Australian National University

How to use the GIWL WLP Evaluation Framework

Elise Stephenson; Gosia Mikolajczak; Michelle Ryan; Victor Sojo; Alexandra Fisher; Jack Hayes; Morgan Weaving; Mai Tanjitpiyanond

Step 1: Establish a Theory of Change (ToC) for your program

When developing a WLP, or any program for that matter, it is important to first consider what your program is aiming to achieve and at what level this change is expected to occur (i.e., individual, organisational or structural)? Make sure to define clear, measurable goals for your program.

It is also critical to connect these goals to program activities; to be explicit about how program activities will contribute to these goals and affect change. For example, the goal of changing organisational biases, will likely require partnering with organisations to hold workshops, seminars, or other organisation-wide interventions. Whereas affecting change at the individual level would require program activities that focus on skill building for individual women.

During this phase, it may be helpful to consult with relevant stakeholders and members of the communities that your program intends to serve. What do individuals in these communities see as the biggest barriers to their leadership pursuits? And what do they want and expect from the program (e.g., what kinds of supports and resources do they perceive as most beneficial)?

Once goals and program activities have been defined, it's time to define how success will be measured. For example, success could be defined as a significant increase in self-reported leadership skills or aspirations in women from pre- to post-program.

Based on your ToC (or whether the program will have an impact at the individual, organisational or structural level), choose the indicators/measures in the Framework through which your program will assess change. For example, some programs aim to educate women on different leadership styles (e.g., transformational leadership at the individual level). As such, an individual level indicator might be most relevant, in particular, an indicator assessing leadership knowledge and skills (e.g., How would you rate your leadership skills, 1 = Very Weak to 7 = Very Strong). On the other hand, other programs may aim to address gender bias in organisational processes (i.e., organisational level), and as such should include organisational indicators such as experience and perception of bias indicators.

It is important to note that assessing and affecting organisational change may require multiple participants from the same organisation to complete the program, as broader organisational change may require a critical mass (multiple participants) working together to enact change. At the structural level, WLPs for women in Victoria may aim to increase women's knowledge about different leadership styles with the expectation that this knowledge overtime will also increase women's representation in leadership roles in the state. For this program, relevant measures to include in the survey would be leadership skills at the individual level as well as proportion of women leaders in Victoria at the structural level. For your program, the type of measures you choose will depend on your ToC.

Suggested Citation: Stephenson, et al (2024), "How to Use the GIWL Women's Leadership Program Evaluation Framework", Global Institute for Women's Leadership at the Australian National University.

Step 2: Plan a method for data collection

The Framework is designed to collect program data at two or more time points, in particular, before and after program commencement. Data collection after program commencement can be done during or immediately after the program ends but also at 6 months to multiple years follow up. Data collection can be done via online or paper survey. However, we recommend using online survey tools (e.g., Survey Monkey, Qualtrics) as it is easier to compile data from all participants, both within and across cohorts, to be used for later analysis.

Collecting data at more than two time points can allow for a more accurate assessment of change over time, but it will also require more sophisticated methods of data analysis. Thus, when deciding on the number of data collection time points, evaluators should consider their goals and the resources they have available and plan accordingly.

Important points to consider before data collection

For some measures and indicators, you might not be able to collect data using a survey. For instance, measures on the proportion of women leaders in Victoria might only be available via a public data repository. For these types of measures, the Framework includes suggestions for data repositories where you can retrieve relevant information. This information (e.g., the proportion of women leaders in Victoria), can be collected before, during, and after program implementation to provide a sense of change over time. Although it is not possible to explicitly link the implementation of one program to observed changes at the structural level, it can still offer a useful yardstick to get an idea of broader structural impact.

For the survey, once you pick all measures you want to include, don't forget to assign unique identifiers (id) for each participant of the program and use it to track all their survey completions (e.g., before and after program commencement). This way, when it comes to data analysis, you can match participants responses and track their progress before, during, and after program completion. Critically, however, you will want to consider how you will keep track of participant ids over time and how you will ensure that each participant is assigned the same unique id each time they complete a survey.

One option is to create a master key whereby one member of the research team has a file that links participants' names to their unique ids. Care should be taken to ensure that appropriate measures are in place so that only designated staff members have access to information linking participants' names to their ids (and thus their survey data). This method should also be disclosed to participants in a consent form before they take part in the survey.

Another option is to ask a series of non-identifying but memorable questions (e.g., what is the last two letters of your primary caregiver's middle name?; What are the first three letters of the very first school you attended?, etc.) that you can ask participants in each survey and then combine to create a unique id code.

Once your WLP starts, encourage participants to complete all surveys especially at the start and at the end of the program. Having more participants complete the survey will allow better assessments of program effectiveness.

Be very thorough in creating and designing the surveys. Make sure to familiarise yourself with survey tool of your choice to minimise any mistake made in the survey. For example, make sure all questions are written clearly in plain language. Ask colleagues to test out the surveys before launching it to participants. Once the survey is sent to participants, making changes to the survey could affect the quality of the data collected.

Suggested Citation: Stephenson, et al (2024), "How to Use the GIWL Women's Leadership Program Evaluation Framework", Global Institute for Women's Leadership at the Australian National University.

If you are planning to collect data at multiple time points, for example 6 months or 1 year follow-up after program completion, make sure you communicate this to your participants. You may ask participants if they are willing to be contacted in the future to complete a follow-up survey.

Important points to note after data collection

Once you finish collecting the data, preview the data file to make sure that the responses are coded correctly before data analysis. For instance, if a response of 1 equal to participants selecting "Very unlikely" for a question.

Check if any questions in the survey have a low response rate from the participants. If this is the case, check if the question is worded incorrectly or is being shown correctly to participants.

Step 3: Decide on a data analysis strategy

You can analyse your data at multiple time points, for instance, comparing survey responses before and immediately after program completion (or even at 6 months follow up). The purpose of data analysis is to establish whether your ToC is supported, or in other words whether your program affects change the way you intend it to. If your ToC is that the program will increase women's confidence, your data analysis would focus on comparing participants' responses on a confidence scale (e.g., how confident are you in your leadership skills) before and after program completion (e.g., using paired sample t-test).

Analysing and understanding the data enables you to assess the impact of your program. The data analysis strategy you choose will depend on the type of data you collected, whether it is quantitative, qualitative or both. Quantitative data are numbers that can be analysed using statistics whilst qualitative data is usually in the forms of texts, images, audio or video recordings. It is important that you have a team member (or external personnel) who can assist you with data analysis for you to best understand the data you collected. Some survey tools (e.g., Survey Monkey) may also have a basic built-in analysis function which can help you analyse and visualise data.

Important points to note on data analysis

A note on analysing qualitative data…

- ***Preparing the data:*** *Qualitative data can be analysed in many ways. Often, this process starts by preparing the data, for example, by transcribing interview data into transcripts.*
- ***Understanding the data:*** *The next step is familiarizing yourself with the data, which involves reading through participant responses several times to get a sense of the data keeping in mind your main research question(s).*
- ***Creating a coding framework:*** *Once you have a general understanding of the data, you may wish to develop a coding framework. This framework may be developed iteratively and may be based on previous research and theory that informs the concepts you wish to explore. Or you may take a more inductive approach whereby you allow the data to inform the codes you make. You can then systematically parse and code the data with your pre-defined or developing codes.*
  Once all the data is coded, you can begin to combine codes into shared themes. The codes within each theme should have a clear connecting thread. Try to ensure that these themes are distinct from one another and adequately characterise the underlying data. You may have to return to the coding process several times to refine your codes and themes.

Suggested Citation: Stephenson, et al (2024), "How to Use the GIWL Women's Leadership Program Evaluation Framework", Global Institute for Women's Leadership at the Australian National University.

- ***Drawing out themes:*** *Finally, you can synthesise themes into a coherent account, perhaps integrating other research and perspectives to make sense of the various themes in relation to your program.*

There are many different approaches to qualitative analysis (e.g., grounded, reflexive/thematic, discourse, qualitative content analysis, interpretive phenomenological analysis) so it is important to find the method that is most appropriate for the aims of your program and research question. For more detailed examples of reflexive thematic coding see Clarke and Braun (2006) and Byrne (2022) in the further readings section.

A note on analysing quantitative data…

As with qualitative data, there are many ways to assess change with quantitative data.

T-Tests

For data collected at two time points (i.e., pre- and post-program), a t-test is a straightforward way to assess whether average pre- and post-program scores significantly differ from one another.

T-tests come in several different varieties depending on characteristics of the samples in question. If, for example, the same women complete surveys at both assessment points, a paired samples t-test would be most appropriate. If, however, different women completed the surveys at each time point (as can often happen), an independent sample t-test would be more appropriate. Likewise, one-sample t-tests can also be used to compare participants' scores to a select value. This can be useful if pre-program data is not available or if there are meaningful scores or benchmarks from which to compare participants' outcomes.

When conducting a t-test, a p-value less than .05 suggests that the observed results would be unlikely in the event that the null hypothesis is true (in the event that there is no difference between pre-and post-program scores). In other words, a p-value less than .05 provides evidence that there are meaningful differences between the pre-and post-program scores. A p-value greater than .05 suggests there isn't enough evidence to reject the null hypothesis (i.e., there isn't enough evidence to conclude the two scores are different).

It should be noted, however, that tests of statistical significance require having adequate sample sizes, good measurement, and adequate statistical power to detect effects if they exist. See Bhandari (2022, November 11) for an accessible introduction to statistical power and how to increase it. If feasible, surveying a matched group of women who do not complete program, i.e., a control group of women, can also increase the rigor of your program evaluation, as observing a change among women who complete the program but not among women who did not complete the program would provide more compelling evidence that the observed change can be attributed to the program.

Magnitude of change

Beyond statistical significance, evaluators may also wish to understand the magnitude of change. To this end, most statistical software will calculate the mean differences between pre-and post-program scores, which can be used to calculate measures of effect size such as "Cohen's d." Often, effect sizes like Cohen's d can be directly requested from statistical software programs or other measures of effect size can be requested and then converted into Cohen's d with online calculators. Rough guidelines exist to quantify the size of effects. Typically, a Cohen's d of 0.20, 0.50, and 0.80 are interpreted as small, medium, and large effect sizes, respectively. For an accessible introduction to effect sizes, see Bhandari (2022, November 17), for a more advanced discussion see Lakens (2013).

Finally, when data is collected at more than two time points, evaluators may wish to use more sophisticated analyses such as repeated measures analysis of variance (ANOVA) or latent growth modeling. While t-tests and ANOVAs are designed to assess average levels of change across time points, latent growth modeling and other multilevel modeling analyses can allow evaluators to assess individual variation in patterns of change over time. For more information on these kinds of methods, see Hess (2000) and Lagarde (2012).

Analysis software

Free, open source, and user-friendly software such as JASP (JASP team, 2023) or Jamovi (The jamovi project, 2020) can be downloaded and used to conduct analyses. These software programs can easily accommodate data in .csv format, a commonly exported format across data collection websites. These programs also offer tutorials on how to conduct different types of quantitative analyses.

Step 4: Report and present your findings

Now that you have evaluated your program, you will want to share your findings with a larger audience. It is important to be transparent and honest about what you find even if the findings are not positive.

Evaluation findings can be used to assess gaps and shortcomings in program delivery that can be addressed in subsequent iterations of the program. Transparency therefore allows others to learn from your evaluation as well as work towards improving future programs and evaluation.

You can share your findings in a report with clear sections outlining program overview, your ToC, indicators, as well as your evaluation methodology.

Some Additional Resources

Bhandari, P. (2022, November 11). Statistical Power and Why It Matters | A Simple Introduction. Scribbr. Retrieved May 1, 2023, from https://www.scribbr.com/statistics/statistical-power/

Bhandari, P. (2022, November 17). What is Effect Size and Why Does It Matter? (Examples). Scribbr. Retrieved May 4, 2023, from https://www.scribbr.com/statistics/effect-size/

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. Quality & Quantity, 56(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Hess, B. (2000). Assessing program impact using latent growth modeling: A primer for the evaluator. Evaluation and Program Planning, 23(4), 419–428. https://doi.org/10.1016/S0149-7189(00)00032-X

Lagarde, M. (2012). How to do (or not to do) … Assessing the impact of a policy change with routine longitudinal data. Health Policy and Planning, 27(1), 76–83. https://doi.org/10.1093/heapol/czr004

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. https://doi.org/10.3389/fpsyg.2013.00863

Open Source and User-Friendly Statistical Software

The jamovi project (2022). jamovi (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org

JASP Team (2023). JASP (Version 0.17.1)[Computer software]. Retrieved from https://jasp-stats.org

Suggested Citation: Stephenson, et al (2024), "How to Use the GIWL Women's Leadership Program Evaluation Framework", Global Institute for Women's Leadership at the Australian National University.