous designs that can determine intervention effectiveness and have the advantage of offering lessons about the real-world context within which the intervention was tested. Furthermore, evidence that policymakers need before they can make evidence-based decisions, in addition to political palatability, include the source and weight of the evidence and whether the intervention is affordable and scalable.

In 2017, a TAG subgroup on standards of evidence discussed potential use of a methodology for classifying evidence—the Modified Gray Scale—that had been used previously by the What Works Association to assess evidence for HIV and AIDs interventions for women and girls[8-9] and, subsequently, for postabortion care,[10] education sector responses to early and unintended pregnancy,[11] and female genital cutting.[12] What has been termed the Gray Scale by the What Works Association[9] is a five-level strengths of evidence scale (Table 1) introduced by Gray,[13(p61)] linked to early work on the Cochrane Collection (https://www.cochranelibrary.com/), and grounded on the philosophy that evidence-based medicine and, by extension, evidence-based public health interventions, should rest on the best available systematic evidence and clinical or practitioner expertise.[14,13,15]

Table 1. Five Strengths of Evidence, AKA the Gray Scale

Type	Strength of Evidence
I	Strong evidence from at least one systematic review of multiple well-designed, randomized controlled trials
II	Strong evidence from at least one properly designed, randomized controlled trial of appropriate size
III	Evidence from well-designed trials/studies without randomization, single group pre-post, cohort, time series or matched case control studies
IV	Evidence from well-designed, nonexperimental studies from more than one center or research group
V	Opinions of respected authorities, based on clinical evidence, descriptive studies, or reports of expert committees

Source: Gray, 1997, p. 61

In presenting the five strengths of evidence scale, Gray explained that "the absence of excellent evidence does not make evidence-based decision-making impossible; in this situation, what is required is the *best evidence available,* not the best evidence possible".[13(p61)] Selection of the Gray Scale and its subsequent modification for the work on HIV and AIDS and the other topics was based on a review of several existing evidence frameworks at the time, including GRADE, SORT (Strength of Recommendation Taxonomy),[16] and Levels of Evidence from the Oxford Center for Evidence-based Medicine (Centre for Evidence Based Medicine, 2009).[17] Early systems focused on evidence-based clinical medicine while newer systems have broadened the focus to evidence-based public health.[15,18] Table 2 illustrates the difference. Clinical practice is more likely to be a single intervention compared to public health practice that tends to include more complex and multiple interventions. The scope of evidence to show effectiveness is wider for public health interventions and the evidence can come from gray literature in addition to published literature.

Table 2. Nature and Role of the Evidence Base in Clinical Practice and Public Health Practice

	Clinical practice	Public health practice and health promotion
Nature of the intervention	Mainly single or simple	Mainly complex or multiple
Nature of evidence to show effectiveness	• Systematic review • RCT	• Systematic review • RCT • Cohort study • Controlled before and after study • Interrupted time series
Sources of evidence	• Published literature	• Published literature • Grey literature
Need for other types of knowledge	Tacit knowledge from clinicians' experience	Tacit knowledge from practitioners and end-users
Contextual factors	Emotional context of the decision	• Sociopolitical context of intervention • Local context

Source: Gray, 2009, p. 322.

Tacit knowledge—from clinicians' experience for clinical practice and from implementers and beneficiaries for public health practice—is also key. Contextual factors are important for both, with emotional context being paramount in clinical practice, and both sociopolitical and local context being important for public health practice. The Modified Gray Scale was appealing since the HIP briefs are not based on systematic reviews. RCTs would not be the most appropriate method for testing impact for some of the interventions. In some cases, program data from a health management information system (HMIS) could be the best available data to assess the intervention. Also, HIPs purposefully include relevant grey literature, e.g., from evaluation reports, rather than relying exclusively on published, peer-reviewed literature.

Prior to its adoption by the TAG, the Gray Scale methodology was reviewed and endorsed for HIV and AIDS programming for women and girls in an expert meeting in 2010, hosted by the Open Society Foundations' Public Health Program, along with criteria for designating practices as 'what works' and 'promising'.[8] Subsequently, the scale was modified to distinguish between studies with control groups and those without to provide additional information for classifying the evidence supporting interventions, based on a recommendation from another expert group meeting in 2011, hosted by the United States Office of the Global AIDS Coordinator. The resulting Modified Gray Scale includes the five levels of evidence outlined by Gray (1997) for assessing bodies of evidence. Level I relates to systematic reviews; level II to RCTs, level IIIa to well-designed studies without randomization that include a control group; level IIIb to well-designed studies without randomization that do not include a control group; level IV to nonexperimental studies; and level V to opinions of respected authorities. The original Gray Scale was designed for assessing bodies of evidence; the Modified Gray Scale can be used to assess the strength of individual studies and, with added criteria, for example, the number of studies and their geographic scope, the Modified Gray Scale rating can assess a body of programmatic evidence.

The TAG subgroup on standards of evidence also discussed having an "evidence dashboard" document associated with each HIP to show the evidence used for each brief. Some examples discussed included the 3ie evidence gap map for adolescents (https://gapmaps.3ieimpact.org/evidence-maps/adolescent-sexual-and-reproductive-health-evidence-gap-map*)* and the Ready, Steady, Go typology of interventions.[19] The dashboard developed for the HIPs as part of the HIP Criteria Tool, is described below.

The TAG agreed that the Modified Gray Scale looked useful for assessing HIPs and informing its deliberations on HIP briefs.[20] The TAG recommended testing its use on upcoming HIP briefs, with tables based on application of the Modified Gray Scale classification to the impact section only, including level of evidence, geographic representation, scale of implementation, and result. The first briefs to include assessment through the Modified Gray Scale were Social Franchising, Mass Media, and Immediate Postpartum Family Planning. Table 3A and 3B were developed for the HIP briefs on Community Health Workers and on Drug Shops and Pharmacies. This experience made it clear that while the Modified Gray Scale was a useful tool, it required further adaptation for the HIP Initiative.

Table 3A. Strength of Evidence in the Impact Section of the 2015 Community Health Workers HIP brief (Proven), presented at June 2017 HIP TAG meeting

HIP brief impact section and Gray Scale level of evidence	# of studies per Gray Scale level	Country(s)
I (systematic review)	1	Multi-country
IIIa (experimental with a control group)	6	Sub-Saharan Africa, Madagascar, Ghana, Bangladesh (2), Ethiopia, India
IIIb (experimental with no control group)	8	Afghanistan, Nigeria (2), India (2), DRC, Guatemala, Philippines
IV (non-experimental)	5	Bangladesh, Indonesia, multi-country (2), Ethiopia (2)
V (expert opinion)	3	Multi-country (3)
Total	**23**	**11 countries, 1 regional, 5 multi-country**

Table 3B. Strength of Evidence in the Impact Section of the 2013 Drug Shops and Pharmacists HIP brief (Promising), presented at June 2017 HIP TAG meeting

HIP brief impact section and Gray Scale level of evidence	# of studies per Gray Scale level	Country(s)
IIIa (experimental with a control group)	1	India
IIIb (experimental with no control group)	5	India; Indonesia; Zambia; UK; USA
IV (non-experimental)	4	Global; Kenya; South Africa; Nigeria
V (expert opinion)	1	Zambia
Total	**11**	**9 countries, 1 global**

The HIP Evidence Scale

As TAG members and other staff and consultants used the Modified Gray Scale to assess evidence for service delivery and social behavior change HIP briefs, additional types of evidence emerged that were not explicitly reflected on that scale, but were appropriate to include, for example, propensity score matching, which is a robust methodology for studies of mass media, a social behavior change HIP. Table 4 shows the HIP Evidence Scale that evolved from the Modified Gray Scale, tailored for the programmatic orientation of the HIPs. Two additional types of evidence were added to level IIIa, namely other rigorous design, and systematic review of quantitative, although non-RCT, studies. Routine, or program data, such as service statistics from HMIS, or other monitoring and evaluation data, were assigned to level IV, and qualitative data, including from qualitative studies or systematic reviews of qualitative studies, were assigned to level V. For the HIPs, levels I, II, and IIIa are designated as studies that include a control group, while levels IIIb, IV and V are those that do not include a control group.

Table 4. HIP Evidence Scale

Level	Type of Study
Evidence with a control group	
I	Systematic review of randomized control trials (RCTs)
II	Randomized control trials
IIIa	Control with pre/post design (non-randomized/quasi-experimental)
	Control with post-only design (non-randomized)
	Other rigorous design (e.g., propensity score matching)
	Systematic review of non-RCTs (quantitative)
Evidence without a control group	
IIIb	Pre/post design, no control
IV	Routine/program data (e.g., service statistics or other monitoring and evaluation data)
V	Qualitative
	Systematic review of non-RCTs (qualitative)

Given that language associated with evidence can be loaded, e.g., strong vs. weak, high quality vs. low quality, the HIP Evidence Scale intentionally uses the distinction of studies with and without a control group to assess strength of evidence—with the recognition that studies need to be well conceived and implemented. In the original Gray Scale, level V was designated as "Opinions of respected authorities, based on clinical evidence, descriptive studies or reports of expert committees" (see Table 1). For HIPs, this evidence is included in the section of the brief on tips for implementation. As noted above, this section, based on experiential evidence, has been considered one of the most valuable. Carrasco et al. confirmed in 35 in-depth interviews with users of the HIP Briefs that they help "address an important need for accessible, practical, and useful information to support the design and implementation of evidence-based policies and programs."[21(p8)]

Building the HIP Criteria Tool

Starting in 2017, the subgroup on standards of evidence began building an Excel-based tool to use for characterizing the evidence in HIP briefs related to the five HIP criteria and assessing the evidence to determine both proven and promising practices. This tool, custom built for assessing service delivery and social behavior change HIP briefs, along with guidance for its use is available at https://www.fphighimpactpractices.org/hip-development/.

Assessing Proven vs. Promising HIPs

Five criteria are used to determine if a practice is proven or promising and how it is assessed by the TAG (Table 5). The first three—impact; applicability, reliability, and generalizability; and scalability—come from a summary of the evidence in the HIP brief, while the other two—affordability and sustainability—

are based on experience and expert opinion. The TAG produced a white paper that includes a checklist for assessing HIPs from the perspective of sustainability.[22]

Table 5. Five Criteria for Assessing High Impact Practices as Proven or Promising

Criteria	How defined for HIP review purpose	Source
Impact	Sufficient evidence of impact as per the HIP Evidence Scale	Based on summary of evidence included in the impact section of the HIP brief
Applicability, Reliability, Generalizability	Range of contexts or settings showing broad evidence of impact from multiple contexts or settings	Based on summary of evidence included in the HIP brief
Scalability	Evidence of scale of the practice from impact being implemented at scale (not only from pilots)	Based on summary of evidence included in the HIP brief
Affordability	Qualitative rating based on what is known about cost and affordability. This is not the same as cost effectiveness.	Experience/expert opinion
Sustainability	Based on HIP sustainability paper https://www.fphighimpactpractices.org/hip-sustainability-paper/	Experience/expert opinion

To further refine the criteria for assessing proven vs. promising, the standards of evidence subgroup undertook an analysis of existing proven and promising service delivery and social behavior change HIPs to ascertain if there were clear differences in the evidence in the proven and promising briefs in terms of the types of studies as indicated by the HIP Evidence Scale, along with distinctions in the other four criteria.

The first step in the analysis was to determine which HIPs had been subjected to some version of the HIP Evidence Scale. This required accessing existing Excel files with analysis for individual HIPs where available, reviewing relevant HIP TAG meeting report sections of TAG review of HIP briefs (this is done prior to publication of briefs), and, in some cases, filling in information to facilitate comparison of the evidence across briefs (going back to the original literature review for the HIP and to original studies, as needed). The analysis included six service delivery HIPs (three proven: Community Health Workers, Immediate Postpartum Family Planning, and Social Marketing; and three promising: Pharmacies and Drug Shops, Family Planning and Immunization Integration, and Social Franchising) and five social behavior change HIPs (four proven: Mass Media; Couples Communication; Social Norms; Knowledge, Attitudes, and Behavior; and Self-Efficacy; and one promising: Digital Health for Social Behavior Change).

Figure 3 shows the analysis of the five criteria for the six service delivery HIPs. The patterns were similar for social behavior change HIPs (not shown). This visual snapshot of existing proven and promising HIPs showed that proven HIPs tended to have more evidence in the 'with a control group' category (green shading in the impact section) than did promising HIPs (light orange shading). The snapshot also shows the number of studies in each level. Furthermore, proven HIPs tended to have more evidence of applicability, reliability, and generalizability, measured through the number of countries and regions with evidence and the populations included (e.g., general population or specific populations). Proven HIPs also showed more evidence of scale beyond pilots or small-scale implementation. For example, in

the case of Pharmacies and Drug Shops, a promising HIP, the impact section includes 12 studies, 10 with no control group and two systematic reviews of non-RCT quantitative studies (that showed positive results, but without tests of significance). Regarding the criteria of applicability, reliability, and generalizability, the evidence was largely from studies of injectables and emergency contraception, with decent geographic spread. Scalability rated highly given the large number of pharmacies and drug shops in many countries. The brief did not present any direct evidence of affordability or sustainability but did note that the interventions could be affordable with a caution about the potential financial burden on clients. For sustainability, the brief noted that the practice could be sustainable if using existing pharmacies and drug shops.

Figure 3. Visual Snapshot of Evidence Related to Five HIP Criteria for Six Service Delivery HIPs

Selected Service Delivery HIPs

			PROVEN						PROMISING					
	HIP Criteria		CHW (see note 1)		Immediate Postpartum FP (see note 2)		Social Marketing		Pharmacies and Drug Shops		FP & Immunization (see note 3)		Social Franchising	
	1. Impact		Proven (2015)		Proven (2022)		Proven (2021)		Promising (2021)		Promising (2021)		Promising (2018)	
	Assessment of impact section evidence													
	Level	Type of study	Positive / Significant	Positive / No significance test	Positive / Significant	Positive / No significance test	Positive / Significant	Possitive / no significance test	Positive / Significant	Outcomes other than contraceptive use	Positive / Significant	Positive / no significance test	Positive / Significant	Not significant
Evidence with control group	I	Systematic Review of RCT		1										
	II	RCT										1		
	IIIa	Control with pre/post (non randomized/quasi-experimental)		6								1	1	2
		Control with post only (not randomized)												1
		Other Rigorous Design (e.g. propensity score matching)												
		Systematic Review of non-RCTs (quantitative)						4		2				
Evidence w/out control group	IIIb	Pre/post no control		8						4				
	IV	Routine/program data		5	1	4						4		
		Other non-rigorous design								6				
	V	Qualitative		3										
		Systematic Review of non-RCTs (qualitative)												
	n/a	Other/unsure												
					The impact section of the IPFP brief is based on evidence from analysis of routine program data, but in the case of this practice, that is appropriate evidence to use. Also, there is other strong data related to PPFP outcomes other than contraceptive use.				Limited evidence on impact; need more definition of actual practice (e.g. task-sharing, over-the-counter provision). Potential for additional secondary analysis of DHS data to look more at causal relationships?		KH reconstructed this scoring on the HIP Evidence Scale from the brief (Excel not found if it was done)			
	2. Applicability, Reliability, Generalizability		11+ countries across regions		Multiple countries across regions		Multiple countries across regions		Evidence largely from injectables and EC; but decent geographic spread		Evidence in the impact section across 6 countries in 3 regions		Evience in the impact section is from 4 countries	
	3. Scalability		Broad evidence of scale		Broad evidence of scale		Broad evidence of scale		Scalability rates highly given large #s of pharmacy and drug shop networks in many countries		Mostly pilots		Wide evidence of social franchising networks: "There are more than 70 known family planning franchise programs worldwide, largely in Africa and South/Southeast Asia."	
	4. Affordability		Evidence of cost-effectiveness		Affordability not covered in the brief		Can be affordable if well designed		No direct evidence; depending on intervention can be affordable but need caution on potential financial burden on clients.		Cost included as a priority research question: Does integration lead to cost savings or other efficiencies in terms of organization of care or deployment of staff resources in various settings?		Cost-effectiveness information is not comparable across countries	
	5. Sustainability		Broad evidence of sustainability		Broad evidence of sustainability		Evidence of sustainability		No direct evidence; depending on intervention can be sustainable if using existing pharmacies and drug shops.		No direct evidence but theoretically sustainable		No direct evidence of sustainability; one research question is: What are the associated costs with maintaining a social franchise?	
	Notes		Note 1 (CHW, 2015). The CHW assessment of evidence was conducted using an early version of the HIP Evidence Scale before significance was included. The brief is undergoing an update and the		Note 2 (PPFP, 2022). The TAG agreed that evidence related to measurement of use of only one contraceptive method related to PPFP should not be included in the evidence of impact, further limiting the evidence in the impact section.						Note 3 (FP and Immunization) KH compiled this information from the brief itself. FP and Immunization provides an example of a practice with some strong evidence of impact and evidence across a range of countries but without evidence of scale,			

Some practices were outliers in the analysis, for example, Immediate Postpartum Family Planning (a proven service delivery HIP) and Social Norms (a social behavior change HIP). In the case of immediate postpartum family planning, the evidence is based on routine data (HMIS), which is in the category of 'without a control group.' In this case, the TAG, which makes the final determination of an HIP being proven or promising, agreed that the evidence for this practice seems limited largely based on the TAG's longstanding decision not to include evidence from studies based on single contraceptive methods. The TAG has since decided that where relevant, evidence from studies of single methods is warranted.[23] For social norms, which is based primarily on qualitative data, the brief itself notes that, "measurement challenges are a factor in the limited evidence available to demonstrate how interventions can successfully address family-planning-related social norms."[24(p5)] In both cases, the TAG exercised its

prerogative to make the final determination of whether the practice is proven or promising, based on extensive discussion of the evidence from studies and from experience in implementing programs, just as Gray and colleagues[14] had recommended. TAG discussions are documented in TAG meeting reports on the HIP website (https://www.fphighimpactpractices.org/hip-technical-advisory-group-meetings/).

Based on this analysis, the TAG approved a recommendation from the subgroup on standards of evidence at its June 2023 meeting for the conditions for each of the HIP criteria corresponding to a proven and promising designation for service delivery and social behavior change practices (Table 6). These tips guide the TAG's discussion and final determination of the designation for each HIP. For the impact criteria, using the HIP Evidence Scale, proven practices should have at least four studies with positive evidence at level I, II, or IIIa on the HIP Evidence Scale (with at least three studies with statistically significant results), with explanation for exceptions, while promising practices should have at least one study at levels I, II, and IIIa, and/or at least four studies at levels IIIb, IV, or V, with explanation for exceptions.

Table 6. Tips for Determining Proven/Promising Designation for HIPs Using the Five HIP Criteria

HIP Criteria	Proven	Promising
Impact	At least four studies with positive evidence at level I, II, or IIIa on the HIP Evidence Scale (with at least three studies with statistically significant results), with explanation for exceptions	At least one study at levels I, II, and IIIa and/or at least four studies at levels IIIb, IV, or V, with explanation for exceptions
Applicability, reliability, generalizability	At least four countries across more than one region	Fewer than four countries or evidence from only one region
Scalability	Broad evidence of implementation at reasonable scale (for the HIP, e.g., at least 50% of studies implemented at a reasonable scale)	Evidence largely from pilots and/or small-scale implementation (greater than 50% of the studies show implementation from pilots and/or small-scale implementation)
Affordability	Not included in determining proven/promising designation given paucity of evidence on costs. Authors of HIP Briefs encouraged to include existing evidence of affordability.	
Sustainability	Not included in determining proven/promising designation. Authors of HIP Briefs encouraged to review the sustainability checklist in the White Paper and to include evidence of sustainability.	

The HIP Criteria Tool includes space to record the explanation from the TAG should an exception be made (e.g., as noted above for Immediate Postpartum Family Planning and Social Norms). The criteria of applicability, reliability, and generalizability uses geographic spread, with proven practices having evidence in at least four countries across more than one region, and promising practices having evidence in fewer than four countries or evidence from only one region. Scalability is distinguished between broad implementation at reasonable scale for proven practices (for the HIP, e.g., at least 50% of studies implemented at a reasonable scale) and evidence largely from pilots and or small-scale implementation for promising (i.e., greater than 50% of the studies show implementation from pilots and/or small-scale implementation). Given the lack of evidence across the HIPs on affordability and sustainability these two criteria, although important, do not have explicit proven or promising designations, although they may factor into the TAG discussions about the practice. In reviewing the

three scores, if ratings are mixed across the criteria, the TAG will need to make a decision anchored on impact. Unless the TAG makes an exception with an explanation for the rationale, to be proven, a practice should show proven impact and proven for at least one of the other two criteria. Figures 4A–C show the Review Evidence Scale tab of the HIP Criteria Tool updated to incorporate the TAG decision on proven vs. promising. Figure 5 shows an illustration of a completed Summary of HIP Criteria tab of the HIP Criteria Tool, the 'evidence dashboard,' including a box to document the discussion of the TAG regarding the practice and the rationale for the TAG's final determination for the HIP. This version of the tool will be used to assess service delivery and social behavior change HIPs developed or revised in the future.

Figure 4. Impact Summary, Replicability and/or Generalizability Summary, and Scalability Summary from Tab 2 in HIP Criteria Tool

4A. Impact Summary Using the HIP Evidence Scale

Fill in details of each study used as evidence in the HIP on the "1_Enter Study Details" tab; the summary tables below will populate automatically.

Impact Summary using the HIP Evidence Scale for *insert name of practice*

Level	Type of study	**# positive** significant results	**# positive** results but no significant test	**# non-significant** results	**# no difference** results*	**# negative** results*	Other (inc. mixed results)	Total # Results*	Total # Unique Studies*
Studies with a Control									
I	Systematic Review of RCT	0	0	0	0	0	0	*0*	*0*
II	RCT	0	0	0	0	0	0	*0*	*0*
IIIa	Control with pre/post (non randomized/quasi-experimental)	0	0	0	0	0	0	*0*	*0*
	Control with post only (not randomized)	0	0	0	0	0	0	*0*	*0*
	Other Rigorous Design (e.g. propensity score matching)	0	0	0	0	0	0	*0*	*0*
	Systematic Review of non-RCTs (quantitative)	0	0	0	0	0	0	*0*	*0*
Studies without a Control									
IIIb	Pre/post no control	0	0	0	0	0	0	*0*	*0*
IV	Routine/program data	0	0	0	0	0	0	*0*	*0*
	Other non-rigorous design	0	0	0	0	0	0	*0*	*0*
V	Qualitative	0	0	0	0	0	0	*0*	*0*
	Systematic Review of non-RCTs (qualitative)	0	0	0	0	0	0	*0*	*0*
n/a	Other/unsure	0	0	0	0	0	0	*0*	*0*
	Total Results*	**0**	**0**	**0**	**0**	**0**	**0**	***0***	***0***
	Distribution of studies by result	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	**#DIV/0!**	***#DIV/0!***		
		#DIV/0!		**#DIV/0!**					

The HIP Evidence Scale was adapted from the Gray Scale (see: Gray, J. 2009. Evidence-Based Health Care and Public Health: How to Make Decisions About Health Services and Public Health. 3rd Edition. Edinburgh, Scotland: Churchill Livingston Elsevier)

**In some cases one study may contribute multiple results; this table shows the total number of outcomes included not unique studies*

†Includes studies where this result was significant, or no significance test was conducted

Select the rating based on the HIP evidence summary and tips for determining proven/promising

Agreed Impact Score	**MAKE SELECTION**

If an exception was made to the proven/promising designation please explain below:

Tips for determining proven/promising designation:

Proven At least 4 studies with positive evidence at level I, II, or IIIa on the HIP Evidence Scale (with at least 3 studies with statistically significant results), with explanation for exceptions

Promising At least one study at levels I, II and IIIa and/or at least 4 studies at levels IIIb, IV or V, with explanation for exceptions

4B. Replicability and/or Generalizability Summary

Replicability and/or Generalizability Summary for *insert name of practice*

The summary below is based on the number of studies, not results (as for the Impact section)

Focus of the evidence: # studies by focus area

	Specific	General
Sub-populations (specific (e.g. sex workers) vs general)	0	0
Contexts (specific (e.g. refugee camps) vs general)	0	0

Geographic coverage of the evidence

	#
# different countries represented in the evidence	1

based on individual results not studies

# studies by region:	
Africa	0
Asia	0
LAC	0
Multiple	0

Select the rating based on the context of evidence base: how broad? Impact across multiple contexts?

Agreed Replicability/Generalizability Rating	**MAKE SELECTION**

Tips for determining proven/promising designation:

Proven	At least 4 countries across more than one region
Promising	Fewer than 4 countries or evidence from only one region ✓

If an exception was made to the proven/promising designation please explain below:

4C. Scalability Summary

Scalability Summary for *insert name of practice*

The summary below is based on the number of studies, not results (as for the Impact section).

Context of evidence base: # studies by scale

	#
Pilot	0
Implemented at small scale (e.g. single clinic)	0
Implemented at reasonable scale	0

Select the rating based on the context of evidence: evidence from beyond a small pilot?

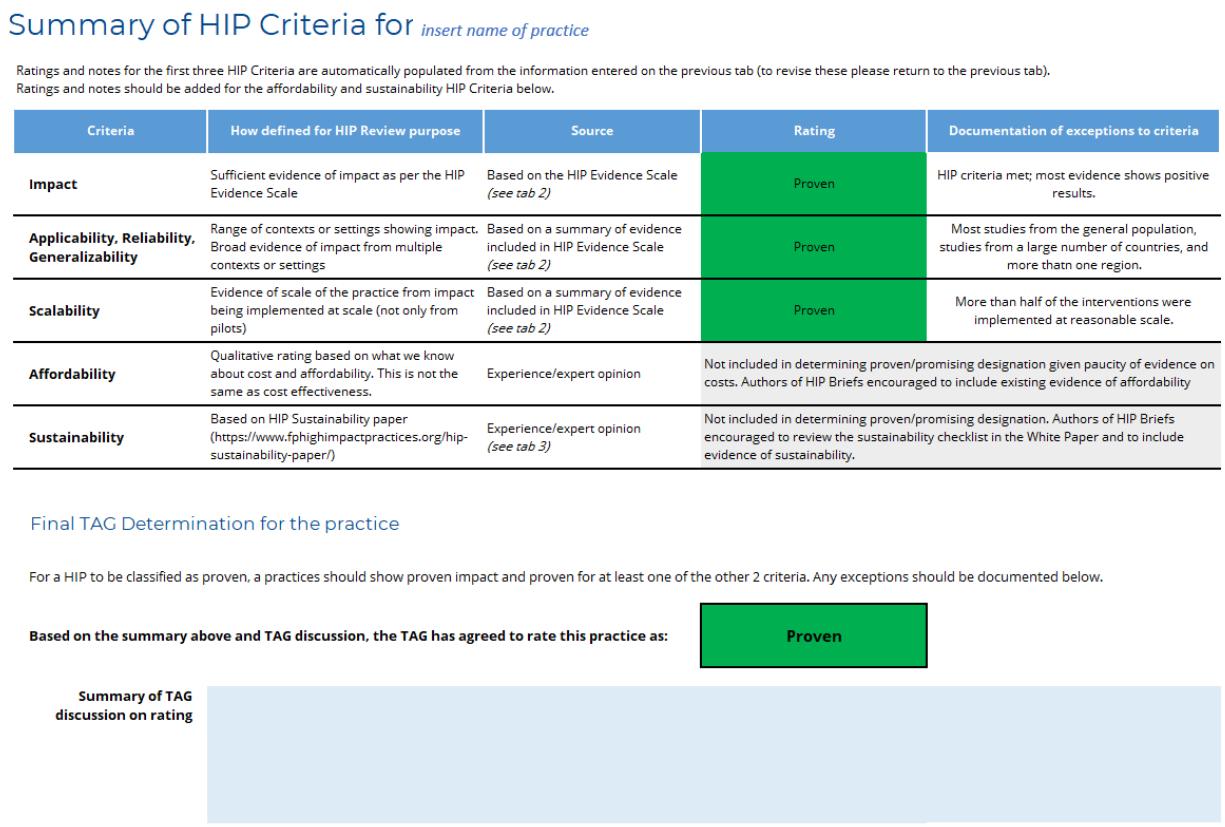
Agreed Scalability Rating	**MAKE SELECTION**

Tips for determining proven/promising designation:

Proven	Broad evidence of implementation at reasonable scale for the HIP (at least 50% of studies implemented at a reasonable scale).
Promising	Evidence largely from pilots and/or small scale implementation (greater than 50% of the studies show implementation from pilots and/or small scale implementation) ✓

If an exception was made to the proven/promising designation please explain below:

Figure 5. Illustrative Example of Summary of HIP Criteria Tab of HIP Criteria Tool



Summary of HIP Criteria for *insert name of practice*

Ratings and notes for the first three HIP Criteria are automatically populated from the information entered on the previous tab (to revise these please return to the previous tab).
Ratings and notes should be added for the affordability and sustainability HIP Criteria below.

Criteria	How defined for HIP Review purpose	Source	Rating	Documentation of exceptions to criteria
Impact	Sufficient evidence of impact as per the HIP Evidence Scale	Based on the HIP Evidence Scale *(see tab 2)*	Proven	HIP criteria met; most evidence shows positive results.
Applicability, Reliability, Generalizability	Range of contexts or settings showing impact. Broad evidence of impact from multiple contexts or settings	Based on a summary of evidence included in HIP Evidence Scale *(see tab 2)*	Proven	Most studies from the general population, studies from a large number of countries, and more thatn one region.
Scalability	Evidence of scale of the practice from impact being implemented at scale (not only from pilots)	Based on a summary of evidence included in HIP Evidence Scale *(see tab 2)*	Proven	More than half of the interventions were implemented at reasonable scale.
Affordability	Qualitative rating based on what we know about cost and affordability. This is not the same as cost effectiveness.	Experience/expert opinion	Not included in determining proven/promising designation given paucity of evidence on costs. Authors of HIP Briefs encouraged to include existing evidence of affordability	
Sustainability	Based on HIP Sustainability paper (https://www.fphighimpactpractices.org/hip-sustainability-paper/)	Experience/expert opinion *(see tab 3)*	Not included in determining proven/promising designation. Authors of HIP Briefs encouraged to review the sustainability checklist in the White Paper and to include evidence of sustainability.	

Final TAG Determination for the practice

For a HIP to be classified as proven, a practices should show proven impact and proven for at least one of the other 2 criteria. Any exceptions should be documented below.

Based on the summary above and TAG discussion, the TAG has agreed to rate this practice as: **Proven**

Summary of TAG discussion on rating

Discussion

Since its inception in 2010, the purpose of the HIPs has been to provide the field with practices in family planning that both demonstrate impact and have the potential to be scaled in a range of country contexts and program settings. The need for a rigorous and transparent process of assessing the evidence for practices deemed high impact was clear from the beginning of the HIP Partnership. Furthermore, the HIP Partnership needed clear criteria to guide the HIP TAG in their determination of whether a service delivery or social behavior change HIP is proven or promising. Based on assessment of the range of evidence in early briefs and existing evidence frameworks, a HIP TAG working group on standards of evidence developed the HIP Evidence Scale for inclusion in a custom-built HIP Criteria Tool to assess the five criteria on which HIPs are based (impact; applicability, reliability, and generalizability to a range of settings; scalability; affordability; and sustainability).

The HIP Evidence Scale and HIP Criteria Tool can accommodate a range of programmatic interventions as well as outcomes (centered, but not exclusively, around contraceptive use), as well as a range of data sources. The HIP Evidence Scale and HIP Criteria Tool were formulated based on the philosophy espoused by Sackett, Gray, and colleagues that evidence-based public health interventions should be based on the best available systematic evidence together with practitioner expertise.[14] The HIP Evidence Scale to assess impact has evolved from Gray's five strengths of evidence, adhering to Gray's view that "what is required is the best evidence available, not the best evidence possible."[13(p61)] Examples of use of the Modified Gray Scale for reviews of evidence of HIV programming for women and girls, postabortion

care, female genital cutting, and education sector responses to early and unintended pregnancy gave the HIP TAG further confidence in starting with it in developing the HIP Evidence Scale. Analysis of evidence in more recent proven and promising service delivery and social behavior change HIPs showed general differences in the evidence between proven and promising, yielding tips for the HIP TAG to use in making its final determination for each HIP.

Conclusion

Careful work over a decade to find an appropriate evidence framework has ensured that the HIP Evidence Scale and HIP Criteria Tool are tailored for the HIP Partnership. The HIP Evidence Scale and Excel-based HIP Criteria Tool, built on a philosophy of using the best available evidence along with practitioner expertise to make decisions on programmatic interventions, can be adapted for other health areas. By describing both the scale and the tool, along with the process for vetting evidence and the tips for determining proven vs. promising HIPs, this paper contributes to the transparency of the HIP Partnership.

References

1. High Impact Practices in Family Planning (HIPs). *Family Planning High Impact Practices List*. Washington, DC: The High Impact Practices Partnership; August 2022a. https://www.fphighimpactpractices.org/briefs/family-planning-high-impact-practices-list/.
2. STEP UP Research Programme Consortium (STEP UP). *Second Consultation on Developing Standards for Identifying Evidence-Based Practices in Reproductive Health, STEP UP Report*. New York: Population Council; 2014. https://knowledgecommons.popcouncil.org/cgi/viewcontent.cgi?article=1269&context=departments_sbsr-rh.
3. Guyatt G, Oxman A, Vist G, et al. for the GRADE Working Group. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008;336 (7650):924-926.
4. Lewin S, Bosch-Capblanch X, Oliver S, et al. Guidance for evidence-informed policies about health systems: assessing how much confidence to place in the research evidence. 2012. *PLoS Med*. 9(3):e1001187.
5. Boon MH, Thomson H, Shaw B, et al. 2021. Challenges in applying the GRADE approach in public health guidelines and systematic reviews: a concept article from the GRADE Public Health Group. *J Clin Epidemiol*. 2021;135:42-53. doi:10.1016/j.jclinepi.2021.01.001.
6. HIPs. HIP Technical Advisory Group (TAG) Meeting Report. June 7–8, 2013. Accessed May 10, 2023. http://www.fphighimpactpractices.org/wp-content/uploads/2017/05/hip_tag_meeting_report_june_2013.pdf.
7. HIPs. HIP Technical Advisory Group (TAG) Meeting Report. June 4–5, 2014. Accessed May 10, 2023. https://www.fphighimpactpractices.org/wp-content/uploads/2017/05/hip_tag_report_2014_jul_24.pdf.
8. Gay J, Croce-Galis M, Hardee K. *What Works for Women and Girls: Evidence for HIV/AIDS Interventions.* Washington, DC: Population Council, The Evidence Project, and What Works Association; 2016. http://www.whatworksforwomen.org/pages/methodology.
9. What Works Association. *Strength of Evidence Methodology*. Arlington, VA: WWA; n.d. http://www.whatworksassociation.org/strength-of-evidence-methodology.html.
10. Huber D, Curtis C, Irani L, Pappa S, Arrington L. Postabortion care: 20 years of strong evidence on emergency treatment, family planning, and other programming components. *Glob Health Sci Pract*. 2016;4(3):481-94. doi:10.9745/GHSP-D-16-00052.
11. United Nations Educational, Scientific, and Cultural Organization (UNESCO). *Early and Unintended Pregnancy & the Education Sector: Evidence Review and Recommendations*. Paris: UNESCO; 2017. https://unesdoc.unesco.org/ark:/48223/pf0000251509.
12. Matanda DJ, Van Eekert N, Croce-Galis M, Gay J, Middelburg MJ, Hardee K. What interventions are effective to prevent or respond to female genital mutilation? A review of existing evidence from 2008–2020. *PLoS Glob Public Health*. 2023;3(5):e0001855. doi:10.1371/ journal.pgph.0001855.
13. Gray JA. *Evidence-Based Healthcare: How to Make Health Policy and Management Decisions*. London, UK: Churchill Livingstone; 1997.

14. Sackett DL, Rosenberg WM, Muir Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: What it is and what it isn't. *BMJ*. 1996;312:71-72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2349778/.
15. Gray JA. *Evidence-Based Health Care and Public Health: How to Make Decisions about Health Services and Public Health*. 3rd ed. Edinburgh, Scotland: Churchill Livingston Elsevier; 2009.
16. Ebell M, Siwek J, Weiss B, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician*. 2004;69(3):548-556.
17. Centre for Evidence-Based Medicine. Oxford Centre for Evidence-Based Medicine: Levels of Evidence. Oxford, UK: Oxford University. www.cebm.net. March 2009. Accessed May 10, 2023.
18. Yamey G, Feachem R. Evidence-based policymaking in global health – the payoffs and pitfalls. *Evid Based Med*. 2011 Aug 16(4):97-99.
19. UNAIDS Interagency Task Team on HIV and Young People/Ross D, Dick B, Ferguson J, eds. *Preventing HIV/AIDS in Young People: A Systematic Review of the Evidence from Developing Countries*. World Health Organization Technical Report Series 938. Geneva: WHO; 2006. https://apps.who.int/iris/bitstream/handle/10665/43453/WHO_TRS_938_eng.pdf.
20. HIPs. HIP TAG Meeting Report. November 29–30, 2017. Accessed May 10, 2023. https://www.fphighimpactpractices.org/wp-content/uploads/2021/08/HIP-TAG-Report-Nov-2017.pdf.
21. Carrasco M, Ohkubo S, Preaux A, et al. Assessing use, usefulness, and application of the High Impact Practices in Family Planning briefs and strategic planning guides. *Glob Health Sci Pract*. 2023;11(4):e2200146. doi:10.9745/GHSP-D-22-00146.
22. Hardee K, Sulzbach S, Chatterji M, Reier S, Malarcher S. Guidance on assessing the potential sustainability of practices as part of an evidence review: considerations for High Impact Practices in Family Planning. Washington, DC: USAID; 2017. https://www.fphighimpactpractices.org/hip-sustainability-paper/.
23. HIPs. 2023. HIP TAG Meeting Report. June 12–14, 2023. Accessed May 10, 2023. https://www.fphighimpactpractices.org/wp-content/uploads/2023/09/June2023-HIPS-TAG-Meeting-Report.pdf.
24. HIPs. Social norms: promoting community support for family planning. Washington, DC: USAID; 2022b. https://www.fphighimpactpractices.org/briefs/social-norms/.