A PRACTICAL GUIDE TO MEASURING WOMEN'S AND GIRLS' EMPOWERMENT IN IMPACT EVALUATIONS

Rachel Glennerster*, Claire Walsh, Lucia Diaz-Martin

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

*The views expressed here are those of the authors and do not necessarily reflect those of the funders or the Department for International Development.

OVERVIEW

Impact evaluations can produce useful insights on how to design programs and policies that can increase women's and girls' empowerment and help us better understand the process of empowerment itself. Yet, it can be challenging to design a measurement strategy and to identify indicators that capture changes in empowerment, that are tailored to the local context, and that minimize reporting bias. Pulling insights from diverse disciplines and the experience of J-PAL affiliated researchers around the world, this guide offers practical tips for overcoming these challenges in impact evaluations. We emphasize the importance of conducting in-depth formative research to understand gender dynamics in the specific context before starting an evaluation, developing locally tailored indicators to complement internationally standardized ones, and reducing the potential for reporting bias in our instruments and data collection plan. We do not provide a single set of ready-to-go survey instruments; instead, we outline a process for developing indicators appropriate to each study along with extensive examples. In this way, we hope this guide can help provide researchers and practitioners the tools to select or develop their own indicators of empowerment that are right for their impact evaluations.

ACKNOWLEDGEMENTS

We thank Christopher Boyer, Mayra Buvinic, Jan Cooper, Aletheia Donald, Sarah Gammage, Markus Goldstein, Seema Jayachandran, Hazel Malapit, Danielle Moore, Agnes Quisumbing, and Innovations for Poverty Action (IPA) who provided invaluable comments. All errors are our own.

LOCATION: INDIA. PHOTO: LIBBY ABBOT | J-PAL/IPA
COVER LOCATION: RWANDA. PHOTO: TOM GILKS | J-PAL/IPA

TABLE OF CONTENTS

An introduction to this guide 1

When should we consider gender in measurement? 3

What is empowerment and when should we consider measuring it in an impact evaluation? 4

What are some challenges in measuring women's and girls' empowerment? 7

How can we build a reliable strategy for measuring empowerment in an impact evaluation? 11

- Step 1. Formative research 12
- Step 2. Theory of change, outcomes, and indicators 18
- Step 3. Data collection instruments 30
- Step 4. Data collection plan 41

Conclusion 45

AN INTRODUCTION TO THIS GUIDE

A growing number of policymakers are investing in women's and girls' empowerment—"the process by which those who have been denied the ability to make strategic life choices acquire such an ability."[1] Policymakers have pursued empowerment as a policy goal in its own right and as a means to unlock greater development potential in low-and middle-income countries, and the number of programs and policies aimed at increasing empowerment has rapidly increased. As a result, many practitioners and researchers are grappling with how to more precisely measure something as complex as women's and girls' empowerment in impact evaluations.

Listening to women to learn about their aspirations and the barriers they face is a critical starting point. Good impact evaluations on empowerment start with rich discussions with women, girls, and other community members about what it means to be empowered in that specific context. As part of the formative research that my research team and I (Rachel) conducted for a randomized evaluation on girls' empowerment and child marriage in Bangladesh, we interviewed young women about their daily lives, aspirations, and what they wanted to do that they were currently prohibited from doing.[2] We also interviewed local NGOs and community partners, asking: "What does an empowered girl do differently than a girl who is not empowered?" Some of the answers we heard again and again were, "she can go where she likes," "she can negotiate with her parents," "she is not the last to eat at home," and "she thinks differently about what girls can do."

Distilling these rich conversations into a set of indicators that could be measured in a survey questionnaire was not simple or straightforward. Many people we spoke to said that empowered girls had the ability to travel in their communities. Yet, when we talked to girls, we quickly learned that a general question like "how far away from your home can you travel alone?" would not accurately capture how their mobility was constrained. Many girls described being unable to travel outside the home by themselves, yet they also mentioned regularly walking to school alone or with friends. When we asked them about this discrepancy, the girls responded—"but that's different, that's school!"

Through these interviews, we soon understood that a girl's ability to travel alone depended on what she was doing and for whom. She could travel to and from school alone every day, but she could not attend local fairs on her own. To measure mobility as an indicator of empowerment, we cared most about a young woman's ability to go somewhere for an activity that only had value to her. In our survey, then, we asked whether young women could travel alone to a list of common locations and activities, including some that were for no one's benefit but the girl herself. Yet, by using a question and indicator that was so locally tailored, we lost the ability to compare our findings with data on women's mobility in other contexts.

Researchers face many such trade offs and challenges in measuring empowerment. Which outcomes should we prioritize from a longer list of plausible outcomes? How should we measure women's and girls' ability to make choices when we rarely observe the decision-making process? Should we ask people about past decision-making, try to observe a real choice, and/or assume some outcomes are good proxies for people's ability to choose? There is already a wealth of scholarship by economists, feminist scholars, psychologists, anthropologists, sociologists, and practitioners addressing these conceptual challenges. In this guide, we will highlight practical insights from these disciplines that economists and others can use when measuring empowerment in quantitative impact evaluations.

---

[1] Kabeer, Naila. 1999. "Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment." *Development and Change* 30(3): 435-464. Abstract. https://doi.org/10.1111/1467-7660.00125.

[2] Buchmann, Nina, Erica Field, Rachel Glennerster, Shahana Nazneen, Svetlana Pimkina, and Iman Sen. "Power vs Money: Alternative Approaches to Reducing Child Marriage in Bangladesh, a Randomized Control Trial." Working Paper, May 2017. The study was funded by the International Initiative for Impact Evaluation, the US National Institutes of Health, the Nike Foundation, and the International Development Research Center.

We will also dive deeper into the many practical challenges of measuring women's and girls' empowerment and how to overcome them. We explore how to turn indicators into good survey questions; how to test whether survey respondents can understand our questions and answer them accurately; how to decide whom to interview and when to get the best information; how to measure empowerment in ways that are less subject to reporting bias; and how and when non-survey instruments can generate more reliable outcome measures than surveys. We also indicate how different instruments or approaches are likely to affect research costs.

WHO IS THIS GUIDE FOR?

This guide is designed to support the work of monitoring and evaluation practitioners, researchers, and students who are interested in learning how to measure women's and girls' empowerment in an impact evaluation. We primarily discuss strategies for developing good quantitative data collection instruments, including surveys and non-survey instruments, so this guide may be most relevant for practitioners and researchers interested in using these tools. Most of the examples are from randomized evaluations by J-PAL affiliated researchers and their co-authors that took place in low- or middle-income countries, but the underlying intuition may be relevant in high-income countries too. These insights may also be useful for program staff designing interventions focused on increasing women's and girls' empowerment, particularly when it comes to integrating monitoring and evaluation processes into program design.

WHAT DOES THIS GUIDE COVER?

We will share insights on the following key questions and topics:

1. When should we consider gender in measurement?
2. What is empowerment and when should we consider measuring it in an impact evaluation?
3. What are some challenges in measuring women's and girls' empowerment and how can we overcome them?
4. How can we build a reliable strategy for measuring empowerment in an impact evaluation?

   Step 1. Formative research: conduct formative research to understand gender and empowerment in the specific context

   Step 2. Theory of change, outcomes, and indicators: map a theory of change to select appropriate outcomes and indicators

   Step 3. Data collection instruments: develop and validate data collection instruments that minimize reporting bias

   Step 4. Data collection plan: design a data collection plan that minimizes measurement error

We also include two in-depth appendices. Appendix 1 provides examples of survey questions related to women's and girls' empowerment from J-PAL affiliated researchers' randomized evaluations. Appendix 2 features examples of a range of different types of non-survey instruments that can be used in quantitative analysis and tips for deciding when and how to use them.

LOCATION: INDONESIA. PHOTO: JURIST TAN | J-PAL

1. WHEN SHOULD WE CONSIDER GENDER IN MEASUREMENT?

We should always consider gender in measurement in an impact evaluation, even when a program is not targeting one gender. For instance, who we choose to interview matters. A survey of heads of households, politicians, or business owners may miss women's views if they are underrepresented in these positions. Even within a household, different members may be more or less informed about various aspects of family and economic life. It is important, then, to identify who has the information we want to collect before we begin our survey. We also need to consider how the enumerator's gender, ethnicity, and class may influence respondents' answers, as these factors could affect who consents to participate, and what individuals say.

LOCATION: UGANDA. PHOTO: ALINA XU | J-PAL/IPA

Moreover, it is always important to consider gender in our analysis. Analyzing the overall impact of a program may mask important gender dynamics that disaggregating our analysis by gender could uncover. For example, a program that provides fertilizer to farmers may also increase weeding responsibilities, a chore often done by women. A program could both improve educational attainment for children overall and decrease the gender gap in school participation. It is critical to consider gender in the measurement and analysis of any impact evaluation to uncover whether and how the program affects people differently by gender. When planning our impact evaluation, we should consider getting a large enough sample size to be able to detect any important gender differences. We should calculate and report impact estimates by gender in our analyses, along with whether the differences are statistically significant. We should also include a discussion of the main reasons why the program did or did not have differential effects by gender.

2. WHAT IS EMPOWERMENT AND WHEN SHOULD WE CONSIDER MEASURING IT IN AN IMPACT EVALUATION?

WHAT IS EMPOWERMENT?

There are many different definitions of empowerment, but most of the seminal definitions emphasize agency and gaining the ability to make meaningful choices.[3] Many definitions draw on Amartya Sen's concept of an agent as "someone who acts and brings about change, and whose achievements can be judged in terms of her own values and objectives."[4] In this guide, we primarily use Naila Kabeer's definition of empowerment as "the process by which those who have been denied the ability to make strategic life choices acquire such an ability."[5] Kabeer's seminal "resources, agency, and achievements" framework also provides a practical intuition for measuring empowerment, which involves three interrelated dimensions (see Figure 1):[6]

Resources: gaining access to material, human, and social resources that enhance people's ability to exercise choice, including knowledge, attitudes, and preferences

Agency: increasing participation, voice, negotiation, and influence in decision-making about strategic life choices

Achievements: the meaningful improvements in well-being and life outcomes that result from increasing agency, including health, education, earning opportunities, rights, and political participation, among others[7]

Measuring indicators related to resources, agency, and achievements over the course of an evaluation can be an intuitive and practical way to measure the process of empowerment.

Like all power relations, the process of empowerment is also shaped by and interacts with the norms and institutions (cultural, social, political, and economic) that define an individual's possibilities in a given context. These institutionalized "structures of constraint" shape the choices available to women and girls at every step of the empowerment process.[8]

For example, norms about women's mobility in a place like Bangladesh can shape women's resources like social capital. These norms can also affect women's agency in terms of what decisions they are actually able to make, to what extent they need to ask for permission, and what kinds of decisions are actually empowering. For example, making decisions about household purchases may be empowering in some contexts where women are denied this kind of choice. However, it may be disempowering to other women who feel that taking charge of these decisions is an additional burden rather than a choice. Norms can also affect women's achievements. For instance, social norms about women and work shape whether women who want to start businesses or work outside the home are able to do so. In our evaluations, we should consider how these structures of constraint might limit the success of empowerment interventions by impeding people's ability to translate resources into agency, and agency into achievements. They may also influence the extent to which gains for women in the household or private sphere translate to women's collective improvements in the public sphere and vice versa.[9]

---

[3] Malhotra, Anju, Sidney R. Schuler, and Carol Boender. 2002. "Measuring Women's Empowerment as a Variable in International Development." Background paper prepared for the World Bank Workshop on Poverty and Gender: New Perspectives, June 28, 2002, 6. http://siteresources.worldbank.org/INTGENDER/Resources/MalhotraSchulerBoender.pdf.

[4] Sen, Amartya. 1999. *Development as Freedom*. New York: Alfred A. Knopf, 19.

[5] Kabeer 1999, Abstract.

[6] Kabeer 1999, 437-438.

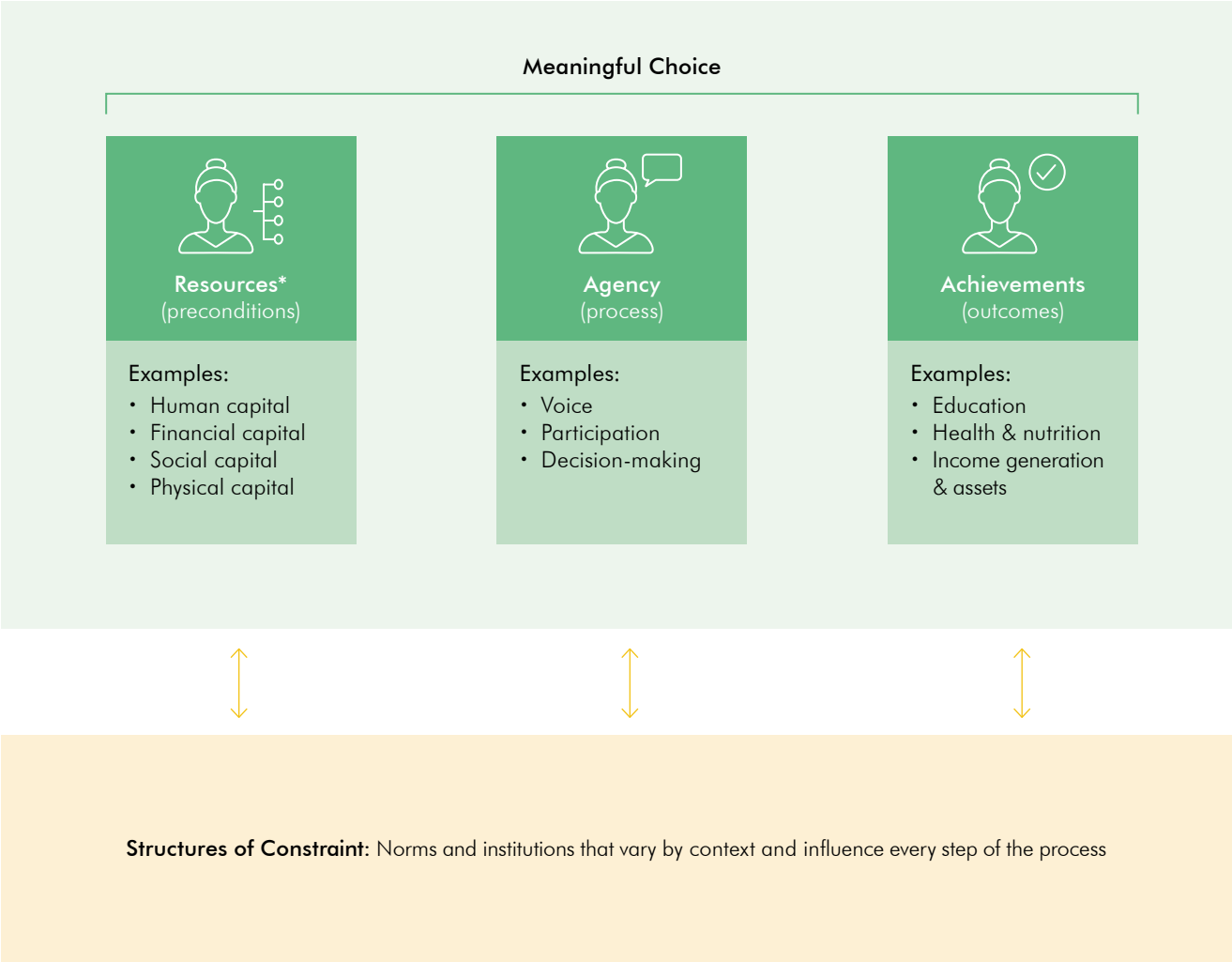
[7] Many outcomes, such as health and educational attainment, can be considered both resources and achievements.

[8] Kabeer, Naila. "Paid Work, Women's Empowerment and Gender Justice: Critical Pathways of Social Change." Working Paper, Institute of Development Studies, January 2008, 24. https://www.ids.ac.uk/publication/paid-work-women-s-empowerment-and-gender-justice-critical-pathways-of-social-change.

[9] Kabeer 2008.

2. WHAT IS EMPOWERMENT AND WHEN SHOULD WE CONSIDER MEASURING IT IN AN IMPACT EVALUATION?

FIGURE 1. VISUAL REPRESENTATION OF KABEER'S *RESOURCES, AGENCY, AND ACHIEVEMENTS* FRAMEWORK OF EMPOWERMENT



*Resources may also be achievements.

Citation: Kabeer, Naila. 1999. "Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment." *Development and Change*, 30 (3): 435-464, Abstract. https://doi.org/10.1111/1467-7660.00125.

Kabeer, Naila. "Paid Work, Women's Empowerment and Gender Justice: Critical Pathways of Social Change." Working Paper, Institute of Development Studies, January 2008, 24. https://www.ids.ac.uk/publication/paid-work-women-s-empowerment-and-gender-justice-critical-pathways-of-social-change.

2. WHAT IS EMPOWERMENT AND WHEN SHOULD WE CONSIDER MEASURING IT IN AN IMPACT EVALUATION?

WHEN SHOULD WE CONSIDER MEASURING EMPOWERMENT IN AN IMPACT EVALUATION?

We should measure empowerment in an impact evaluation when it is the primary or secondary objective of the program. Many programs specifically target women's and girls' empowerment as the main outcome, like negotiation skills programs for young women or trainings on women's rights. Besides these clear cases, there are many programs for which empowerment is a hoped for but secondary outcome, which is also key to measure. For instance, many financial inclusion programs that increase access to credit, savings, and payments services aim to improve the performance of small enterprises, but also aim to increase women's economic empowerment. It may also be important to consider measuring empowerment when there is a risk that a program may disempower or have negative consequences for women. We may want to measure, for example, whether a program unintentionally adds to women's or girls' domestic duties, care work, or chores.[10]

We may also expect program impacts to vary among women according to how empowered they are. For example, access to free contraception may only increase contraceoption use when women have enough agency to negotiate the ability to use it with their partners. In these cases, it can be useful to measure an indicator of agency or decision-making power before implementing the program in order to capture how impacts vary according to initial differences among women. Since empowerment programs span many different domains, from finance to education, health, and political participation, the barriers and opportunities women face in becoming more empowered in these various aspects will necessarily be different. It is therefore important that our outcomes, indicators, and data collection tools are tailored to the specific domain(s) under study.[11]

While it is important to use indicators specific to the interventions and domains being studied, empowerment measures tend to include a core set of concepts. Topics often include women's access to and control over resources including income and assets; participation in important decisions at the personal, household, and community level; control over reproductive health and fertility choices; subjective well-being and happiness; mobility; time use and sharing domestic work; freedom from violence; community and political participation; and well-being outcomes in domains like education, health, and labor.[12]

---

[10] We may find it useful in many cases to consider how social norms related to gender roles in unpaid domestic work interact with the program being evaluated and how it affects women's time use. The World Bank publication below on time use describes two measurement approaches: stylized questions and time diaries. There are many other resources available on conceptualizing and measuring time use, time poverty, and unpaid domestic work, including the Levy Institute's resources and Marilyn Waring's seminal book *If Women Counted* below.

Seymour, Greg, Hazel Jean Malapit and Agnes R. Quisumbing. "Measuring Time Use in Development Settings (English)." World Bank Policy Research Working Paper No. WPS 8147, July 2017. http://documents.worldbank.org/curated/en/443201500384614625/Measuring-time-use-in-development-settings;

Levy Economics Institute of Bard College. n.d. "Publications on Time Poverty." Accessed March 19, 2018. http://www.levyinstitute.org/topics/time-poverty;

Waring, Marilyn. 1998. *If Women Counted: A New Feminist Economics*. San Francisco: Harper & Row.

[11] Many researchers emphasize the importance of using domain-specific measures of empowerment. Some useful examples include International Center for Research on Women's (ICRW) modules for measuring economic empowerment (Golla, Anne Marie, Anju Malhotra, Priya Nanda, Rekha Mehra, Aslihan Kes, Krista Jacobs, and Sophie Namy. 2011. "Understanding and Measuring Women's Economic Empowerment." International Center for Research on Women. https://www.icrw.org/publications/understanding-and-measuring-womens-economic-empowerment/) or International Food Policy Research Institute's (IFPRI) index for measuring women's empowerment in agriculture (Alkire, Sabina, Ruth Meinzen-Dick, Amber Peterman, Agnes Quisumbing, Greg Seymour, and Ana Vaz. 2013. "The Women's Empowerment in Agriculture Index." *World Development* 52: 71-91. https://doi.org/10.1016/j.worlddev.2013.06.007).

For more on the importance of domain-specific indicators see: Alkire, Sabina. 2005. "Subjective Quantitative Studies of Human Agency." *Social Indicators Research* 74 (1): 217–260. 236. https://doi.org/10.1007/s11205-005-6525-0.

[12] Categories synthesized based on a review of the following publications: Alkire et al. 2013; Golla et al. 2011;

Donald, Aletheia, Gayatri Koolwal, Jeannie Annan, Kathryn Falb, and Markus Goldstein. "Measuring Women's Agency." World Bank Policy Research Working Paper no. 8148, July 2017, 34-35. http://documents.worldbank.org/curated/en/333481500385677886/Measuring-womens-agency;

Laszlo, Sonia, and Kate Grantham. "Measurement of Women's Economic Empowerment in GrOW Projects: Inventory and User Guide." McGill University GrOW Working Paper, December 2017;

Lombardini, Simone, Kimberly Bowman, and Rosa Garwood. 2017. "A 'How-To' Guide To Measuring Women's Empowerment: Sharing Experience from Oxfam's Impact Evaluations." Oxfam. https://policy-practice.oxfam.org.uk/publications/a-how-to-guide-to-measuring-womens-empowerment-sharing experience-from-oxfams-i-620271.

3. WHAT ARE SOME CHALLENGES IN MEASURING WOMEN'S AND GIRLS' EMPOWERMENT?

1. MEASURING PEOPLE'S ABILITY TO MAKE IMPORTANT LIFE CHOICES IS CHALLENGING BECAUSE WE RARELY OBSERVE DECISION-MAKING DIRECTLY.

Empowerment is not just about changes in well-being; it is also about people's agency in achieving these changes. Some of the most common options for measuring agency and decision-making power, including asking people about how past decisions were made in a survey, are prone to reporting bias if not carefully designed and implemented. Moreover, since we can often only observe the outcomes of choices and not the real decision-making process itself, it is hard to know whether changes in well-being are indeed the result of women's increased ability to make choices or not. For example, in Rachel and co-authors' evaluation in rural Bangladesh, we found that more married women took on income-generating activities as a result of an empowerment program, which initially appeared to be an empowering labor market choice. Yet, in qualitative interviews with a subset of these women, we found that some were working out of severe economic necessity and many had limited autonomy in choosing their income-generating activity. Women were also often limited to working within the home.[13]

2. EMPOWERMENT IS A PROCESS.

Unlike many other outcomes, empowerment is a continuous process so it requires more effort and creativity on our part to measure it well. Many theories of empowerment emphasize the importance of conceptualizing it as a process, including Kabeer's "resources, agency, achievements" framework. She argues that it is important to measure the resources that could enhance women's ability to make choices, women's agency (voice, participation, etc.) in making those choices, and the final changes in well-being that could result from increased agency.[14] To measure this multistep process, we need to select and measure short-term, intermediate, and final outcomes that can credibly make this causal link. For instance, say we want to measure whether a negotiation program empowers girls to complete secondary school. We will need to measure their resources (such as their negotiation ability), the decision-making process between parents and girls that determines whether they stay in school, and their ultimate educational achievement.

3. MANY ASPECTS OF EMPOWERMENT ARE SUSCEPTIBLE TO REPORTING BIAS.

Asking people about sensitive topics like gender attitudes, aspirations, reproductive health, contraception use, marriage, violence, and decision-making can lead to reporting bias if our survey instruments are not well designed. Social desirability bias is one type of reporting bias that is particularly challenging to mitigate when measuring empowerment. It occurs when respondents give answers that they think the surveyor wants to hear or that are in line with generally accepted social norms rather than reality. Respondents' stated preferences may also be different from their revealed preferences—meaning that people can say one thing but do another. For example, parents may say that they believe girls should be given the same education opportunities as boys in a survey, but they may not actually enroll all of their daughters in school. It takes a good deal of time and effort to design data collection instruments that mitigate social desirability bias and other types of reporting bias.

---

[13] Field, Erica, Rachel Glennerster, and Shahana Nazneen. 2018. "Economic Empowerment of Young Women in Bangladesh: Barriers and Strategies." *Women at Work: Addressing the Gaps*. International Policy Centre for Inclusive Growth Policy in Focus 15 (1): 31-32. http://www.ipc-undp.org/publication/28507.

[14] Kabeer 1999, 437-438.

3. WHAT ARE SOME CHALLENGES IN MEASURING WOMEN'S AND GIRLS' EMPOWERMENT?

4. EMPOWERMENT MEANS DIFFERENT THINGS IN DIFFERENT CONTEXTS, BUT WE MAY ALSO WANT TO COMPARE ACROSS CONTEXTS.

While the underlying concept of empowerment (increasing one's ability to make important life choices) is not specific to a particular context, its concrete manifestations differ by location. For example, a woman's ability to sell goods in a local market may be a relevant constraint in parts of rural Bangladesh, but not in a context where this activity is common for women, like urban Mexico. Since the barriers women and girls face vary considerably by context, so must the way that we measure empowerment, particularly in impact evaluations. If we do not include locally tailored indicators, we run the risk of failing to capture the changes that occur as a result of the intervention. Yet if we only use locally-tailored indicators, we may lose the ability to compare our findings with other studies on empowerment from different contexts or contribute our data to a meta-analysis, which are both useful for drawing broader lessons about effective empowerment programs.

5. PRIORITIZING OUTCOME MEASURES IS DIFFICULT.

Empowerment spans nearly every realm of a woman's life, making it challenging to decide what to focus on, particularly because changes in empowerment can be unpredictable. However, selecting too many outcome measures may lead us to collect data we do not end up using in our analysis, which also increases the cost of research and unnecessarily uses more of the study participants' time. Furthermore, with a large number of outcomes, there is a likelihood that at least one indicator will show a significant change just by chance rather than being a true impact of the program. We need to prioritize the most relevant outcomes and to be realistic about what a given program can change.

6. MEASURING WOMEN'S PREFERENCES IS CHALLENGING IN CONTEXTS WHERE WOMEN HAVE INTERNALIZED SOCIETY'S VIEWS.

Power often operates invisibly through our everyday social institutions, norms, and habits. Women may internalize society's views about their status and have preferences that reflect and accept inequality between men and women. For example, fourteen percent of women responding to a global ILO/Gallup poll in over a hundred countries responded that it is not acceptable for a woman in their families to have a paid job outside the home if she wants one.[15] Researchers need to be careful not to impose an outsider view of what women should want, and yet we also need to keep in mind that women's preferences may reflect society's views about gender rather than their own true preferences. Even though women's preferences are an important component of empowerment, measuring preferences alone may not always fully reflect women's ability to make a meaningful choice.

7. DISEMPOWERMENT CAN HEIGHTEN DATA COLLECTION CHALLENGES.

Related to the challenge above, when women lack power and voice, it may be more difficult to collect data about their aspirations, opinions, and desires. In some cases, for example, women may find it hard to answer questions about goal-setting or their plans for the future. It can also sometimes be challenging to find and hire women enumerators to interview women. When working with particularly vulnerable populations, like women who may have experienced abuse or violence, we must be even more thoughtful about the ethical implications of our research and ensure that the consent process takes unequal power dynamics into account (see page 11 for more information).

There are many approaches to overcoming these seven challenges. Table 1 on the next page summarizes some key tips for overcoming them, along with references to sections in this guide with more information.

[15] Gallup, International Labor Organization. 2017, "Towards a Better Future for Women and Work: Voices of Women and Men." ILO-Gallup Report, March 8, 2017. 74. http://www.ilo.org/global/publications/books/WCMS_546256/lang--en/index.htm.

3. WHAT ARE SOME CHALLENGES IN MEASURING WOMEN'S AND GIRLS' EMPOWERMENT?

TABLE 1. SEVEN CHALLENGES IN MEASURING EMPOWERMENT AND APPROACHES TO OVERCOMING THEM

This table lists approaches to overcoming seven common challenges in measuring women's empowerment, along with the page numbers where you can find more in-depth information. Challenges one through five are addressed on pages 21-29, while challenges six and seven are addressed in various sections which are noted in the table.

CHALLENGE	APPROACHES FOR OVERCOMING THE CHALLENGES	PAGE NUMBER
1. Measuring people's ability to make important life choices is difficult because we rarely observe decision-making directly.	**Combine one or more of the following approaches:** • Ask people about specific decision-making processes. • For joint decisions, ask more than one person about the decision-making process. • Measure women's and men's preferences at baseline and track whether outcomes move in the direction of women's preferences after the program. • Measure the psychological components of agency: women's ability to set goals in line with what they value and act on them. • Try to observe a choice directly or create a situation in which you can observe a choice. • Measure fundamental outcomes related to well-being.	21-25
2. Empowerment is a process.	• Track each major step along the causal chain: resources, agency, and achievements. • Use panel data to observe changes in the same people over time.	25-26
3. Many aspects of empowerment are susceptible to reporting bias.	• When possible, complement indicators that are subject to reporting bias with more objective indicators or proxy indicators. • Triangulate an outcome using multiple indicators or perspectives when we don't have the ideal measure. • Frame the question indirectly by asking about a hypothetical situation. • Consider using a non-survey instrument that is less susceptible to reporting bias.	27-28 34-36, Appendix 2

3. WHAT ARE SOME CHALLENGES IN MEASURING WOMEN'S AND GIRLS' EMPOWERMENT?

CHALLENGE	APPROACHES FOR OVERCOMING THE CHALLENGES	PAGE NUMBER
4. Empowerment means different things in different contexts, but we may also want to compare across contexts.	• Use findings from formative research to select or develop locally tailored indicators and questions. • Pilot extensively in the field before launching a survey. • Complement context-specific indicators of empowerment with more standard ones.	12-17 37-40 28
5. Prioritizing outcome measures is difficult.	• Prioritize indicators that are related to the main objectives of the program (or can be expected as an important spillover), de-prioritize indicators that are not.	29
6. Measuring women's preferences is challenging in contexts where women have internalized society's views.	• Measure inequalities along with the extent to which these inequalities are accepted. • Measure changes in important life outcomes like health, education, or economic status that are widely accepted as signals of improved well-being.	25 21, 25
7. Disempowerment can heighten data collection challenges.	• In addition to university IRB approval, seek feedback and approval from local IRB institutions, local leaders, and respondents' family members if needed. • Consider enumerator identity carefully, create private spaces for interviews, and/or use technologies that allow respondents to answer privately.	11, 39, 42 41-43

4. HOW CAN WE BUILD A RELIABLE STRATEGY FOR MEASURING EMPOWERMENT IN AN IMPACT EVALUATION?

In the rest of the guide, we will walk through lessons we have learned about how to measure women's and girls' empowerment during the different stages of an impact evaluation. Box 1 lists the main steps that we cover for developing a good measurement strategy. While the steps are listed in approximate chronological order, in reality we often work on several simultaneously. Before an evaluation begins, it is important to conduct formative research to understand gender and empowerment issues in a particular context prior to designing our evaluation and data collection instruments. Using our findings from this pilot research, we can map a theory of change for the intervention being evaluated, which will help us prioritize outcomes and indicators. Once we have decided what to measure, we need to develop and pilot our data collection instruments in communities similar to ones where the evaluation will take place. This is an important reality check to make sure our surveys and non-survey instruments work in local conditions and pick up the information we are interested in. Based on our piloting, we can design a data collection plan to minimize measurement error, including deciding where, when, and how often to gather data, and ensuring people are comfortable speaking with our enumerators.

ETHICAL CONSIDERATIONS

Before initiating any fieldwork, we must ensure that our data collection plans meet international standards and norms intended to protect research participants. In the academic context, our research design, data collection instruments, and consent forms must be reviewed and approved by the Institutional Review Boards (IRBs) at our host academic institutions. In many cases, they must also be reviewed and approved by local IRBs or government research councils in the countries where we are conducting research. Many research institutions outside academia have their own IRBs. As IRB requirements can vary, we should check with our host institution to see what is required before starting any interaction with human subjects, even before the formative research phase.

BOX 1. DEVELOPING A MEASUREMENT STRATEGY

Step 1. Conduct formative research to understand gender and empowerment in the specific context

Step 2. Map a theory of change and use it to select appropriate outcomes and indicators

Step 3. Develop and validate data collection instruments that minimize reporting bias

Step 4. Design a data collection plan that minimizes measurement error

If we are studying girls' empowerment, we need to incorporate additional protections for working with children, including getting approval from parents or guardians. The risks of participating in research may be elevated for women and girls when they are in marginalized positions, which requires careful consideration when developing research approaches. For example, if we are asking women about their experiences of violence, we should research laws for mandatory reporting of violence, phrase survey questions to minimize distress, consider how to protect survey staff if needed, and get guidance on if and under what conditions referrals for care and support should be given. Note that social science and medical ethics as well as national guidelines can differ on some of these questions (for example, referral to care) and enumerators in many cases may not be qualified to make medical judgments. For more in-depth guidance on measuring domestic violence and intimate partner violence, see Annex 1 of the WHO Multi-country Study on Women's Health and Domestic Violence against Women and Innovations for Poverty Action's resource on conducting violence research.[16] For more information and resources on ethics in impact evaluations in general, see J-PAL's ethics page.[17]

[16] García-Moreno, Claudia, Henrica A.F.M. Jansen, Mary Ellsberg, Lori Heise, and Charlotte Watts. 2005. "WHO Multi-Country Study on Women's Health and Domestic Violence against Women: Summary Report of Initial Results on Prevalence, Health Outcomes and Women's Responses." Geneva: World Health Organization, 101.

Innovations for Poverty Action. 2018. "The Safe and Ethical Conduct of Violence Research: Guidance for IPA Staff and Researchers." Updated June 2018. https://www.poverty-action.org/publication/ipv-ethical-guidance.

[17] Abdul Latif Jameel Poverty Action Lab (J-PAL). N.d. "Ethics." Research Resources. Accessed March 19, 2018. https://www.povertyactionlab.org/research-resources/ethics.

LOCATION: UGANDA. PHOTO: CAMERON BRESLIN | J-PAL/IPA

STEP 1. FORMATIVE RESEARCH

Conduct formative research to understand gender and empowerment in the specific context

We conduct formative research prior to starting an impact evaluation to better understand the local context, the underlying problem that the program is trying to address, and the rationale behind the proposed solutions. As part of this process, we should work closely with our implementing partners to either carefully design and pilot the intervention, or to ensure that the program is being implemented as designed. It also helps us better formulate the research questions for the evalution. Formative research is a combination of field and desk research, both qualitative and quantitative, that helps us gain a deeper understanding of empowerment and barriers to it in a particular context.[18]

Good formative research creates repeated opportunities to listen to people living in the communities in which we are working. Some common formative research exercises include needs assessments or stakeholder analyses; common research methods include semi-structured interviews, focus groups, and direct observation.[19]

Whatever the methods we choose, it is critical to devote a significant amount of time to better understanding local gender dynamics and the barriers to empowerment. This should include analyzing resources, agency, and achievements that women want, currently have, and do not have. The findings from our formative research will shape nearly every aspect of the evaluation design, from our theory of change to indicator selection, what questions we ask, and whom to survey. Because it is so integral to every step of the impact evaluation, it is not advisable to simply subcontract this work out to others. Collaborating with researchers who specialize in qualitative methods can also help improve the quality and rigor of our formative research.

[18] Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press, 66-81.

[19] Many economists that work on impact evaluations conduct this kind of formative research, some more and less formally, and they may use different names to describe these activities. In addition, some of this background research is conducted prior to designing the intervention itself and can help shape the design of the program, as well as the research.

STEP 1. FORMATIVE RESEARCH

WHAT ARE SOME COMMON FORMATIVE RESEARCH EXERCISES?

Needs assessments and stakeholder analyses are formal tools we can use to systematically gather and analyze information on gender dynamics and barriers to empowerment in a specific context. It is not necessary to structure our formative research process in this manner, but it can be helpful if we are working in a new context and need to gain an in-depth understanding of these issues.[20]

A needs assessment[21] documents the gap between the current and desired state of the world for a particular issue by gathering information from many different stakeholders through interviews, focus groups, direct observation, and in some cases existing or new surveys. It helps us hone in on the key needs or barriers people face in solving their problems and how the program implementers have in mind could be designed to address them. It can also be particularly useful for understanding the nature, consequences, and potential drivers of problems in order to link them to potential solutions. The World Bank's free book *A Guide to Assessing Needs* is a helpful resource for conducting a needs assessment.[22]

A stakeholder analysis uses data from interviews, focus groups, and direct observation to document the relative power and interest of the various actors involved in a particular issue. It is another tool that can help us understand and map gender dynamics relevant to our evaluation, as well as important actors to include in our evaluation. For example, a stakeholder analysis for a program that seeks to increase women's participation in local politics might help us identify the possibility of opposition from certain groups of community members that could undermine the success of the program. We might consider adding additional survey modules measuring these community members' reactions as a result. Kammi Schmeer's "Stakeholder Analysis Guidelines" and Bryson and Quinn Patton's "Analyzing and Engaging Stakeholders" are useful resources for conducting a stakeholder analysis.[23]

For an impact evaluation on empowerment, we can use these frameworks to assess what resources are available or denied to women or girls, where they lack the ability to make life choices that are important to them, what choices they would make if they could, and what changes are possible. It is also important to assess how gender norms and constraints may interact with the program being tested.

HOW DO WE COLLECT DATA IN FORMATIVE RESEARCH?

We should use a combination of field and desk research during this stage. We can start by reading the existing literature on gender in our context and consulting other implementers and researchers with expertise in the study context. In addition to reading work from anthropology, history, and/or sociology, it is also useful to read descriptive quantitative studies or other impact evaluations from the same context.

Listening to the perspectives of people in communities similar to those where we will conduct the evaluation is the most important element. We typically use direct observation, semi-structured interviews, and/or focus groups to collect these kinds of data. Some researchers also use immersion exercises and/or participatory

[20] Needs assessments and stakeholder analyses typically occur as inputs for program or intervention design. Thus, if we are evaluating an existing program, we may be able to get this information from our implementing partners. However, if we are designing new treatment arms or cannot access these preexisting analyses, it may be helpful to structure our formative research in this way.

[21] This can also be called a situation analysis, landscape analysis, or need statement, among other terms.

[22] Watkins, Ryan, Maurya West Meiers, and Yusra Laila Visser. 2012. "A Guide to Assessing Needs: Essential Tools for Collecting Information, Making Decisions, and Achieving Development Results." Washington, DC: World Bank. http://hdl.handle.net/10986/2231.

[23] Bryson, John M., and Michael Quinn Patton. 2010. "Analyzing and Engaging Stakeholders." *Handbook of Practical Program Evaluation* Fourth Edition: 36-61. https://experts.umn.edu/en/publications/analyzing-and-engaging-stakeholders.

Schmeer, Kammi. 2000. "Stakeholder Analysis Guidelines." Policy Toolkit for Strengthening Health Sector Reform 2: 1-43. http://www.who.int/workforcealliance/knowledge/toolkit/33.pdf.

research methods. Collaborating with trained qualitative researchers as co-investigators can be valuable for applying the necessary rigor in the data collection and analysis process. However, we should not subcontract this work out to others, as there is simply no substitute for spending time in the field ourselves.

We only give a brief summary of common qualitative methods here, as there are many existing in-depth resources by researchers and practitioners with expertise in qualitative methods, including Michael Quinn Patton's *Qualitative Research & Evaluation Methods* and FHI360's free field guide on qualitative methods.[24]

Direct observation is critical and enables us to get a more in-depth picture of issues and opportunities related to women's and girls' empowerment in a particular context. For example, if there are groups of women that regularly meet in the community, a representative from the research team may be interested in observing several of these meetings to get detailed information about the issues important to them.

Semi-structured interviews allow for open-ended discussion, enabling new ideas or issues to emerge that we may not have previously thought of. They are one-on-one interviews that follow a general script of open-ended questions but allow flexibility to ask impromptu follow-up questions. We should strategically interview people from different parts of the community so that we receive diverse perspectives. In addition to helping identify the barriers women face, we can also use semi-structured interviews to generate a list of qualities that people in the community perceive as signs that a woman is empowered. The most frequent answers to these questions can help inform the indicators we select to measure resources, agency, and achievements. There is no magic number for how many interviews to conduct; the goal is to repeat them until they no longer yield new meaningful information. Oxfam's "Conducting Semi-Structured Interviews" provides some useful guidance on carrying out semi-structured interviews.[25]

LOCATION: PHILIPPINES. PHOTO: AGNESE CARRERA | J-PAL/IPA

[24] Mack, Natasha, Cynthia Woodsong, Kathleem M. MacQueen, Greg Guest, and Emily Namey. 2005. "Qualitative Research Methods: A Data Collector's Field Guide." Research Triangle Park: Family Health International. https://www.fhi360.org/sites/default/files/media/documents/Qualitative%20Research%20Methods%20-%20A%20Data%20Collector%27s%20Field%20Guide.pdf.

Patton, Michael Quinn. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* Fourth Edition. Thousand Oaks, California: SAGE Publications.

Hennink, Monique, Inge Hutter, and Ajay Bailey. 2011. *Qualitative Research Methods*. Thousand Oaks, California: SAGE Publications.

[25] Raworth, Kate, Caroline Sweetman, Swati Narayan, Jo Rowlands, and Adrienne Hopkins. 2012. "Conducting Semi-Structured Interviews." Oxfam. https://policy-practice.oxfam.org.uk/publications/conducting-semi-structured-interviews-252993.

LOCATION: KENYA. PHOTO: THOMAS CHUPEIN | J-PAL

Focus groups are semi-structured interviews with groups of people led by trained moderators. We can use them to understand how communities or societies think about a particular issue in general, like women's access to education or health services, rather than getting an individual's opinion. We should keep in mind that charismatic or powerful individuals can shape or drown out others' opinions. For this reason, it is important to complement focus groups with individual interviews. We can use our discretion to decide how many focus groups to conduct to get the right amount of information. We may find that five are enough if we start to hear the same answers over and over again, or we may want to implement more if each one is still yielding new and interesting information. Just like interviews, focus groups should be conducted with different kinds of people in order to get different perspectives on the same issue. Participative Ranking Methodology (PRM) is one kind of focus group methodology that may be useful in developing locally grounded indicators of empowerment (see Box 2 for more details).

BOX 2. PARTICIPATIVE RANKING METHODOLOGY (PRM)

PRM uses a ranking process that can be used to gather data on how members of a community understand a specific concept and how they view the relative importance of the categories used to explain it.[26] Researchers formulate a key research question, such as "how can you tell if a girl under the age of eighteen is doing well in your community?" A trained moderator gathers a small focus group of roughly seven to fifteen participants and leads a discussion on the question until the group agrees on approximately five to eight categories and ranks them in order of importance. The focus groups are repeated with different groups of community members. Researchers can then analyze how often each category was chosen across groups and use the most frequently cited categories to develop locally tailored indicators.

[26] Ager, Alastair, Lindsay Stark, and Alina Potts. 2010. "Participative Ranking Methodology: A Brief Guide: Version 1.1." New York, NY: Program on Forced Migration & Health, Mailman School of Public Health, Columbia University. https://doi.org/10.13140/RG.2.2.34356.45448.

Immersion visits: Immersion visits are a more in-depth form of direct observation. In an immersion visit, researchers and/or practitioners travel to live in a community similar to where the evaluation will take place for a short period of time. See Box 3 for an example of an immersion visit from a program and evaluation in Burkina Faso, led by the implementing partner that designed the program, Development Media International (DMI).

In some cases, the formative research itself can change people's opinions or views. For example, asking people to talk about topics they don't normally discuss in focus groups or interviews might shift how they think about these topics. We conduct formative research in areas outside but similar to where the evaluation will take place to avoid introducing an unintended intervention in our evaluation sample.

BOX 3. CASE STUDY: USING IMMERSION VISITS TO INFORM QUESTIONNAIRE DESIGN

Rachel, Joanna Murray, and Victor Pouliquen, are working with Development Media International (DMI) to conduct a randomized evaluation measuring the impact of DMI's mass media campaigns on family planning and gender norms in Burkina Faso. DMI sent staff (who were mostly based in the capital Ouagadougou) to visit different rural parts of the country for week-long immersion visits to gain a better understanding about how women and men think and talk about family planning. In these visits, DMI discovered critical information that helped the research team and DMI develop locally tailored and culturally appropriate mass media campaigns, indicators, and survey questions:

> *"In a country where only 11% of women in rural areas use modern contraceptives, we assumed that awareness about contraception methods would be fairly low. Our research showed that almost everyone knew of at least one modern method—yet there were a lot of myths about the side effects. Rather than just encouraging people to use contraception methods and educating them about the different options available, we will need to address those myths in our campaign…We also found some major differences in behaviors and knowledge across our intervention areas. These are partially reflected in the local languages—in three of the local languages we studied, there were different words to describe family planning, birth spacing, and contraception. By contrast, in one of the language groups the same word 'Maapedi/ Mapè' was used for all sof these concepts."*[27]

DMI's in-depth work revealed many new insights that helped shape the survey questionnaires for the evaluation. First, it helped identify the right words to use for "family planning" in the different local language versions of the survey. Second, it helped us realize that we needed to collect data on the prevalence of certain myths about contraceptives in the evaluation because this could be driving low use. DMI's immersion visits taught us which myths were prevalent in these villages. Based on this formative research, our baseline questionnaire asked respondents if they believed that birth control pills, implants, or injections could make a woman sterile or cause sickness to see if these beliefs were driving low use and if they changed over time.

[27] Development Media International. 2016. "Sending Our Writer to Live in a Village for a Week." DMI News, May 23, 2016. http://www.developmentmedia.net/news/sending-our-writers-to-live-in-a-village-for-a-week.

SYNTHESIS AND ANALYSIS OF QUALITATIVE DATA

Once all the information is gathered, researchers and practitioners can work together to code and analyze the data. We can then summarize the current state of women's empowerment in a particular domain, the biggest barriers women face in becoming more empowered, and how the program being designed could potentially address them. It is also beneficial to present the findings to some of the original respondents to get their feedback on our conclusions and make sure they accurately reflect their opinions. We should also document and save the data and analysis from our formative research in a place and format that is easy to access and use throughout the evaluation.

QUANTITATIVE ANALYSIS OF EXISTING OR NEW SURVEY DATA

In addition to these qualitative research exercises, it is useful to examine data from representative surveys. This can help validate our qualitative findings and assess if they apply to the broader population. For example, in the qualitative interviews for Rachel and co-authors' evaluation on young women's empowerment in Bangladesh, many community members and partners stated that menstruation was a major reason why girls miss school. Yet when we conducted our baseline survey in several hundred villages, we found that very few girls missed school due to their periods. Instead, the most common reason for missing school was that the teacher was not present during class time.

We don't necessarily need to collect new data to do this type of analysis. We can often find the information we need, such as the prevalence of particular issues or problems, in publicly available datasets from nationally representative surveys. These include surveys from government statistics bureaus, or USAID's Demographic and Health Surveys, which include modules on everything from women's status and empowerment, to family planning and domestic violence, among others, from many low- and middle-income countries.[28]

MAIN TAKEAWAYS ON FORMATIVE RESEARCH

- Our measurement strategy is only as good as our understanding of gender dynamics and empowerment in the local context. Spending significant time conducting formative research in the field will improve our theory of change, outcome indicators, and survey questions.
- Formative research is a combination of field and desk research that uses methods like semi-structured interviews, focus groups, and direct observation.
- In an evaluation on empowerment, our goal is to use these tools to understand what resources are available or denied to women or girls, where they lack the ability to make life choices that are important to them, what choices they would make if they could, and what changes are possible.
- Collecting qualitative information from many different sources, including women and girls, men and boys, NGOs, and other community stakeholders and comparing our qualitative findings to new or existing representative surveys can help us validate our findings.
- Formative research can also help us consider how influential people and organizations should be incorporated into our data collection strategy.

[28] Demographic and Health Surveys Program. N.d. The DHS Program. Accessed March 19, 2018. https://www.dhsprogram.com.

Demographic and Health Surveys Program. N.d. "Women's Status and Empowerment." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/Topics/Womens-Status-And-Empowerment.cfm.

Demographic and Health Surveys Program. N.d. "Family Planning." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/Topics/Family-Planning.cfm.

Demographic and Health Surveys Program. N.d. "The Gender Corner." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/topics/gender-Corner/index.cfm.

LOCATION: INDIA. PHOTO: NIKITA TANIPARTI | J-PAL

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS

Map a theory of change to select appropriate outcomes and indicators

MAPPING A THEORY OF CHANGE

The next step is to use the findings from our formative research to refine our theory of change for the program being tested. To identify good outcomes (the change or impact we expect to see) and indicators (observable signals we use to measure that change), we need a deep understanding of the pathways through which the program can affect people's lives. A theory of change framework provides a structured approach to mapping these potential pathways using the findings from our formative research, along with relevant theory or lessons from completed impact evaluations. It documents a program's logical chain of results—from the inputs to the outputs to the intermediate and final outcomes—along with the indicators to track each major step along the chain. We can also use it to specify the necessary assumptions to get from one step to another and the possible risks that could break the causal chain. It is beneficial for researchers and implementers to jointly develop and refine the theory of change for a program, even when they are part of the same organization.

In an evaluation about empowerment, we can use a theory of change to map changes in women's resources, agency, and achievements, which often correspond to short-term, intermediate, and final outcomes. Box 4 shares an example of a logical framework summarizing the theory of change for an evaluation of a government policy instituting quotas for women leaders in village councils in India, in which researchers measured changes in women's resources, agency, and achievements. For a more detailed guide on building a theory of change, see Chapter 5.1 of *Running Randomized Evaluations: A Practical Guide*[29] or J-PAL's research resources on Measurement and Data Collection.[30]

[29] Glennerster and Takavarasha 2013, 180-190.

[30] Abdul Latif Jameel Poverty Action Lab (J-PAL). *N.d.* "Measurement & Data Collection." Research Resources. Accessed March 19, 2018. https://www.povertyactionlab.org/research-resources/measurement-and-data-collection.

BOX 4: BUILDING A THEORY OF CHANGE TO MEASURE THE IMPACT OF WOMEN LEADERS IN INDIA

In the early 1990s, a constitutional amendment in India called for a random one-third of village council leader, or pradhan, positions to be reserved for women. The village council, which encompasses between five and fifteen villages, is responsible for the provision of local infrastructure—such as public buildings, water, and roads—and for identifying government program beneficiaries. The village council is required to organize two village meetings per year, during which they present their proposed budget and report on their activities in the previous six months.

J-PAL affiliated researchers Raghabendra Chattopadhyay (Indian Institute of Management) and Esther Duflo (MIT) took advantage of the random assignment embedded in the reform to study the impact of mandated representation of female policymakers on the provision of social services between male- and female-led village councils. Table 2 shows a logical framework depicting the policy's theory of change. In short, if quotas for women leaders are implemented and more women made chairs of their village council, then women would become more involved in council policy discussions, and the council's investment decisions will better reflect women's preferences. Finally, the quality of the public goods preferred by women would increase.

The researchers used several innovative indicators for measuring outcomes related to empowerment. Rather than rely on self-reported measures of women's participation in village council meetings, they used minutes of the village meetings to count how many women spoke up. They analyzed records of the requests submitted to the village council by women and men to identify whether women's preferences for investments in local public goods differed from those of men. They also tracked how women's preferences varied by state. This classification allowed them to measure whether councils' ultimate investments better reflected women's preferences when leadership positions were reserved for women. Results from the evaluation suggest that reservations for female leaders affected policy decisions in ways that better reflected women's preferences.[31]

[31] Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica* 72(5): 1409-1443. https://doi.org/10.1111/j.1468-0262.2004.00539.x. Read a full summary of the study at: https://www.povertyactionlab.org/evaluation/impact-women-policy-makers-public-goods-india.

BOX 4 CONTINUED: BUILDING A THEORY OF CHANGE TO MEASURE THE IMPACT OF WOMEN LEADERS IN INDIA

Table 2. Logical framework summarizing the theory of change of a policy to increase women's participation in village councils in India.[32]

		DESCRIPTION	INDICATOR	ASSUMPTIONS/RISKS
RESOURCES	INPUTS	Quotas for women are passed	Passage of legislation in state legislature	Supreme court mandate is translated into effective legislation at the state level
RESOURCES	OUTPUTS	There are more women leaders in village councils (gram panchayats)	Number of women leaders in council chair positions	Quota legislation is implemented as designed in villages
AGENCY	OUTCOMES	Political participation	Number of women speaking at general village council meetings	Seeing women leaders emboldens women to speak up at meetings.
ACHIEVEMENTS	IMPACTS	Public goods investments more closely match women's priorities	Types of public goods mentioned in women's queries vs. men's queries Number of public goods of different types Repairs to public goods by type Recently built public goods by type	Women's preferences for public goods differ from men's. The system is democratic enough to respond to an increase in women's queries Responses to political pressure from women will impact what gets repaired. New investments will be more in line with women's needs than older investments.
ACHIEVEMENTS	IMPACTS	The quality of public goods that are priorities for women improves	Reduced presence of microbes in drinking water	Greater investment in women's priority areas will improve quality of services.

[32] Glennerster and Takavarasha 2013, 189.

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS: CHALLENGES AND TIPS

Tips for selecting outcomes and indicators for empowerment that can help address common measurement challenges

There are many challenges in measuring empowerment. The following section outlines some tips to address the challenges related to the selection of outcomes and indicators.

CHALLENGE 1: MEASURING PEOPLE'S ABILITY TO MAKE IMPORTANT LIFE CHOICES IS DIFFICULT BECAUSE WE RARELY OBSERVE DECISION-MAKING DIRECTLY.

Five approaches to identifying outcomes and indicators that capture agency and decision-making

When it comes to measuring empowerment, the outcomes related to agency and decision-making are often the most challenging to select. How can we measure the ability to make a choice when we only observe the outcomes of choices and not the process itself? There are several ways researchers try to tackle this challenge: measuring fundamental outcomes and tracking change over time, asking people about specific decision-making processes, measuring women's and men's preferences to see if outcomes move in the direction of women's preferences, measuring psychological aspects of women's abilities to set goals in line with their values and act on them, and observing choices directly. Each option has strengths and weaknesses, and combining two or more can help overcome their individual limitations. In the ideal scenario, we can measure the full process: tracking whether a program changes resources available to women, if this leads to changes in women's agency, and finally if it also results in changes in well-being.

1. Measuring fundamental outcomes: Some outcomes are so fundamental to well-being that we often assume they reflect people's status and decision-making ability. For example, if women are more malnourished than men, this likely reflects women's lack of choice and agency. Measuring objective outcomes like this and tracking changes over time can help avoid reporting bias, yet this approach is most relevant in resource-poor settings or situations of severe poverty. For example, in contexts where malnutrition is not prevalent, it is likely not a good indicator of whether women have choice.

2. Asking people about decision-making processes: A more direct way to measure choice is to ask people about the decision-making process itself. The most commonly used survey modules, often adapted from USAID's Demographic and Health Survey (DHS), ask women and men general questions about who makes decisions in their household. For example: "Who usually makes decisions about [healthcare for yourself]/[major household purchases]/[visits to your family or relatives] you, your husband/partner, you and your husband jointly, or someone else?"[33]

Asking this kind of general question is useful for comparing participation in household decision-making across contexts. While internationally standardized metrics are valuable, we should be thoughtful about what kinds of indicators will realistically pick up changes in our context. Asking more specific questions about a concrete scenario tailored to the choices women care most about in our study context may be easier for them to answer accurately, and may tell us more about whether they can make choices that matter to them. For example, we could ask, "if your child is sick and needs immediate health care, but your husband is not home, what would you do?" Or, "if you ever need medicine for yourself (for a headache, for example), could you go buy it yourself?" In the evaluation Rachel and coauthors

[33] Demographic and Health Surveys Program. 2017. "DHS Model Questionnaire – Phase 7." The DHS Program. w-64 row 922. Accessed March 19, 2018. https://dhsprogram.com/pubs/pdf/DHSQ7/DHS7-Womans-QRE-EN-07Jun2017-DHSQ7.pdf.

conducted in Bangladesh, our team got different answers when we asked the specific versus general questions about health care decisions, implying that the general decision-making questions were not picking up on the characteristics we expected by themselves.[34]

A fruitful area for future measurement research is to conduct more validation exercises comparing different methods for asking about complex concepts like decision-making. More systematic investigations could help us determine whether it is worth making improvements or additions to current widely used questions. A recent example is an analysis of DHS data from twenty countries by World Bank researchers, which finds women's responses to the general DHS decision-making questions are correlated with several other empowerment indicators.[35] Another useful example from International Food Policy Research Institute (IFPRI) compares responses to decision-making questions with questions about autonomy and makes the case for calibrating questions to specific contexts.[36]

Until there is more data on concrete ways to improve decision-making questions, combining these kinds of questions with more objective outcome indicators (like education, health, income, etc.) or asking questions about the same concept in two different ways (e.g., asking about a concrete scenario and decision-making in general) can help us determine whether changes in general decision-making indices are meaningful.

Many researchers and practitioners have also created useful domain-specific survey modules for decision-making, like International Center for Research on Women (ICRW)'s modules for measuring economic empowerment and IFPRI, USAID, and The Oxford Poverty & Human Development Initiative (OPHI)'s index for measuring women's empowerment in agriculture (WEAI).[37] Oxfam's "A 'How To' Guide to Measuring Women's Empowerment" provides guidance on constructing a women's empowerment index that includes decision-making and can be applied in different domains.[38]

3. Comparing gender preferences to changes in outcomes: We can ask women and men about their preferences before an intervention and track whether the outcomes we observe after it move more in the direction of women's preferences. For example, in the study on women policymakers in India, researchers found that women's preferences for investments in local infrastructure were different than men's, and tracked new infrastructure projects in communities to see whether they moved to be more in line with women's preferences.[39] This approach is useful because it explicitly measures what women want, a key part of the definition of empowerment, along with how their preferences relate to changes in outcomes. Yet it only works if men and women have different preferences. In an evaluation in Sierra Leone, for example, Rachel and co-authors found no systematic differences in gender preferences for the uses of community driven development

[34] In response to the standard decision-making question, sixteen percent of women said they usually make decisions about their healthcare alone or jointly with their husbands. Given this response, we would call this group more empowered—yet nearly a quarter of this group also said they could not take a sick child to the doctor until their husbands came home. We also found discrepancies in the other direction: Over half of the women who appeared disempowered according to the standard question said that they could take a sick child to the doctor on their own, and even more telling, could buy medicine for themselves. For more see, Glennerster, Rachel, and Claire Walsh. 2017. "Is It Time to Re-Think How We Measure Women's Household Decision-Making Power?" Abdul Latif Jameel Poverty Action Lab, September 7, 2017. https://www.povertyactionlab.org/blog/9-6-17/it-time-rethink-how-we-measure-women's-household-decision-making-power-impact. The 2017 book *Measuring Women's Economic Empowerment: Critical Lessons from South America* also describes scenarios in which standardized indicators failed to reflect women's decision-making constraints (Martínez-Restrepo, Susana, and Laura Ramos-Jaimes (eds.). 2017. *Measuring Women's Economic Empowerment: Critical Lessons from South America*. Springfield, VA: IDRC, Fedesarrollo. http://hdl.handle.net/11445/3482).

[35] Donald et al. find that women who report having greater sole or joint-decision making power were also more likely to own land, work outside the home, earn more than their husbands, and not condone domestic violence—outcomes we typically think of as signs of empowerment. Donald et al. 2017, 34-35.

[36] Seymour, Greg, and Amber Peterman. 2017. "Understanding the Measurement of Women's Autonomy: Illustrations from Bangladesh and Ghana." IFPRI Discussion Paper No. 1656. https://www.ifpri.org/publication/understanding-measurement-womens-autonomy-illustrations-bangladesh-and-ghana.

[37] Golla et al. 2011; Alkire et al. 2013.

[38] Lombardini et al. 2017.

[39] Chattopadhyay and Duflo 2004, 1411.

grants, so this approach was not an option. Furthermore, measuring preferences can be difficult if women's preferences change over time or if people's stated preferences reflect current social norms more than they reflect their true preferences (see Box 5 on page 25). It is useful to combine this approach with intermediate outcomes measuring women's participation in decisions in order to draw the causal link between participation in decisions and changes in outcomes.

4. Measuring the psychological aspects of agency: We can also use tools from psychology to try to measure the psychological components of agency. For example, in Mexico, Manuela Angelucci (UT Austin) and J-PAL affiliated researchers Dean Karlan (Northwestern) and Jonathan Zinman (Dartmouth) found that access to microcredit reduced depression among potential borrowers as measured by a depression index.[40] We can also try to measure a woman's ability to: 1) set goals to achieve things she values; 2) perceive that she can achieve these goals; and 3) act on these goals. One tool for measuring the ability to set goals that are aligned with one's own preferences and values is the Relative Autonomy Index (RAI). It measures whether a person believes their actions are driven by their own goals or by external factors like social norms or coercion.[41] Locus of control scales quantify the degree to which a person thinks life events are caused by their own behavior or external factors, and may be useful for measuring whether people perceive that they have the capacity to achieve their goals. Self-efficacy scales ask people about their confidence in completing actions and are another popular way to measure whether people perceive they can achieve their goals.[42] Many surveys also just ask people about whether they feel they have freedom of choice.[43] To measure a person's ability to act on their goals, we can use some of the other approaches we have already discussed, such as measuring their participation in decision-making and outcomes related to well-being.

LOCATION: MEXICO. PHOTO: ALEJANDRA ROGEL | J-PAL/IPA

One note of caution is that people in different contexts may find that the questions from tools like the RAI, locus of control, or self-efficacy scales don't make sense to them or are difficult to answer. Before adding them to our questionnaire, we should be sure to first extensively pilot, validate, and if needed, adapt them to make sure they work in our context. For an in-depth discussion of the pros and cons of these measures, see the recent World Bank paper "Measuring Women's Agency."[44]

[40] Angelucci, Manuela, Dean Karlan, and Jonathan Zinman. 2015. "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." *American Economic Journal: Applied Economics* 7 (1): 151-82. 174. http://dx.doi.org/10.1257/app.20130537.

[41] Donald et al. 2017, 7.

[42] Donald et al. 2017, 17.

[43] Donald et al. 2017, 20.

[44] Donald et al. 2017.

LOCATION: SIERRA LEONE. PHOTO: GLENNA GORDON | J-PAL/IPA

5. Observing choices directly: The last approach is to observe women and men making choices directly, either in a real-world setting or through a game or structured community activity. These methods are beneficial because by observing actions directly we do not have to rely on people's reports of whether they participated in decisions. This type of direct observation can occur when decision-making is taking place in large or small group settings or between individuals. For example, we could count how many times women speak up in community meetings and see if it increases after the intervention takes place.

We could also create a scenario in which people have to make a decision during a survey and examine whether women exercise more agency. For example, in an evaluation of a community-driven development program in Sierra Leone, Rachel and co-authors gave communities the choice of two different thank-you gifts after we finished surveys. Enumerators recorded whether and how women participated in this community decision and whether they had an effect on which option was ultimately chosen.[45] We could also prompt a similar decision at the household level and record how men and women's preferences are incorporated into the final decision. For example, researchers created a proxy measure of women's bargaining power in a marriage by offering women a choice of getting a slightly smaller cash transfer delivered directly to her or getting a slightly larger cash transfer delivered to her husband.[46] While these approaches avoid reporting bias, they tend to be more expensive to implement than surveys and require additional enumerator training to do well. See Appendix 2 for more details on these and other quantitative non-survey instruments and when they may be appropriate to use.

[45] Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127 (4): 1755-1812. https://doi.org/10.1093/qje/qje027.

[46] Almås, Ingvild, Alex Armand, Orazio Attanasio, and Pedro Carneiro. Measuring and Changing Control: Women's Empowerment and Targeted Transfers." National Bureau of Economic Research Working Paper No. w21717, 2015.

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS: CHALLENGES AND TIPS

BOX 5: SOME CHALLENGES IN MEASURING PREFERENCES

In contexts with high inequality between men and women, women may internalize their society's views that they are of lower status, which may be reflected in the preferences they share in surveys. For example, fourteen percent of women stated in a recent global ILO Gallup poll that it is not acceptable for a woman in the family to have a paid job outside the home if she wants one.[47] It is important to measure inequalities in basic needs and opportunities along with the extent to which these inequalities are accepted. We need to be careful not to impose an outsider view of what women should want, yet we also have to be careful that women may be reflecting society's view of them rather than their actual preferences. In these cases, we may not want to use preferences as a way to help us measure agency.

Another challenge with measuring preferences to understand agency is that empowerment programs often try to change women's preferences. The empowerment literature discusses the ability to "imagine the possibility of having chosen differently" as an important dimension of meaningful choice and empowerment.[48] This implies women recognizing that there is another way that they could be treated and that women may be able to change their own situations. Economists tend to think about this using the idea of changing preferences, but the challenge is that it is hard to make clear statements about welfare when people's preferences change.[49] One way to avoid this issue is to decide that meaningful choices are ones that lead to changes in important outcomes like health, education, and income, as discussed above. We could then measure both changes in women's ability to make a choice and changes in important outcomes that correspond to those choices.

[47] Gallup, International Labor Organization 2017, 31.

[48] Kabeer 1999, 441.

[49] Grüne-Yanoff, Till, and Sven Ove Hansson, eds. 2009. *Preference Change: Approaches from Philosophy, Economics and Psychology*. Dordrecht: Springer Science & Business Media.

CHALLENGE 2: EMPOWERMENT IS A PROCESS

Tip 1: Track each major step along the causal chain along with changes over time.

Measuring both intermediate and final outcomes can help us make a credible claim that changes in women's outcomes are the result of their increased agency. For example, in the evaluation of reservations for women in village councils in India, it would have been more difficult to claim that the increase in spending on drinking water and road infrastructure was the result of women's increased political participation without evidence that women voiced more opinions related to these public goods in village council meetings.[50] We can also capture steps in the causal chain by measuring changes in women's resources, agency, and achievements (see Table 3). The United Nations Foundation's guidance on Measuring Women's Economic Empowerment, for example, identifies direct, intermediate, and final outcomes to prioritize.[51]

Collecting panel data that tracks changes in the lives of the same people over time can also help us better measure the process of empowerment. We can also first ask women about their goals and plans, and measure progress over time. For example, we asked young women in Bangladesh about their education and income-generating goals for the future and tracked changes in these outcomes over time. However, given that the empowerment process itself can lead women to reshape their plans and visions of the future, our interpretations should be flexible. In Bangladesh, for example, we spoke to one woman who described how having a job in the city changed some of her perceptions: rural society felt confining when she returned home from working in Dhaka.

[50] Chattopadhyay and Duflo, 2004.

[51] Buvinic, Mayra, and Rebecca Furst-Nichols. 2013. "Measuring Women's Economic Empowerment: Companion to A Roadmap for Promoting Women's Economic Empowerment." United Nations Foundation and Exxon Mobil Foundation. http://womeneconroadmap.org/sites/default/files/Measuring%20Womens%20Econ%20Emp_FINAL_06_09_15.pdf.

For more information on the roadmap, see the website: United Nations Foundation and Exxon Mobile Foundation. 2013. "Women's Economic Empowerment: A Roadmap." Last accessed March 20, 2018. http://womeneconroadmap.org/.

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS: CHALLENGES AND TIPS

TABLE 3. MEASURING THE PROCESS OF EMPOWERMENT BY MEASURING RESOURCES, AGENCY, AND ACHIEVEMENTS

A good way to measure the process of empowerment is to measure women's access to resources to make a decision, their agency and participation in the decision itself, and the final outcomes of these choices in terms of well-being. The table below summarizes how we attempted to do this when we measured young women's empowerment related to reproductive health in our evaluation on child marriage and empowerment in Bangladesh.

	CONCEPT	EXAMPLE INDICATOR(S)
RESOURCES	Knowledge	Knowledge about different forms of contraception, risks of early marriage and teen pregnancy, and availability of health services
	Preferences	Preferred number of children and spacing between births
	Access to health care	Past use of health services, use of contraception
AGENCY	Negotiation	Talks to spouse about contraception
	Decision-making	Has a say in decisions about contraception (whether and which type to use)
ACHIEVEMENTS	Child marriage	Age of marriage
	Reproductive health	Age of first birth, maternal morbidity

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS: CHALLENGES AND TIPS

CHALLENGE 3: MANY ASPECTS OF EMPOWERMENT ARE SUSCEPTIBLE TO REPORTING BIAS

Tip 1: When possible, complement indicators that are subject to reporting bias with more objective indicators or proxy indicators.

People may not feel comfortable speaking freely in response to survey questions about many topics related to women's empowerment. Measuring a more objective or proxy indicator of the behavior in addition to gathering self-reported data can help us mitigate the consequences of this reporting bias. Say we want to examine the effect of a program on sexual behavior, sexually transmitted infection (STI) incidence, and decisions about childbearing. STI tests are far more reliable than self-reported data, but they are also more expensive. The incidence of childbearing among young women can be used as a proxy for the incidence of unprotected sex, and childbearing is much easier to observe objectively than sexual behavior. In Uganda, to measure whether an empowerment program increased young women's control over decisions about sex, marriage, and having children, J-PAL affiliated researchers Oriana Bandiera (London School of Economics), Robin Burgess (London School of Economics), Imran Rasul (University College London), and co-authors used more objective indicators, such as incidence of teen pregnancy, early marriage, and cohabitation. They complemented these indicators with self-reported measures, such as the share of girls reporting having sex against their will and the ages at which they aspire to get married and/or have children.[52]

Tip 2: Triangulation can help capture outcomes that are challenging to measure or susceptible to reporting bias.

When we don't have one ideal indicator, it can be helpful to use more than one metric. One survey question cannot give us a comprehensive understanding of whether a woman has a say in household spending decisions, for example. Combining responses from several survey questions into indices or families of indicators can help when we are trying to capture different aspects of an overall concept or cross-validate self-reported responses. This can also help when a more objective or proxy indicator is not available. For example, in a randomized evaluation of commitment savings accounts in the Philippines, J-PAL affiliated researchers Nava Ashraf (London School of Economics), Dean Karlan (Northwestern University), and Wes Yin (UCLA) specified women's influence in household spending decisions as their main outcome of interest. They collected data on two indicators: 1) an index of women and men's responses about who decides in nine common household spending decisions and, 2) household expenditures on what respondents identified as typically "male" or "female" goods.[53] Using indices or outcome families can also help prevent us from cherry picking a single indicator that shows a significant impact by chance during analysis. However, indices have costs too. They force us to give relative weights to our various indicators. In the above example, using an index required researchers to put an implicit weight on women's influence in decisions about spending at the market and spending on children's schooling. Another great way to triangulate is to ask different people the same question and compare their answers.[54]

[52] Bandiera, Oriana, Robin Burgess, Markus Goldstein, Niklas Buehren, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa." World Bank Group Working Paper, July 2017. http://documents.worldbank.org/curated/en/707081502348725124/Women-s-empowerment-in-action-evidence-from-a-randomized-control-trial-in-Africa.

[53] Ashraf, Nava, Dean Karlan, and Wesley Yin. 2010. "Female Empowerment: Impact of a Commitment Savings Product in the Philippines." World Development 38 (3): 333-344. 336. https://doi.org/10.1016/j.orlddev.2009.05.010.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Household Decision Making and Savings Impacts: Further Evidence from a Commitment Savings Product in the Philippines." Paper Presented at Yale University Economic Growth Center, New Haven, CT, June 2006, 29. https://ideas.repec.org/s/egc/wpaper.html.

[54] For example, see Donald et al. 2017, 39 or Ambler, Kate, Cheryl Doss, Caitlin Kieran, and Simone Passarelli. 2017. "He Says, She Says: Exploring Patterns of Spousal Agreement in Bangladesh." International Food Policy Research Institute Discussion Paper 01616, March 2017. http://ebrary.ifpri.org/cdm/ref/collection/p15738coll2/id/131097.

BOX 6. ASK ABOUT A HYPOTHETICAL SITUATION IN A VIGNETTE

J-PAL affiliated researcher Seema Jayachandran (Northwestern University) and co-authors Diva Dhar (Bill & Melinda Gates Foundation) and Tarun Jain (Indian School of Business) are evaluating the impact of a school-based program in India promoting gender equity on students' attitudes about gender and fertility decisions later in life. To measure attitudes about marriage and women's labor market participation, researchers included a vignette to ask respondents' their opinions on a hypothetical decision:

Pooja, a 21-year-old girl belongs to a village in Haryana. Since childhood, she has aspirations of becoming a police officer. After graduating from college, she appears for the Haryana police examination and is offered a job as a police officer. Her parents are worried about her job as they think that is not suitable for a woman. They also believe that it is her age to get married and they have found a prospective groom for her from a good family. Pooja, however, wants to take up the job and does not wish to get married. According to her parents, Pooja would not need to work after she gets married as her husband will take care of her. Pooja should, instead, focus on household work, help out her mother in law and eventually have children. Finally, her parents decide that instead of taking up the job, she should get married. Do you agree with the parent's decisions?[55]

The vignette was tied to the specific context and designed to resemble a decision that respondents' relatives or family friends may have faced. While the researchers also asked questions about general attitudes (e.g. "should women be allowed to work outside the home?"), the vignette may potentially elicit more honest or nuanced views because of the concreteness and complexity of the setup.

Tip 3. Frame the question indirectly by asking about a hypothetical situation.

If we anticipate that survey respondents might find it hard to answer certain questions about their own lives and families, we can try framing the questions differently. One option could be to pose the question by asking, "for someone similar to you in your community…" Another option is to use a vignette to describe a hypothetical decision or scenario a fictional person is facing and ask about that. See Box 6 for an example of this type of vignette.

CHALLENGE 4: EMPOWERMENT MEANS DIFFERENT THINGS IN DIFFERENT CONTEXTS, BUT WE MAY ALSO WANT TO COMPARE ACROSS CONTEXTS

Tip: Complement context-specific indicators of empowerment with internationally standardized ones.

In general, it can be useful to select commonly used outcomes and indicators to facilitate comparison with other empowerment research. However, indicators that are valid measures of empowerment in the local context are likely to give us a more accurate and precise measure of empowerment than standardized outcomes that are not tailored to our context. For example, women's formal labor market participation is a standard measure of empowerment in many middle- and high-income countries. However, this indicator may not be sensitive to changes in empowerment in rural communities in many low-income countries where there are few formal labor market opportunities for women or men. In these contexts, we may want to ask about informal work. In addition, certain kinds of work may be disempowering depending on the working conditions, so we may also want to consider job quality along with asking women about whether they want to work and how and why they entered their current jobs.

[55] Dhar, Diva, Tarun Jain, and Seema Jayachandran. "Intergenerational Transmission of Gender Attitudes: Evidence from India." NBER Working Paper No. 21429, July 2015.

STEP 2. THEORY OF CHANGE, OUTCOMES, AND INDICATORS: CHALLENGES AND TIPS

CHALLENGE 5: PRIORITIZING OUTCOME MEASURES IS DIFFICULT

Tip: Prioritize indicators that are related to the main objectives of the program, de-prioritize indicators that are not.

One program is not likely to empower a woman in every realm of her life. It is important to be realistic about what a program can change and to prioritize a few main outcomes for which there is a strong logical connection to the program being evaluated according to a well-articulated theory of change. For example, if we are evaluating a program trying to increase women's access to and use of family planning services in Burkina Faso, it may not be relevant to include a series of questions on household consumption, unless household consumption is an integral part of our plan for analysis. If we want to pick up where effects could spill over into other domains, we should use our theory of change to identify the specific areas that might experience change and prioritize measuring those. Beyond prioritizing outcomes, it can help to state them in terms of the specific attitudes, behaviors, and/or achievements we expect to see change. Incorporating too many indicators opens up the risk that some of the indicators will change just by chance, compromising the validity of our findings. Thus, it is critical to be thoughtful about how many metrics are necessary to capture changes in empowerment.

MAIN TAKEAWAYS FOR THEORY OF CHANGE, OUTCOMES AND INDICATORS

- We use the findings from our formative research to create a theory of change and select indicators that are relevant to barriers women and girls face in a particular context.
- To measure agency, we can try to observe choices directly. When this isn't possible, we can ask about the decision-making process, compare gender preferences to changes in outcomes, and/or measure psychological aspects of goal-setting and agency.
- We can measure the process of empowerment by selecting indicators that track the major steps in the causal chain, including short-term, intermediate, and final outcomes, which can correspond to changes in women's resources, agency, and achievements.
- We should think strategically about using locally tailored indicators along with internationally standardized ones.
- It can be useful to measure more than one indicator for outcomes in cases where one indicator doesn't tell the full story or in cases where our indicators are susceptible to reporting bias.
- We can partially overcome the challenge of social desirability bias by asking about a hypothetical scenario or complementing subjective indicators with objective ones or proxy outcomes.

LOCATION: UGANDA. PHOTO: JAVIER MERELO DE BARBERÀ LLOBET | J-PAL/IPA

STEP 3. DATA COLLECTION INSTRUMENTS

Develop and validate data collection in instruments that minimize reporting bias

Once we have selected our outcome measures and indicators, the next step is to select, develop, and pilot our data collection instruments.

COLLECTING DATA WITH SURVEYS

Surveys are the most common method for collecting data in an impact evaluation using quantitative analysis. They are good for collecting a large amount of data from an individual on a range of topics. Relative to many non-survey instruments, they can cover a lot of ground quickly and usually at a lower cost. However, it is important to remember that for some data points, surveys may produce unreliable information. Mayra Buvinic (United Nations Foundation) and Ruth Levine (Hewlett Foundation) have written about how surveys have the potential to minimize the role of women when the questions themselves have a gender bias—for example, when survey instructions guide surveyors towards assuming that a man is the head of household.[56] Another potential concern is that respondents may not be able to accurately recall the past. They may also misreport information on potentially sensitive topics, including gender attitudes, family planning, or how decisions are made at home. For example, in Rachel and coauthors' evaluation on reducing child marriage in Bangladesh, the research team found that since dowry was illegal, we had to approach survey questions on the topic very carefully to get an accurate response. First, we asked, "Often in weddings, gifts are exchanged between the two families. Did you give any gifts to your daughter's in-laws household?" Then, if they said yes, we asked what kinds of gifts they gave. If they mentioned cash, we asked how much.[57]

[56] Buvinic, Mayra, and Ruth Levine. 2016. "Closing the Gender Data Gap." *Significance* 13 (2): 34-37, 35.

[57] Buchmann et al. 2017.

STEP 3. DATA COLLECTION INSTRUMENTS

To help address the kinds of challenges common to surveys, we can complement subjective indicators with objective ones, consider using non-survey instruments, or give respondents options for answering questions privately. We can also ask the question in different ways or ask different people.

SELECTING RESPONDENTS

Before designing our instruments, it is important to determine who our respondents will be. Important considerations include: Who is the target population of the program? Who knows the information we want to collect? Who is unlikely to manipulate the information? Who can we get the largest amount of information from at once? Should we only survey women or girls, or will it be valuable to gather data from other members of the households or communities? Who knows what piece of information is likely to vary by context? For example, in Rachel and co-authors' evaluation on girls' empowerment in Bangladesh, the research team surveyed girls' parents about whether the girls were married, the age of marriage, and age of first birth. Since many of the girls had grown up and left their home communities by the time we conducted the follow-up survey, it was more efficient to collect this information from the parents who remained in the home communities. We also learned that girls typically came back home to visit their mothers for the first births, so mothers had fairly accurate information about whether and when the daughters had given birth.

LOCATION: INDIA. PHOTO: THOMAS CHUPEIN | J-PAL

Collecting information from both spouses in a household can be useful for examining bargaining power related to household spending decisions, family planning, and investment in children's education and health. Interviewing men and women separately can also reveal how husbands and wives' preferences differ. For example, J-PAL affiliated researchers Mushfiq Mobarak (Yale University) and Grant Miller (Stanford University) measured the demand for nontraditional cookstoves in Bangladesh. In the evaluation, a team of two enumerators visited each household and developed a way to measure household decision-making. While one enumerator interviewed the husband, the other conducted a separate interview with the wife. After completing the survey, either the husband or wife was given the opportunity to order a stove that directed smoke away from the cook, but was not able to consult his or her spouse before making the decision. They found that women had much higher demand for health-protecting cookstoves, but that they lacked authority to make purchases.[58]

Asking more than one person the same question can serve as a check to the validity of the main respondents' answers—for example, by asking both parents and young women about why they are missing school. In addition, it can be a useful way to identify systematic differences in perceptions. For example, Aletheia Donald, Gayatri Koolwal, and Markus Goldstein (World Bank) along with Jeannie Annan and Kathryn Falb (International Rescue Committee) found that husbands' and wives' responses to DHS survey questions about who makes decisions about

[58] Miller, Grant, and A. Mushfiq Mobarak. "Gender Differences in Preferences, Intra-household Externalities, and Low Demand for Improved Cookstoves." NBER Working Paper No. 18964, April 2013. http://www.nber.org/papers/w18964.pdf.

large purchases and use of the husband's earnings differ systematically within the same household across many countries.[59] Asking husbands about their wives' preferences and seeing how closely they match her own (and vice versa) can provide a measure of the flow of information between spouses. In Rachel and co-authors' evaluation in Bangladesh, we used questions about how well husbands understood their wives' preferences as a signal of better marriages.

TURNING INDICATORS INTO GOOD SURVEY QUESTIONS

Well-constructed survey questions can go a long way in mitigating measurement error and reporting bias. Good survey questions are easy to answer, pick up variation, minimize the risk of social desirability bias, have realistic recall periods, and tailored to the local context. They should also be specific, neutral, understandable, clearly framed and relevant.

Specific. Each question should only ask one thing at a time. For example, we may not want to ask a young woman, "at what age do you want to get married and have a child?" Instead, we should ask two separate questions, one about desired age of marriage and the other about desired age for having a child.

Neutral. The wording of the question should not bias respondents to give us a particular answer one way or another. For example, we would never ask the following question because it could lead to biased answers, "don't you think that women can be good village council leaders?" We could instead ask, "who is your village council leader?" and then follow up with, "how effective a leader do you think he/she is?" and see how perceptions of effectiveness correlate with the leader's gender.

Understandable. All survey questions should be relatively easy to comprehend by anyone in our sample. They should not contain unfamiliar terms or concepts that are not clearly defined in the survey prompt. Generally speaking, we should avoid abstract concepts. For example, in many cases, women may not be very familiar with the term "empowerment." It could therefore be risky to ask a question like, "do you feel empowered to make decisions about how much to spend on your children's schooling in your household?" Instead we could ask, "who decides how much to spend on your children's schooling in your household?" We may find that we never use the term empowerment once in our survey measuring empowerment.

Clearly framed. The best questions have clear boundaries and a well-defined time frame. If the responses are multiple choice, the list of possible answers should be mutually exclusive—no two responses should overlap in meaning in a way that could confuse respondents. They should also be collectively exhaustive—no frequent answers should be missing from the list. For example, we may find that women respond "no" to the question, "have you visited the doctor in the past month?" but report that they visited other kinds of health care providers that we are interested in documenting. Asking "have you visited the doctor, a nurse, health clinic, or healer in the past month?" could capture a fuller picture of the health services women use. Second, any question asking a respondent to recall something from the past should include a clear time frame and this recall period should be short enough that the respondent can remember accurately. For example, instead of asking, "how many times did you take your child to the doctor in the past year," we could ask, "have you taken your child to the doctor in the past month?" If she responds yes, we can then ask, "how many times have you taken your child to the doctor in the past month?"

[59] Donald et al. 2017, 39.

Relevant. It is important to be respectful of study participants' time, so all of our survey questions should measure indicators that will actually be used in our analysis. Several studies have shown that the accuracy people's answers can be affected by survey fatigue, which we can avoid by being judicious in the number of questions we include.[60] Our questions should also be relevant to the local context. Say we want to examine women's access to technology. We might ask about internet use in some contexts but ask about cell phone ownership and usage in others depending on the availability of services. We can sort out these details during the formative research phase through semi-structured interviews, participant observation, and spending time in the communities. We can also double-check this information when pretesting our survey instruments.

The information we gather during our formative research can also help us write better questions. It can help us select the right words from the local language(s) to use, identify sensitive words or topics to avoid, and figure out the most appropriate way to ask about a particular topic.

LOCATION: COLOMBIA. PHOTO: PAUL SMITH | J-PAL/IPA

Appendix 1 features a catalogue of examples of survey questions and modules related to women's and girls' empowerment from completed impact evaluations by J-PAL affiliated researchers. Another useful resource is the University of California San Diego's Evidence-Based Measures of Empowerment for Research on Gender Equality (EMERGE) website, which compiles survey questions that have been used to measure gender equality and empowerment along with guidelines on developing metrics.[61]

OTHER WAYS TO MITIGATE REPORTING BIAS

In addition to writing good questions, we can design the survey to make respondents feel more comfortable and therefore perhaps more willing to answer accurately. It is ideal to start the questionnaire with simpler questions, such as basic demographic information. Once the enumerator and respondent have had the opportunity to build a rapport, we can begin to ask questions about more sensitive topics. For highly sensitive or personal topics, it is better to have women enumerators interview women and men enumerators interview men. In some cases, we can ask respondents to enter their answers to sensitive questions privately and anonymously into a computer, phone, or tablet device, so that fear of judgment by the enumerator has less of an influence on their responses. We can also mitigate reporting bias by comparing self-reported answers to data from other sources. If we are interested in learning whether a woman redeemed a voucher for a family planning consultation at a local health clinic in the past month, for example, we can ask her directly and also check the administrative data from the clinic.

---

[60] Krosnick, Jon A., and Stanley Presser. 2010. "Question and Questionnaire Design." *In Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright, 263-315. Bingley: Emerald Group Publishing, 291-292.

[61] The University of California San Diego's Evidence-Based Measures of Empowerment for Research on Gender Equality (EMERGE) website is another useful resource. The website compiles survey questions that have been used to measure gender equality and empowerment along with guidelines for selecting metrics.

Evidence-Based Measures of Empowerment for Research on Gender Equality (EMERGE) University of California San Diego. 2018. Accessed June 19, 2018. Emerge.ucsd.edu.

LOCATION: INDIA. PHOTO: JASLEEN KAUR | J-PAL

QUANTITATIVE NON-SURVEY INSTRUMENTS

Non-survey instruments that allow for quantitative analysis, including direct observation, games, experimental vignettes, implicit association tests, and more, can offer more objective measures of some outcomes than surveys.[62] They can also be helpful for quantifying things that are difficult to measure. This includes attitudes or activities that participants may not report honestly in a survey (e.g., prejudice, attitudes about gender norms), are highly subject to recall error (e.g., how many times women spoke in a community meeting), or things about themselves that respondents may not even be aware of (e.g., subconscious gender bias). Non-survey instruments are more complicated for enumerators to administer than surveys and tend to cost more. See Box 7 for two examples of non-survey instruments used to measure an indicator related to empowerment. Appendix 2 contains a catalogue with examples of non-survey instruments that could be useful in evaluations about empowerment, along with their pros and cons, and tips on when to use them. It covers direct observation and structured community activities, implicit association tests, vignettes, list randomization, purchase decisions, games, social interaction and network effects, participatory research methods, and biomarkers.

[62] Some of these instruments, such as vignettes and list randomization, can be implemented in a survey context. We include them in this section because they require different processes and analyses than standard survey questions.

BOX 7. EXAMPLES OF TWO NON-SURVEY INSTRUMENTS

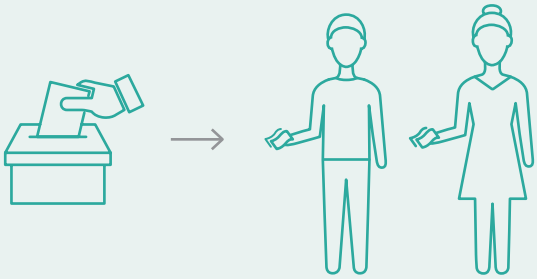
Games: Having study participants play a game can help measure qualities like altruism, cooperation, and trust. They are useful when we want to test theories of how people will respond to various incentives and scenarios, or categorize people into different groups based on their behavior in the game. However, games are only approximations of decisions in the real world, and they may not reflect what people would do in scenarios with bigger stakes or when they are not being observed.

In a randomized evaluation in Kenya, J-PAL affiliate Simone Schaner (Dartmouth College) examined how differences in spending preferences between husbands and wives affected demand for savings accounts.[63] The research used a game to measure intra-household bargaining power to test whether the savings account programs had different impacts on women with more or less bargaining power. At the end of the survey, husbands and wives, who were being surveyed separately, were asked to divide a small cash prize between themselves and their spouse. Each spouse recorded his or her allocation separately on cards and placed the amount they allocated to themselves in their tin and the amount allocated to the spouse in the spouse's tin. Then they came together to decide how to allocate the prize money between them and recorded it on cards, which they added to each of their tins. To ensure respondents' privacy, the husband and wife also each added an additional envelope to each of their tins with a randomly selected amount of money. The husband and wife then chose one card from each of their tins and were immediately given the cash amount allocated to them on the card they chose. This game allowed researchers to identify women with relatively low- or high-bargaining power and test whether the impact of the intervention was different for women with higher or lower bargaining power. Women with larger differences between their individual and joint preferences for allocating the money were classified as having relatively lower bargaining power.

I have with me Ksh 700 for you to divide. You can keep any amount from Ksh 50 to Ksh 650 for yourself, and send the balance to your spouse. What you decide for yourself, I will put in the tin, hidden in this envelope. I will put the amount you decide for your spouse...into your spouse's tin.

We will ask you and your spouse to come to an agreement over how to divide the Ksh 700. After you make a decision, I will put the amount intended for you into your tin and... the amount for intended for your spouse into your spouse's tin. I [also] have here a bag with all the possible money amounts, ranging from Ksh 50 to Ksh 650. I will ask you to select one of the envelopes in this bag and put it into your tin. This is the "secret-keeping" choice.

Once you have added this final amount, you will be asked to pick one of the envelopes from the tin and whichever envelope you draw, you will receive that amount today in cash.



[63] Schaner, Simone. 2017. "The Cost of Convenience? Transaction Costs, Bargaining Power, and Savings Account Use in Kenya." *Journal of Human Resources* 52 (4): 919-945. https://doi.org/10.3368/jhr.52.4.0815-7350R1.

Schaner, Simone. "Bargaining Game Script: The Cost of Convenience? Transaction Costs, Bargaining Power, and Savings Account Use in Kenya." Research. Accessed June 8, 2018. https://sites.google.com/site/sschaner/home/research.

BOX 7. EXAMPLES OF TWO NON-SURVEY INSTRUMENTS, CONTINUED

Experimental Vignettes: Vignettes are brief descriptions of hypothetical scenarios. They can be useful in many different aspects of research, from reducing the risk of social desirability bias in survey questions about sensitive topics to clarifying the meaning of concepts in our survey questions. Experimental vignettes—in which study participants are randomly assigned to hear one of multiple versions of the same story with a key detail changed—can also be used to measure subconscious biases or prejudices based on gender, race, ethnicity, or other factors. To measure gender bias, for example, we could create two identical versions of a vignette in which only the gender of the subject changes. We would then randomly assign study participants to hear one of the two versions and ask them the same follow-up questions about the vignette, allowing us to isolate the difference in responses caused solely by the gender of the subject in the hypothetical scenario. Typically, only one detail in the vignette can be changed to isolate the source of bias.

J-PAL affiliated researchers Lori Beaman (Northwestern University), Raghabendra Chattopadhyay (Indian Institute of Management), Esther Duflo (MIT), and Rohini Pande (Harvard University) along with Petia Topalova (International Monetary Fund) evaluated whether exposure to female leaders in Indian village councils changed perceptions about women's effectiveness as leaders.[64] As part of the survey, researchers played a short recording of a speech by a local leader responding to a complaint from a villager. Respondents were randomly assigned to hear the recording spoken by a man or woman. After the speech was over, they were asked to rate the leader's effectiveness. This vignette allowed researchers to measure whether there was a subconscious bias that led people to rate female leaders as relatively less effective. They found that exposure to a female leader through the policy that reserved village council head positions for women reduced men's bias against female leaders.

1. A short recording of a speech by a local leader responding to a complaint from a villager is played. Respondents heard the recording spoken by either a man or woman.

2. After the speech, respondents were asked to rate the effectiveness of the leader.

[64] Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?." *The Quarterly Journal of Economics* 124, no. 4: 1497-1540.

INCORPORATING QUALITATIVE RESEARCH

Research teams often conduct semi-structured interviews and focus groups with small subsets of the larger evaluation sample to gain a more in-depth understanding of a particular phenomenon or generate hypotheses about why we observe certain behaviors and how and why a program may have changed behavior. This does not mean asking "How did this program change your life?". Such questions suffer from social desirability bias; but more importantly, people can find it hard to construct their own counterfactual of what their lives would have been like without the program. Instead, qualitative and quantitative researchers working together can use these interviews to form or investigate hypotheses and mechanisms. For example, in Rachel and co-authors' evaluation in Bangladesh, the team of researchers is made up of several economists and a qualitative researcher Shahana Nazneen. We conducted in-depth qualitative interviews with sub samples of young women, their parents, and local matchmakers. This helped us better understand how decisions about young women's marriage are made and why a girls' empowerment program did not affect child marriage rates but an in-kind transfer to unmarried adolescents up to age eighteen did.

Our quantitative survey showed that a woman's decision on when and whom to marry was nearly always made by her parents, and the empowerment program did not change that. The theory of change behind the empowerment program was that either parents did not know the dangers of child marriage and would be informed by their daughters, or that parents were not taking their daughters' interests into account and that daughters would be better able to bargain with their parents after the empowerment program. However, our qualitative interviews showed that parents did know the dangers and agonized over the decision. Many reported not wanting to marry their daughters as young as they did, but felt they were doing what was best for their daughters. Marriage offers came infrequently and parents reported taking offers because they worried they might never again receive an equally good offer with a dowry they could afford, as dowries increase substantially as young women get older.[65]

[65] Field et al. 2018, 31.

They also reported community pressure, with family and neighbors asking, "why isn't your daughter married yet?" Together, these discussions helped us think about marriage decisions differently—as a search process with asymmetric information and signaling. Maybe dowries increased with age because grooms were concerned that there was a (negative) reason why some older girls were not married yet. Maybe the community stopped asking "why is your daughter not married yet?" after the in-kind transfer program because there was now a clear financial reason (the in-kind transfers) to delay. Researchers are now testing these hypothesis in the field. These conversations also helped explain why the impact of the program did not end when the program did. Once the in-kind transfers stopped, it would still take some time to find the right match, and the girls did not all get married immediately when they were eighteen and ineligible to continue receiving the transfers.

PRETESTING AND PILOTING LOGISTICS FOR DATA COLLECTION INSTRUMENTS

Once we have designed our instruments, the crucial next step is to validate them with respondents from the communities in which our evaluation will take place, but outside of the actual sample for our evaluation. We will only know whether we are asking the right questions, whether the questions make sense to the people we will be interviewing, and whether they are picking up the right information after we ask people. Researchers can validate their data collection instruments by pretesting their survey questionnaires and non-survey instruments and piloting data collection logistics. Pretesting and piloting responses should not be included in our final data analysis.

1. Pretesting survey questionnaires involves implementing practice interviews or survey prompts with respondents similar to people in our sample. It can help us check how the target population interprets the survey questions. Are the questions easy to understand and answer? Are there any confusing words that need to be changed or questions that don't pick up the information we're looking for? Based on these practice interviews, we can fix the questions or wording in the survey and test the new version with a new practice respondent. The goal is for enumerators to be able to conduct practice interviews in which respondents

can understand and answer all the questions well. There is no right number of pretests to conduct; we may find that fifteen are enough, or we may need to do up to fifty. It can be useful for several enumerators to participate in pretesting the survey, as differences in enumerator style can affect responses, and different enumerators will notice different problems with the questionnaire. If we are using non-survey instruments during or separate from our survey, it is critical to pilot them the same way we pilot surveys. Some additional pretesting techniques include:[66]

- **Cognitive interviews:** interviewees respond to the enumerator by describing their ongoing ideas and thought processes in addition to answering the survey question. This can help researchers learn about whether and how people understand their survey questions and identify the roots of any confusion or difficulties people have in answering them.[67]
- **Respondent debriefings:** researchers implement the survey and collect comments and feedback from respondents after they complete it. This can be easily and cost-effectively incorporated into pretesting.
- **Expert reviews:** experienced specialists in a particular topic or sector provide comments on the content or style of the survey.

2. Piloting logistics includes test runs of the entire data collection process, enabling us to identify possible sources of measurement error. Enumerators practice executing the full survey from start to finish, including recruiting respondents, conducting surveys, running non-survey instruments, digital data entry, and data quality checks. It is important to test the logistics for complicated or unpredictable data collection processes, such as recruiting survey respondents at events, or for testing technology in the field. Piloting logistics helps ensure that there are sufficient resources to implement the survey, that enumerators fit the job description well, and that the administrative processes work. Researchers can also incorporate pretesting techniques into the piloting. Overall, logistics piloting can be costly and time intensive, involving dozens of respondents. As such, if we are familiar and comfortable with the enumerators and administrative systems in a given context, it might not be necessary to do a full-scale piloting process.

While the time needed to validate the survey instruments varies, it is important to budget at least a few weeks before the actual survey is scheduled to begin to update and finalize the survey instruments.

LOCATION: BANGLADESH. PHOTO: ISHITA AHMED | J-PAL/IPA

[66] For practical insights on pretesting and piloting logistics, see:

Caspar, Rachel, Emilia Peytcheva, Ting Yan, Sunghee Lee, Mingnan Liu, and Mengyao Hu. 2016. "Pretesting." *In Cross-Cultural Survey Guidelines.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Accessed March 19, 2018. http://ccsg.isr.umich.edu/index.php/chapters/pretesting-chapter#one.

Ruel, Erin E., William E. Wagner, III, and Brian J. Gillespie. 2015. *The Practice of Survey Research: Theory and Applications.* Los Angeles: SAGE.

[67] For an example of how cognitive interviewing helped strengthen the wording of survey questions, see: Malapit, Hazel J., Kathryn Sproule, and Chiara Kovarik. 2017. "Using Cognitive Interviewing to Improve the Women's Empowerment in Agriculture Index Survey Instruments: Evidence from Bangladesh and Uganda." *Agri-Gender: Journal of Gender, Agriculture, and Food Security* 2 (2): 1-22. http://agrigender.net/views/cognitive-interviewing-to-improve-women-empowerment-from-bangladesh-and-uganda-JGAFS-222017-1.php.

WHAT ARE WE LOOKING FOR IN PRETESTS AND PILOTING LOGISTICS?

Have we chosen the right people to interview? In the course of piloting we may discover that the people we are surveying do not have the information we want. An adolescent girl may not know how much her school fees are. Her parents and teachers may be more likely to know. If women and men are in charge of planting different crops on a farm, women may only have information about planting decisions for the traditionally female crops and men for the traditionally male crops. If we observe a clear pattern that several respondents do not know the information we are asking them about, we should consider other potential data sources for these questions. In the first example that could mean asking parents, and in the second example it could mean surveying men in addition to women. We must carefully weigh the benefits of receiving more accurate data on a given indicator against the increased cost of adding a new survey or group of respondents.

Have we worded our questions well, and do people understand them? Respondents may express confusion about what a question is asking, or they may respond with an answer that shows they did not understand the question in the same way we did. See Box 8 for one example of a survey question that needed clarification. We should update our survey as frequently as needed during piloting to make sure all questions are well understood.

BOX 8. WORKING TOWARDS PRECISE AND LOCALLY TAILORED SURVEY QUESTIONS

In an evaluation Rachel and co-authors conducted in Sierra Leone, an early version of our survey contained the question, "do you belong to any social groups?" and many pilot participants replied, "no." After asking follow-up questions, our enumerators came to find that people did belong to what we considered social groups—mosques, churches, farming or business cooperatives—they just did not label them "social groups" like we did. We amended the question to prompt people with a list of the most common groups we found people belonged to in our pilot surveys.

Do our instruments pick up variation? Our surveys and other data collection instruments must be sensitive enough to pick up the variation found in the population where our research is taking place. For example, if we are trying to measure women's knowledge about their civil and political rights, the questions we use to assess their knowledge should not be so easy that most respondents get all of the questions right nor so hard that very few get any right. If we anticipate that answers to some questions may be similar, we should pretest the questions first to verify that this is the case. Pretesting can help us find the right combination of questions to get a range of different responses.

Is the plan appropriate given the culture and politics of the local context? Both our survey questions and data collection plan should be tailored to the local context. Say we want to examine whether women have opportunities for collective action in their community. We need to ask questions about types of collective action that make sense and actually occur in the specific context. Women may participate in rallies, protests, and issue-based political campaigns in some contexts, but in others these types of collective action may be rare or inaccessible to women so they would not be relevant to include as survey responses. We can learn these details during the formative research phase through semi-structured interviews, participant observation, and spending time in communities similar to where the evaluation will take place. We can also double check this information when pretesting our survey instruments.

During pretesting, we may find that the way we plan to conduct our survey may not be appropriate given the culture and politics of the area. For example, it may be easier to interview young women at school rather than at home to reduce survey costs. Yet parents and local leaders may not think it is appropriate for enumerators to survey girls alone in a place outside their homes. In this case, surveying girls at home would be more appropriate, though more expensive. We often cannot anticipate all of the community's preferences, so it is important to consult local leaders during the formative research phase and pilot our survey processes before starting the official data collection.

Is our survey too long? People can become fatigued during long surveys and their answers may become less accurate as time passes. Pretesting can help us determine whether our survey is too long and identify questions that could be cut. We may also consider placing the bulk of questions that require more mental energy earlier on in the survey or even conduct the data collection over multiple visits. Pretesting will give us a good estimate of how long the questionnaire takes, which we include in our script asking for respondents' consent to participate in the study prior to beginning the exercise. We should note that as enumerators get used to implementing the survey, the process typically speeds up by about one-third.

Are the recall periods appropriate? It is nearly impossible to assess whether our recall periods are appropriate without posing our survey questions to people. If we find during pretesting that the question is too challenging for most respondents to answer precisely, we may need to reformulate it. During field testing, we can even test two recall period variations to see which one gives us the best information.

MAIN TAKEAWAYS FOR DATA COLLECTION INSTRUMENTS

- Most quantitative impact evaluations will use surveys as the main data collection instrument. There is also a wide array of non-survey instruments that generate quantitative data that researchers can use for outcomes that are hard to measure or that people may not speak about accurately or honestly (i.e., prejudice or gender bias).
- Incorporating qualitative interviews with a subset of our sample can also help us better understand particular phenomena and/or generate hypotheses about why a program did or did not work.
- Before we write our survey questions, we need to pick the right people to interview. In many cases it is useful to survey multiple sources (e.g., girls and their parents; both men and women).
- When writing survey questions, it is important to make them specific, neutral, understandable, clearly framed, and relevant.
- Pretesting our survey questionnaire with people similar to our target population is critical for ensuring questions make sense and pick up the information we are interested in. Cognitive interviewing and respondent debriefings can also be useful in these efforts.
- We need to make sure the sequencing, length, and recall periods work; that the questions are well understood and culturally appropriate; and that they pick up variation.
- We should repeatedly pretest and update the questionnaire until all questions are well understood and generate the information we want.
- If we have less experience collecting data in the context where the evaluation will take place, we should also pilot the survey logistics, including everything from finding and enrolling respondents, administering the questionnaire, entering data, and doing data quality checks.

LOCATION: COLOMBIA. PHOTO: PAUL SMITH | J-PAL/IPA

STEP 4. DATA COLLECTION PLAN

Design a data collection plan that minimizes measurement error

The last step before officially launching the evaluation is to make sure our data collection process itself does not introduce significant opportunities for measurement error. The way we collect data can affect its accuracy. We have to think carefully about how often, when, and where we will collect data, in addition to who we select as enumerators. All of these factors have a gender component that should be considered and included in the data collection strategy. If we're conducting a randomized evaluation, it is also imperative that all aspects of data collection—the enumerators, questionnaires, the timing and frequency of surveys—be identical for the treatment group and the comparison group.[68]

WHO COLLECTS THE DATA?

The identity of the enumerator can affect the answers people give. Study participants may feel less free to speak openly about certain topics with enumerators of a particular gender, class, and/or ethnicity. As a general rule of thumb, enumerators should be from the country where the research is taking place and, when possible, from a similar region of the country. Enumerators should also be fluent in the languages spoken in the areas where they are conducting surveys, including local languages. Pretesting is a good opportunity to investigate whether the identity of enumerators affects respondents' answers.

In general, when we are asking people about topics related to women or girls' empowerment, it is best if the enumerator is the same gender as the respondent. This is particularly true in contexts that have social norms regulating the interaction between men and women. For example, in the evaluation that Rachel and co-authors conducted in Bangladesh, the young women participating were always interviewed by a woman because it was socially inappropriate for a woman and a man from outside

[68] Data collection differences between the two groups could lead to non-classical measurement error, meaning error that it is correlated with treatment status, which leads to biased impact estimates. The good news is that measurement error sources are less troublesome in randomized evaluations if we can reasonably assume that they happened at approximately the same rates in the treatment and comparison groups. For example, if people consistently misunderstand how certain questions were phrased across both groups, our impact estimate won't be biased.

her family to interact without a chaperone. If it is not possible to match respondents with enumerators of the same gender, we can collect and record data on enumerator gender and test whether answers vary systematically by enumerator gender in our analysis. Gender is not the only aspect of identity that can affect data collection. For example, a young woman from a low-income family in a rural community may feel uncomfortable speaking with an enumerator from a middle- or high-income background from a major city.

Nonetheless, some respondents would not feel free to speak honestly about certain topics, such as intimate partner violence or sexual assault, with any enumerator. In contexts where respondents are literate, one alternative to provide increased privacy for survey questions on particularly sensitive subjects could be to offer respondents the ability to answer directly on a tablet, phone, or computer. Audio Computer-Assisted Self-Interview (ACASI) technology involves listening to pre-recorded interview questions on headphones and answering the questions using a digital device and can be adapted to contexts with low literacy.

Because enumerator identity matters, using the same enumerators in both treatment and comparison groups is key. Asking existing program staff to collect data in the treatment group and hiring enumerators or a survey firm to conduct surveys in the comparison group can seem like a good way to save on research costs. Yet, respondents may feel less free to answer questions honestly when being interviewed by a staff member they recognize from an organization that has provided them with a program.[69] Enumerators who are skilled in administering surveys and who know how to avoid asking leading questions are more likely to collect high-quality data than individuals who do not have experience conducting research.

DO WE ALWAYS NEED TO INTERVIEW RESPONDENTS ALONE?

Interviewing someone in the presence of another person can bias his or her answers. To eliminate this potential bias, our best option is to interview each respondent alone. When setting up the interviews, researchers should let respondents know that the interview should take place privately so that they can plan accordingly. If there is not enough space to interview respondents alone in their living quarters, we may need to identify nearby locations where we can use a more private space.

Yet, interviewing respondents alone can be challenging when working with populations that lack power, particularly if family members feel that it is inappropriate. Say we are interested in interviewing young women about their aspirations related to their careers, getting married, and having children, but during piloting we find that some parents express concern about strangers speaking with their daughters alone. One option is to offer to interview the young woman within sight—but not earshot—of her parents, such as on the porch or just outside the house. A well-written parental consent script can also help mitigate some concerns parents have about letting their children participate. The script should thoroughly explain the purpose of the research, the organization the enumerator represents, the complete anonymity and privacy of their daughter's responses, and give parents the opportunity to ask follow-up questions before giving consent.

Women with young children may often need to look after them during the interview. If there is no one else who can look after her young children during the questionnaire, the enumerator can proceed with the survey but make note of who else was present. The questionnaire should include a prompt asking the enumerator to record whether the respondent was alone and, if they were not alone, who else was present and a short explanation of why it was not possible to interview the respondent alone. We can check these notes regularly (on a daily or weekly basis) to develop new strategies that enumerators can try to overcome this and other challenges in future surveys.

[69] This could lead to non-classical measurement error—measurement error that is systematically different between the treatment and comparison groups—which will bias our impact estimates.

STEP 4. DATA COLLECTION PLAN

WHEN AND WHERE DO WE INTERVIEW PEOPLE?

There are certain times when respondents may be more or less available to participate in a survey. Pretesting and logistics piloting helps us document how many people we were able to reach at various times of day and whether there were any days of the week when people were less available. We can use this information to create a survey timetable that fits respondents' needs and availability.

Study participants may be more available during certain times of year. In contexts where many secondary school students study in boarding schools, it may be easier to interview adolescent girls at home during their school holidays. Farmers may be less available once planting season or harvest has started. Since many young women in Bangladesh migrate to cities to work in factories, we found that it was easier to interview them during a month with many holidays when many young women travel home to visit their families.

LOCATION: RWANDA. PHOTO: TOM GILKS | J-PAL/IPA

The time of year can also affect how individuals respond to our surveys. For example, men and women in agricultural households may have different workloads during the planting season compared to the harvest season, meaning that a time use survey could generate very different data if we conduct it during planting season versus harvest season.

Where we interview respondents can also substantially affect the cost of data collection. Interviewing students at school will be much less expensive than interviewing each of them at home. Similarly, interviewing women at the monthly savings group meeting place will cost less than a household survey. We may need to follow up individually with the respondents we cannot reach in these central locations, but this will still be less expensive than individually surveying each respondent at home.

Survey location is also important to pretest. For example, we may think it will be more efficient to interview women entrepreneurs in an outdoor market during the day than to interview them at home, but find that many women cannot stop working in order to answer questions. In this instance, we could consider conducting as many surveys in the market as possible and follow up with the women who could not participate in the survey at a less busy time, a different day of the week, or at home. If we decide to conduct interviews in public places rather than individuals' homes, we should be sure to collect enough personal and contact information to be able to follow up with respondents later in subsequent survey rounds.

WHEN SHOULD WE START AND END DATA COLLECTION AND HOW FREQUENTLY SHOULD WE COLLECT DATA?

A program's effects are usually not instantaneous. We should begin to measure outcomes after the program has had an adequate amount of time to achieve its effects. It can be difficult to predict exactly how long this takes, particularly for complex processes like empowerment. Our theory of change can help us think through how long it will take for the program to start to have an impact. Consider a business training program for women entrepreneurs that aims first to improve their business management knowledge, then improve their business management practices, and as a result, increase their profits. We may expect knowledge

and business practices to improve within the first couple of months after the program ends. Yet, it can take some time for business practice changes to result in changes in profits, so we want to be careful not to measure our final outcomes too early. Measuring outputs and intermediate outcomes along the way can also help us check whether our timeline for measuring final outcomes is reasonable. If the program is taking longer to affect our intermediate outcomes than originally expected, we could consider pushing back the start date of our endline survey.

Several factors can influence the timing for the final round of data collection, including cost, attrition, whether longer-term results are needed, and whether the results are needed by a certain date to make a particular decision. In Rachel and coauthors' evaluation in Bangladesh, the research team was primarily interested in learning whether various programs helped young women delay marriage into adulthood, so we needed to wait several years until the girls began to get married to understand this impact. Research costs and the likelihood of attrition increase as the follow-up period gets longer. A new program could also have novelty effects, where people are more excited about a program when it first starts, automatically raising outcomes in the short term. If this seems likely, we should not end data collection within the first few months after the program starts. Additionally, if we expect our program to have a lasting effect on people's lives, we should not end data collection sooner than six months or a year after the program ends.

What about data collection frequency? Collecting data more frequently increases costs, but it can also help us better capture intermediate outcomes and observe changes over time. If the impact of the program is likely to change (or even decay) over time, and if this information will be used to make a decision about investing more in a program, we may want to collect data more frequently and over longer periods.

MAIN TAKEAWAYS FOR DATA COLLECTION PLAN

- It is not just our survey content that affects the accuracy of our data—who conducts the survey, when, and how often, also matter.
- Enumerator identity can bias people's responses, so we need to ensure that respondents are comfortable with their enumerators. It is often most effective to have women interview women and men interview men.
- Find a reliable but culturally appropriate way to interview respondents alone.
- Select a time that is convenient for respondents but that also allows us to maximize the number of surveys we can complete.
- Collect data following a timeline that allows for outcomes to materialize according to the program's theory of change.
- Use the same data collection processes and enumerators in treatment and comparison groups to avoid measurement error that is systematically different across treatment and comparison groups.

LOCATION: INDIA. PHOTO: JASLEEN KAUR | J-PAL

CONCLUSION

Empowerment is a complex process, and measuring it can be challenging. Yet, collecting good data on where women lack power to make choices that are important to them and using impact evaluations to identify effective ways to help women gain greater agency are critical inputs for reducing gender inequality worldwide.

This guide has just scratched the surface of the many creative strategies researchers and practitioners from diverse disciplines have developed to help us better understand, measure, and track changes in women's and girls' empowerment. Building on this large body of work, we hope we have shed light on how to weigh the trade offs between using different types of measurement tools and outlined a process for developing a measurement strategy that is right for the evaluation at hand. While there are concrete ways to address many of the common challenges of measuring empowerment, there is still a great need for more systematic measurement research and validation exercises to improve on our current approaches. By continuing to experiment with, refine, and improve how we measure empowerment, researchers and practitioners can generate evidence on approaches that can help women and girls realize their own vision of a better life.

REFERENCES

Abdul Latif Jameel Poverty Action Lab (J-PAL). N.d. "Ethics." Research Resources. Accessed March 19, 2018. https://www.povertyactionlab.org/research-resources/ethics.

Abdul Latif Jameel Poverty Action Lab (J-PAL). N.d. "Measurement & Data Collection." Research Resources. Accessed March 19, 2018. https://www.povertyactionlab.org/research-resources/measurement-and-data-collection.

Ager, Alastair, Lindsay Stark, and Alina Potts. 2010. "Participative Ranking Methodology: A Brief Guide: Version 1.1." New York, NY: Program on Forced Migration & Health, Mailman School of Public Health, Columbia University. https://doi.org/10.13140/RG.2.2.34356.45448.

Alkire, Sabina. 2005. "Subjective Quantitative Studies of Human Agency." *Social Indicators Research* 74 (1): 217-260. https://doi.org/10.1007/s11205-005-6525-0.

Alkire, Sabina, Ruth Meinzen-Dick, Amber Peterman, Agnes Quisumbing, Greg Seymour, and Ana Vaz. 2013. "The Women's Empowerment in Agriculture Index." World Development 52: 71-91. https://doi.org/10.1016/j.worlddev.2013.06.007.

Almås, Ingvild, Alex Armand, Orazio Attanasio, and Pedro Carneiro. "Measuring and Changing Control: Women's Empowerment and Targeted Transfers." NBER Working Paper No. w21717, 2015. http://www.nber.org/papers/w21717.pdf.

Ambler, Kate, Cheryl Doss, Caitlin Kieran, and Simone Passarelli. 2017. "He Says, She Says: Exploring Patterns of Spousal Agreement in Bangladesh." International Food Policy Research Institute Discussion Paper 01616, March 2017. http://ebrary.ifpri.org/cdm/ref/collection/p15738coll2/id/131097.

Angelucci, Manuela, Dean Karlan, and Jonathan Zinman. 2015. "Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco." *American Economic Journal: Applied Economics* 7 (1): 151-82. 174. http://dx.doi.org/10.1257/app.20130537.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Household Decision Making and Savings Impacts: Further Evidence from a Commitment Savings Product in the Philippines." Paper Presented at Yale University Economic Growth Center, New Haven, CT, June 2006. 29. https://ideas.repec.org/s/egc/wpaper.html.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2010. "Female Empowerment: Impact of a Commitment Savings Product in the Philippines." *World Development* 38 (3): 333-344. https://doi.org/10.1016/j.worlddev.2009.05.010.

Bandiera, Oriana, Robin Burgess, Markus Goldstein, Niklas Buehren, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. "Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa." World Bank Group Working Paper, July 2017. http://documents.worldbank.org/curated/en/707081502348725124/Women-s-empowerment-in-action-evidence-from-a-randomized-control-trial-in-Africa.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias?." *The Quarterly Journal of Economics* 124 (4): 1497-1540. https://doi.org/10.1162/qjec.2009.124.4.1497.

Bryson, John M. and Michael Quinn Patton. 2010. "Analyzing and Engaging Stakeholders." *Handbook of Practical Program Evaluation* Fourth Edition: 36-61. https://experts.umn.edu/en/publications/analyzing-and-engaging-stakeholders.

Buchmann, Nina, Erica Field, Rachel Glennerster, Shahana Nazneen, Svetlana Pimkina, and Iman Sen. "Power vs Money: Alternative Approaches to Reducing Child Marriage in Bangladesh, a Randomized Control Trial." Working Paper, May 2017.

Buvinic, Mayra, and Rebecca Furst-Nichols. 2013. "Measuring Women's Economic Empowerment: Companion to a Roadmap for Promoting Women's Economic Empowerment." United Nations Foundation and Exxon Mobil Foundation. http://womeneconroadmap.org/sites/default/files/Measuring%20Womens%20Econ%20Emp_FINAL_06_09_15.pdf.

Buvinic, Mayra and Ruth Levine. 2016. "Closing the Gender Data Gap." *Significance* 13_c12j: (2) 34-37.

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127 (4): 1755-1812. https://doi.org/10.1093/qje/qje027.

Caspar, Rachel, Emilia Peytcheva, Ting Yan, Sunghee Lee, Mingnan Liu, and Mengyao Hu. 2016. "Pretesting." *In Cross-Cultural Survey Guidelines*, Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Accessed March 19, 2018. http://ccsg.isr.umich.edu/index.php/chapters/pretesting-chapter#one.

Chattopadhyay, Raghabendra, and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Policy experiment in India." *Econometrica* 72 (5): 1409-1443. https://doi.org/10.1111/j.1468-0262.2004.00539.x.

REFERENCES

Demographic and Health Surveys Program. N.d. "Women's Status and Empowerment." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/Topics/Womens-Status-And-Empowerment.cfm.

Demographic and Health Surveys Program. N.d. "Family Planning." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/Topics/Family-Planning.cfm.

Demographic and Health Surveys Program. N.d. "The Gender Corner." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/topics/gender-Corner/index.cfm.

Demographic and Health Surveys Program. 2017. "DHS Model Questionnaire – Phase 7." The DHS Program. Accessed March 19, 2018. https://dhsprogram.com/pubs/pdf/DHSQ7/DHS7-Womans-QRE-EN-07Jun2017-DHSQ7.pdf.

Development Media International. 2016. "Sending Our Writer to Live in a Village for a Week." DMI News, May 23, 2016. http://www.developmentmedia.net/news/sending-our-writers-to-live-in-a-village-for-a-week.

Dhar, Diva, Tarun Jain, and Seema Jayachandran. "Intergenerational Transmission of Gender Attitudes: Evidence from India." NBER Working Paper No. 21429, July 2015.

Donald, Aletheia, Gayatri Koolwal, Jeannie Annan, Kathryn Falb, and Markus Goldstein. "Measuring Women's Agency." World Bank Policy Research Working Paper no. 8148, July 2017. http://documents.worldbank.org/curated/en/333481500385677886/Measuring-womens-agency.

Field, Erica, Rachel Glennerster, and Shahana Nazneen. 2018. "Economic Empowerment of Young Women in Bangladesh: Barriers and Strategies." *Women at Work: Addressing the Gaps*. International Policy Centre for Inclusive Growth Policy in Focus 15 (1): 31-32. http://www.ipc-undp.org/publication/28507.

Gallup, International Labor Organization. 2017. "Towards a Better Future for Women and Work: Voices of Women and Men." ILO-Gallup Report, March 8, 2017. 74. http://www.ilo.org/global/publications/books/WCMS_546256/lang--en/index.htm.

García-Moreno, Claudia, Henrica A.F.M. Jansen, Mary Ellsberg, Lori Heise, and Charlotte Watts. 2005. "WHO Multi-Country Study on Women's Health and Domestic Violence against Women: Summary Report of Initial Results on Prevalence, Health Outcomes and Women's Responses." Geneva: World Health Organization.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press.

Glennerster, Rachel, and Claire Walsh. 2017. "Is It Time to Re-Think How We Measure Women's Household Decision-Making Power?" Abdul Latif Jameel Poverty Action Lab, September 7, 2017. https://www.povertyactionlab.org/blog/9-6-17/it-time-rethink-how-we-measure-women's-household-decision-making-power-impact.

Golla, Anne Marie, Anju Malhotra, Priya Nanda, Rekha Mehra, Aslihan Kes, Krista Jacobs, and Sophie Namy. 2011. "Understanding and Measuring Women's Economic Empowerment." International Center for Research on Women. https://www.icrw.org/publications/understanding-and-measuring-womens-economic-empowerment/.

Grüne-Yanoff, Till, and Sven Ove Hansson, eds. 2009. *Preference Change: Approaches from Philosophy, Economics and Psychology*. Dordrecht: Springer Science & Business Media.

Hennink, Monique, Inge Hutter, and Ajay Bailey. 2011. *Qualitative Research Methods*. Thousand Oaks, California: SAGE Publications.

Innovations for Poverty Action. 2018. "The Safe and Ethical Conduct of Violence Research: Guidance for IPA Staff and Researchers." Updated June 2018. https://www.poverty-action.org/publication/ipv-ethical-guidance.

Kabeer, Naila. 1999. "Resources, Agency, Achievements: Reflections on the Measurement of Women's Empowerment" *Development and Change* 30 (3): 435-464. https://doi.org/10.1111/1467-7660.00125.

Kabeer, Naila. "Paid Work, Women's Empowerment and Gender Justice: Critical Pathways of Social Change." Institute of Development Studies, Working Paper, January 2008. https://www.ids.ac.uk/publication/paid-work-women-s-empowerment-and-gender-justice-critical-pathways-of-social-change.

Krosnick, Jon A., and Stanley Presser. 2010. "Question and Questionnaire Design." In *Handbook of Survey Research*, edited by Peter V. Marsden and James D. Wright, 263-315. Bingley: Emerald Group Publishing.

Laszlo, Sonia, and Kate Grantham. "Measurement of Women's Economic Empowerment in GrOW Projects: Inventory and User Guide." McGill University GrOW Working Paper, December 2017.

Levy Economics Institute of Bard College. n.d. "Publications on Time Poverty." Accessed March 19, 2018. http://www.levyinstitute.org/topics/time-poverty.

REFERENCES

Lombardini, Simone, Kimberly Bowman, and Rosa Garwood. 2017. "A 'How-To' Guide To Measuring Women's Empowerment: Sharing Experience from Oxfam's Impact Evaluations." Oxfam. https://policy-practice.oxfam.org.uk/publications/a-how-to-guide-to-measuring-womens-empowerment-sharing experience-from-oxfams-i-620271.

Mack, Natasha, Cynthia Woodsong, Kathleem M. MacQueen, Greg Guest, and Emily Namey. 2005. "Qualitative Research Methods: A Data Collector's Field Guide." Research Triangle Park: Family Health International. https://www.fhi360.org/sites/default/files/media/documents/Qualitative%20Research%20Methods%20-%20A%20Data%20Collector%27s%20Field%20Guide.pdf.

Malapit, Hazel J., Kathryn Sproule, and Chiara Kovarik. 2017. "Using Cognitive Interviewing to Improve the Women's Empowerment in Agriculture Index Survey Instruments: Evidence from Bangladesh and Uganda." *Agri-Gender: Journal of Gender, Agriculture, and Food Security.* 2 (2): 1-22. http://agrigender.net/views/cognitive-interviewing-to-improve-women-empowerment-from-bangladesh-and-uganda-JGAFS-222017-1.php.

Malhotra, Anju, Sidney R. Schuler, and Carol Boender. 2002. "Measuring Women's Empowerment as a Variable in International Development." Background paper prepared for the World Bank Workshop on Poverty and Gender: New Perspectives, June 28, 2002. http://siteresources.worldbank.org/INTGENDER/Resources/MalhotraSchulerBoender.pdf.

Martínez-Restrepo, Susana, and Laura Ramos-Jaimes eds. 2017. *Measuring Women's Economic Empowerment: Critical Lessons from South America*. Springfield, VA: IDRC, Fedesarrollo. http://hdl.handle.net/11445/3482.

Miller, Gran, and A. Mushfiq Mobarak. "Gender Differences in Preferences, Intra-household Externalities, and Low Demand for Improved Cookstoves." NBER Working Paper No. 18964, April 2013. http://www.nber.org/papers/w18964.pdf.

Patton, Michael Quinn. 2015. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice* Fourth Edition. Thousand Oaks, California: SAGE Publications.

Raworth, Kate, Caroline Sweetman, Swati Narayan, Jo Rowlands, and Adrienne Hopkins. 2012. "Conducting Semi-Structured Interviews." Oxfam. https://policy-practice.oxfam.org.uk/publications/conducting-semi-structured-interviews-252993.

Ruel, Erin E., William E. Wagner, III, and Brian J. Gillespie. 2015. *The Practice of Survey Research: Theory and Applications*. Los Angeles: SAGE.

Jose, Rupa, Nandita Bhan, and Anita Raj. 2017. "EMERGE Measurement Guidelines Report 2: How to Create Scientifically Valid Social and Behavioral Measures on Gender Equality and Empowerment." Center on Gender Equity and Health (GEH). San Diego, CA: University of California, San Diego School of Medicine.

Schaner, Simone. 2017. "The Cost of Convenience? Transaction Costs, Bargaining Power, and Savings Account Use in Kenya." *Journal of Human Resources* 52 (4): 919-945. https://doi.org/10.3368/jhr.52.4.0815-7350R1.

Schmeer, Kammi. 2000. "Stakeholder Analysis Guidelines." Policy toolkit for strengthening health sector reform 2: 1-43. http://www.who.int/workforcealliance/knowledge/toolkit/33.pdf.

Sen, Amartya. 1999. *Development as Freedom*. New York: Alfred A. Knopf.

Seymour, Greg, Hazel Jean Malapit, and Agnes R. Quisumbing. "Measuring Time Use in Development Settings (English)." World Bank Policy Research Working Paper No. WPS 8147, July 2017. http://documents.worldbank.org/curated/en/443201500384614625/Measuring-time-use-in-development-settings.

Seymour, Greg, and Amber Peterman. 2017. "Understanding the Measurement of Women's Autonomy: Illustrations from Bangladesh and Ghana." IFPRI Discussion Paper No.1656. https://www.ifpri.org/publication/understanding-measurement-womens-autonomy-illustrations-bangladesh-and-ghana.

United Nations Foundation and Exxon Mobile Foundation. 2013. "Women's Economic Empowerment: A Roadmap." Last accessed March 20, 2018. http://womeneconroadmap.org/.

U.S. Department of Health and Human Services Office for Human Research Protections. 2016. "Special Protections for Children as Research Subjects." Guidance. Accessed March 19, 2018. https://www.hhs.gov/ohrp/regulations-and-policy/guidance/special-protections-for-children/index.html.

Waring, Marilyn, and Gloria Steinem. *If Women Counted: A New Feminist Economics*. San Francisco: Harper & Row, 1988.

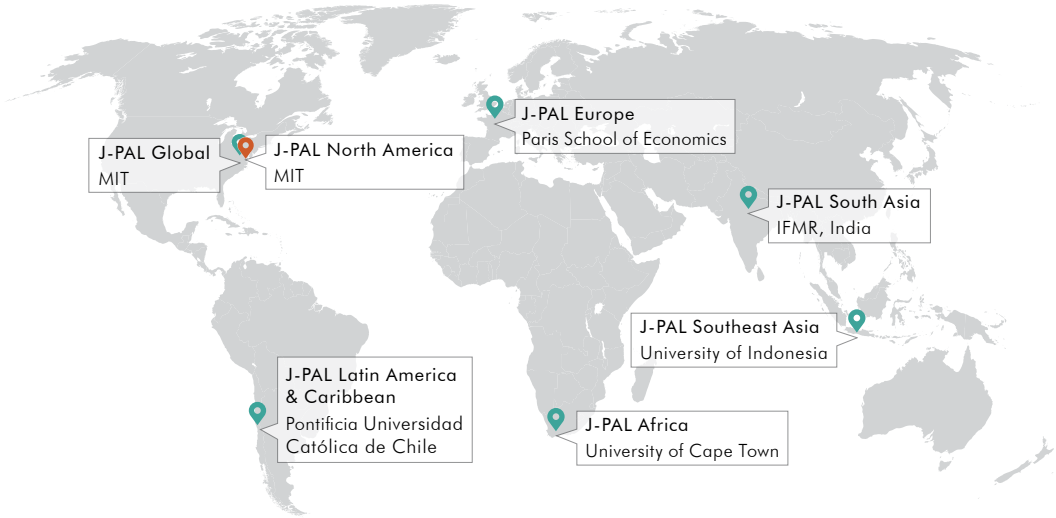
Watkins, Ryan, Maurya West Meiers, and Yusra Laila Visser. 2012. "A Guide to Assessing Needs: Essential Tools for Collecting Information, Making Decisions, and Achieving Development Results." Washington, DC: World Bank. http://hdl.handle.net/10986/2231.

ABOUT J-PAL

The Abdul Latif Jameel Poverty Action Lab (J-PAL) is a global research center working to reduce poverty by ensuring that policy is informed by scientific evidence. Anchored by a network of more than 160 affiliated professors at universities around the world, J-PAL draws on results from randomized impact evaluations to answer critical questions in the fight against poverty.

povertyactionlab.org

ABOUT THE GENDER SECTOR

Gender norms and biases continue to constrain human potential around the world. J-PAL's Gender sector produces cross-cutting insights on promoting gender equality and women's and girls' empowerment and on how social norms related to gender affect the outcomes of social programs.

povertyactionlab.org/gender