The Value of our Evaluations: Assessing Spending and Quality

Report and contributing analyses completed by:

Amy Arbreton, Prithi Trivedi, Kris Helé, Fay Twersky, and Jing-Jing Zheng

I. Executive Summary

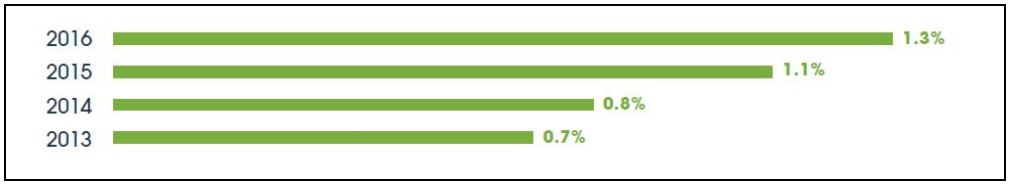
In 2013, the Hewlett Foundation adopted a clear framework for how we evaluate our strategies. We formalized a set of Evaluation Principles and Practices. We hired our first dedicated Evaluation Officer to support programs to commission, manage, interpret, and use evaluation findings. And, we made several recommendations to our Board (which were adopted) regarding evaluation spending and quality. Of course, foundation staff learn in many ways, but this set of activities represented a significant step towards a more disciplined approach for foundation staff to learn from independent third parties about the effectiveness of our grantmaking strategies.

In this report, we take stock of progress made on these recommendations, to ensure that we are not just increasing spending and funding more evaluations, but that we are increasing the utility and value of evaluations for more effective grantmaking. We base our assessment on financial data, ratings of evaluation documents, and interviews with program staff, for a set of 46 evaluations contracted directly by program staff between 2009 and 2016.

Below are the original recommendations we made to the board—and our findings on their progress.

1. *Over the next three years, the foundation should aim to increase its spending on evaluation to approximately 2 percent of program spending.*[1]

   We find that **spending on evaluation as a proportion of grants has almost doubled** between 2013 and 2016, from .7% to 1.3%. While we have not yet reached our goal of 2%,[2] the increase is notable given that the foundation's program grant spending also increased during this period—requiring even more to be spent on evaluation to keep pace.

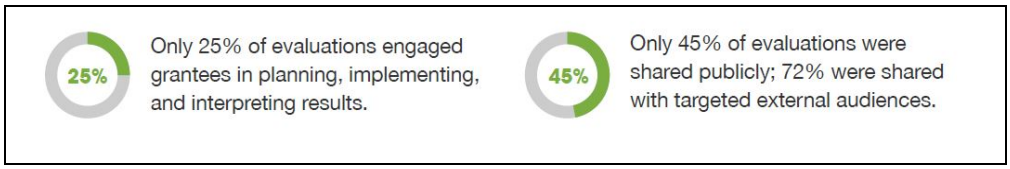


2. *The focus of our increased spending should be on improving the quality and practicality of our evaluations (as opposed to simply funding more of them), thereby producing insights that add value and lead to better grantmaking.*

   We find that **quality has increased overall**—though we are strong in some areas and weaker in others. In terms of quality, staff are typically clear on an evaluation's planned use, evaluations include various perspectives and more than one data source, and

evaluations are in fact used for a variety of purposes—commonly for course corrections and grant decisions. We also find that **efforts to increase spending and to increase quality are mutually reinforcing**. The evaluations that more closely follow our principles and practices tend to cost more.

That said, our evaluations could be improved by sharpening evaluation questions and integrating comparative reference points—for greater methodological rigor. Further, despite improvements in recent years, we can do more to engage grantees and can substantially increase our public sharing, which more recently stand at 25 and 45 percent, respectively.



3. *We will continue to pay for evaluations with a mix of administrative and grant budget funding. This means the additional spending should have minimal impact on our administrative costs, as both grant-funded evaluations and contracts that qualify in whole or in part as direct charitable activities are treated as coming from the grants budget without affecting administrative overhead.*

   As intended, **we are increasingly using grant budget funding to pay for a portion of our evaluations**, enabling us to keep administrative costs low while still increasing evaluation spending.

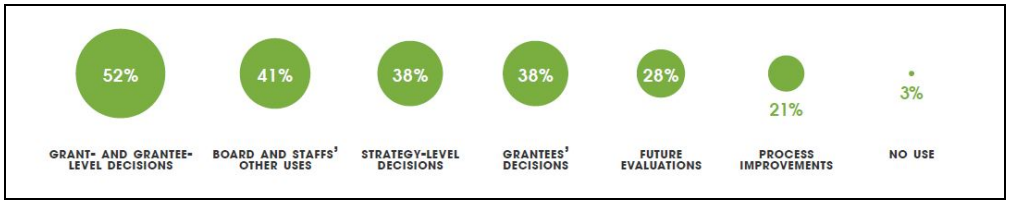
4. *We should improve our systems for tracking evaluation expenditures so we have more accurate data on overall costs and on the costs associated with different types of evaluations.*

   Our Evaluation Officer, in collaboration with the Finance Department, now tracks spending on evaluation contracts in a systematic, ongoing way. In addition, based on this assessment, we now have in-depth information about key features of our evaluations and the associated costs.

5. *We will assess the value we are deriving from evaluation and report back to the Board in three years.*

   Applying "use" as a proxy for value, we find that our evaluations are valuable. **Nearly all of our evaluations are being used in various ways**, including to inform grantmaking and strategic decisions, for board consideration, and by the grantees themselves. Our evaluations were being used in the 2009 to 2012 period as well, but since evaluation quality has increased since 2013, we are using better evaluations—and deriving more

value. We also find that we can further increase value by increasing grantee engagement and sharing.



Given what we have learned through this assessment, we make the following set of recommendations, with the goal of further increasing the quality, practicality, and value of our evaluations, which in turn leads to better grantmaking.

1. **The foundation should use a benchmark of 1.5% to 2% spending on evaluation as a proportion of grant awards—recognizing that rates of spending on evaluation by specific programs and strategies will fluctuate, depending on where they are in a strategy lifecycle.** This analysis showed that increased spending did improve certain aspects of quality. Aiming for a 1.5 to 2% benchmark should be a helpful target for improving the quality of our evaluations in the areas in which we are not yet strong. We suggest that every strategy or sub-strategy begin at least one evaluation within a 3-year time period.

2. **We should focus on increasing the quality of our evaluations in two key areas—engaging grantees and sharing the findings.** Both the Evaluation Principles and Practices and the Hewlett Foundation Guiding Principles[3] stress the importance of these practices. For grantee engagement, we recommend building in the time it takes to involve grantees during the planning, implementing, and interpreting results phases of an evaluation. For sharing, we recommend working closely with Communications staff from the beginning of an evaluation to consider plans, especially for public distribution.

3. Finally, similar to our recommendation in the last board memo, we believe it is important to **continue to track evaluation quality, spending, and value, to ensure we are learning and adapting, and report back to the board in five years—in 2022.** We will also revise our Evaluation Principles and Practices paper based on the lessons we have learned from this analysis.

II. Background and Methods

In a November 2013 memo,[4] Fay Twersky, Director of the Effective Philanthropy Group (EPG), reported on the Hewlett Foundation's evaluation spending and made the following recommendations—which the Board adopted:

1. Over the next three years, the foundation should aim to increase its spending on evaluation to approximately 2 percent of program spending.[5]
2. The focus of our increased spending should be on improving the quality and practicality of our evaluations (as opposed to simply funding more of them), thereby producing insights that add value and lead to better grantmaking.
3. We will continue to pay for evaluations with a mix of administrative and grant budget funding. This means the additional spending should have minimal impact on our administrative costs, as both grant-funded evaluations and contracts that qualify in whole or in part as direct charitable activities are treated as coming from the grants budget without affecting administrative overhead.
4. We should improve our systems for tracking evaluation expenditures so we have more accurate data on overall costs and on the costs associated with different types of evaluations.
5. We will assess the value we are deriving from evaluation and report back to the Board in three years.

In this report, we take stock of the progress we have made on these recommendations. First, we assess how our spending has changed over time, and where we are against the 2% goal. Second, we assess the quality[6] of our evaluations and see how it has evolved over time—to examine whether we are increasing the utility and value of evaluation for more effective grantmaking. To assess quality, we use the foundation's Evaluation Principles and Practices as our guide.

The primary audience for this report is our Board. From this assessment, however, we learned lessons which will prove useful for our program peers and colleagues in the field. We are committed to sharing what we have learned more broadly; not only is this one of the foundation's principles, but it is especially salient at this time given a recent call for greater transparency by foundations about what they are learning from their evaluations.[7]

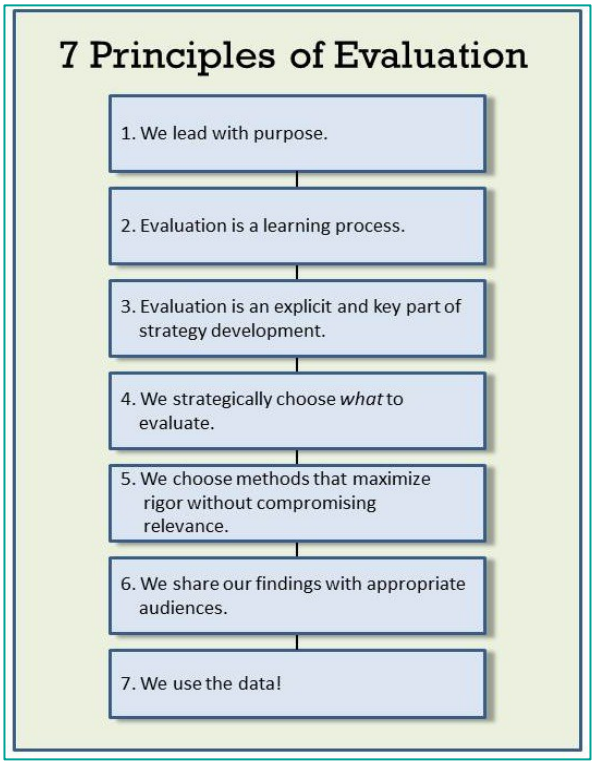
METHODS

For this assessment (and in the Evaluation Principles and Practices paper), we define evaluation as "an independent, systematic investigation into how, why, and to what extent objectives or goals are achieved." All evaluations included in the assessment were conducted by independent third parties and contracted directly by program staff to inform

their grantmaking. We **did not assess:** research studies that are themselves part of our strategies to produce evidence for a field, or projects funded to gather data primarily with the intent of tracking progress on our implementation markers. Since we do not yet have a system in place for systematically tracking other types of evaluations (e.g., funded as part of a larger grant to an intermediary), these are also not included.

Spending: To gauge our spending on evaluation, we used two different data sources. First, we used financial records from 2013 to 2016 to calculate both the annual evaluation spending as a proportion of program grants and the proportion of administrative versus charitable dollars spent. These are the years for which complete and accurate data regarding evaluation expenditures are available in the foundation's financial tracking system. In calculating the proportion, we excluded grants from the denominator if: 1) they were not associated with a particular strategy (e.g., our special projects grants), and therefore we would not expect to conduct an evaluation to improve strategy, or 2) a grant was a significant outlier (for example, a single grant for $100 million), in which case the 2% spending benchmark would not be appropriate. Second, we reviewed available evaluation contracts to examine whether and how our spending on individual evaluations (which sometimes carried over multiple years) has changed between 2009-2012 and 2013-2016.

The main limitation of our spending data is the differing sample sizes and time periods for which reliable data are available. It is likely that we missed some evaluations commissioned between 2009 and 2012 because the expenditure and contract tracking system was not centralized and the Evaluation Officer position did not exist. We had program staff review the list of evaluations and add to it, based on their recollection—but nevertheless, we might have missed some. Since 2013, the Finance department and EPG have worked together to track dollars spent on evaluation contracts; this has greatly improved our ability to understand how much we spent on evaluations for the years 2013 to 2016.



Quality: To assess evaluation quality, first we worked with program staff to identify evaluations initiated at the foundation between 2009 and 2016. This sample was narrowed to those that could deliver a final report to program staff by August 2016 (when interviews

for this assessment were finished). The resulting sample includes 46 evaluations; unless otherwise noted, the quality analyses in this report are based on these evaluations.

Second, we developed a scoring rubric to systematically gauge the quality of evaluations; this rubric is based on the foundation's Evaluation Principles and Practices, and includes categories such as clarity of purpose, rigor and relevance, engagement of grantees, and sharing of evaluations. (Appendix A contains the rubric used.) We also identified information about the characteristics of our evaluations—such as type of evaluation (formative/ongoing/intended to inform adjustment, summative/at the conclusion or inflection point to inform decision-making or strategy refresh, or exit/to "tell the story" at close-out); the unit of evaluation (strategy, cluster, grant); length of the evaluation; whether the evaluator was an organization or an individual; and who initiated the evaluation. (The list of characteristics is in Appendix B.) We then reviewed a range of available documents (e.g., evaluation contracts, Requests for Proposals, design documents, interim and final evaluation reports) and rated each of the 46 evaluations.

Third, we conducted semi-structured interviews with the program staff (current and former) who had commissioned the evaluations (or were recipients of the final report, in the cases where staff transitions took place during an evaluation), and asked them to complete a brief survey about each evaluation. (The protocols we used are included in Appendices C and D.) We used these methods to complement the information gathered through our document review.

Finally, after we collected the data, we conducted statistical and content analyses. We looked more deeply at relationships among key variables including evaluation spending, characteristics, and quality, and assessed whether and how these variables changed over time.[8] We used 2013 as the cutoff in our comparisons over time, because in that year, we had formalized and disseminated our guiding Evaluation Practices and Principles, hired a new Evaluation Officer to support programs to commission evaluations, and made our evaluation goals and recommendations to the Board.

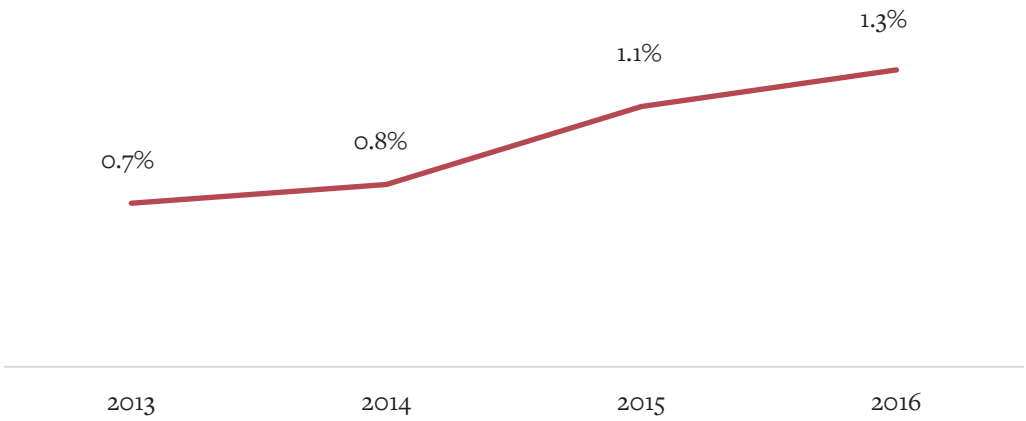
The main limitations of our quality assessment are: variation in the number and type of documents available for each evaluation (e.g., for some evaluations we could not locate interim reports, final contracts, or design documents), particularly for older evaluations, which tended to have fewer documents to rely on for making the assessment; the subjective nature of our assessment of some categories, such as strength of evaluator recommendations; and, the fact that for evaluations that were more recently completed, some may not have had the chance to have been used or shared to the same extent as those completed longer ago. We made every effort to mitigate challenges where possible.

III. Spending on Evaluation

Between 2013 and 2016, annual spending on evaluation as a proportion of program grants almost doubled. While we have not yet reached our goal of 2%,[9] the increase is notable given that the foundation's program grant spending also increased during this period—requiring even more to be spent on evaluation to keep pace.

Our spending has nearly doubled.

Annual Foundation Evaluation Spending as a Proportion (%) of Program & Initiative Grants



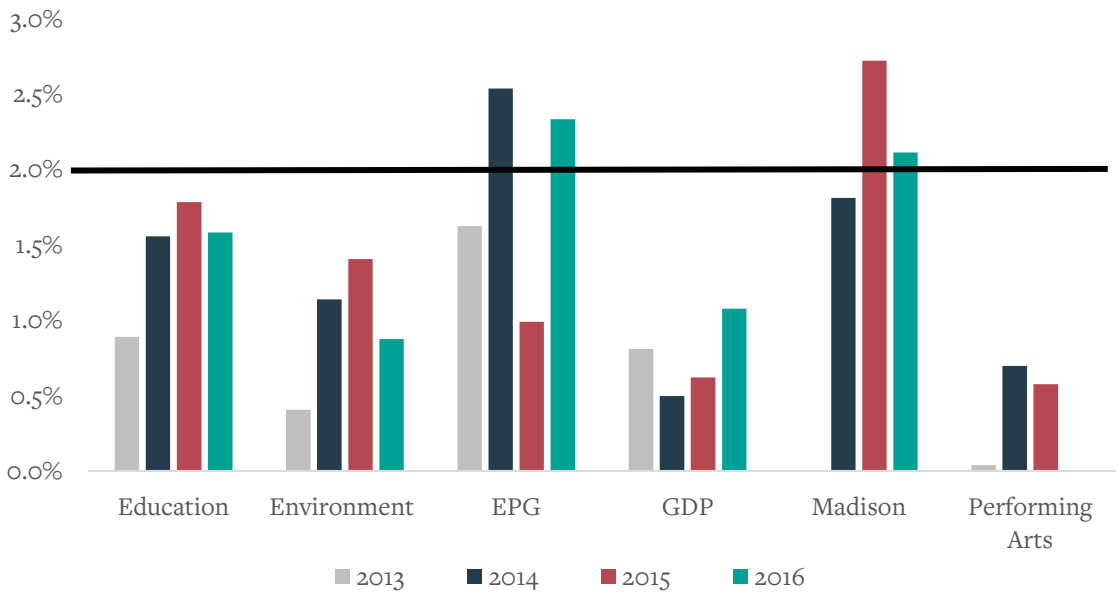
The overall increase in evaluation spending over time is primarily attributable to the following:

1. We are initiating more evaluations per year. For instance, 8 evaluations were initiated in 2013, compared to 13 evaluations initiated in 2016.
2. We are evaluating new initiatives developed post-2013 (e.g., Madison, Cyber), which have relatively large evaluation contracts.
3. We are spending more on individual evaluation contracts. For example, the median contract amount doubled from $69K for contracts initiated in 2013 to $137K for contracts initiated in 2016.[10]

Although the foundation as a whole has seen a steady proportional increase in evaluation spending, there is variation in spending among individual programs and initiatives by year. Only the Madison Initiative and the Effective Philanthropy Group (EPG) (relatively smaller grantmaking programs) have met/exceeded the 2% target, though the Education program has routinely spent at or above 1.6% in the past three years.

Madison and EPG met the 2% target.

Annual Program Evaluation Spending, Proportion (%) of Program & Initiative Grants



The variation in proportional spending by program and year somewhat reflects the episodic nature of when, for what aspect(s) of their strategy or sub-strategies, and over how long of a period programs are commissioning evaluations.

The Madison Initiative, for example, commissioned a developmental evaluation covering the entirety of the strategy; they contracted this evaluation at approximately 2% of the initiative's award level for its first three years. Other programs appear to be spending more in the years leading up to and just after a refresh of a strategy or sub-strategy. EPG shows a spike in 2014, related to simultaneous refreshes for the Knowledge and Organizational Effectiveness strategies. Similarly, Education shows a spike in 2015, related to the refresh of the Deeper Learning and Open Educational Resources strategies. Global Development and Population's proportional spending has been increasing in the last few years, with its peak so far in 2016. In part, this is because program staff have been working with evaluators after refreshing many of the program's sub-strategies (e.g., International Women's Reproductive Health: supporting local advocacy in Sub-Saharan Africa, and Transparency, Participation and Accountability), to plan for evaluations going forward and to build in evaluation earlier. In fact, this increase in spending on evaluation by the Global Development and Population program, given the large size of its grant awards, has been a substantial contributing factor to the foundation's overall increase in evaluation spending. The Performing Arts program's evaluation spending increased in 2014, when staff commissioned an evaluation of the regranting and intermediary organizations they work with, and again in 2015 when conducting a midpoint assessment of their entire program strategy. Alternatively, the Environment program has spent most on evaluation at the end of their strategies; in 2015, staff commissioned relatively large "close out"/exit evaluations of the Clean Transportation

Initiative and of the Nuclear Security Initiative, to summarize achievements and gather lessons.

PAYING FOR EVALUATIONS

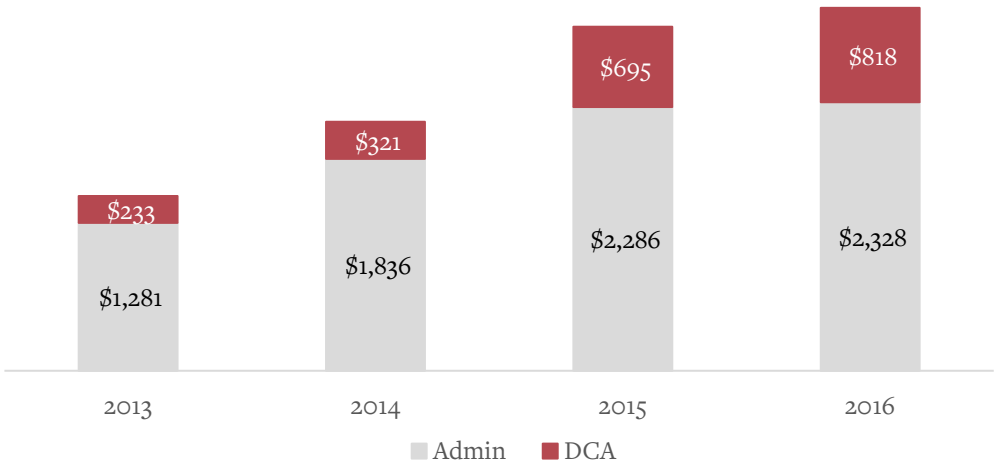
In our 2013 recommendations to the Board, we asserted our intention to fund evaluations with a mix of administrative (non-charitable) and Direct Charitable Activities (DCA)[11] funds. DCA are philanthropic activities that a foundation engages in directly, rather than indirectly through grants to grantees; these must have a charitable purpose that extends beyond the foundation.

Using DCA funds for evaluation rather than administrative dollars is important for two reasons. First, using grant budget funding for part of the evaluation contract keeps our overhead rates low, even while increasing our spending on evaluation. Second, when we create a contract that uses DCA funds, we are required to set the evaluation up in a way that encourages greater engagement of grantees to learn from and use the evaluations, as well as more external sharing of findings—both aligned with our evaluation principles.

We have delivered on our payment intention: the overall annual increase in evaluation spending is accompanied by an increase in the use of DCA funding to fund a portion of the evaluation contracts. Foundation evaluation spending increased from $1.5M in 2013, to $2.2M in 2014, to $2.9M in 2015, and finally to $3.1M 2016—with the proportion of DCA funds increasing from 15% in 2013 to 26% in 2016.

We have increased the use of DCA funds to pay for our evaluations.

Evaluation Spending ('000)



WHY HAVEN'T WE SPENT MORE?

We believed that the conditions we had in place for increasing evaluation spending might have been sufficient to help us reach the 2% goal within the past few years. In particular, the Board's approval for the proposed increase in spending has been a key asset; board leadership support is cited by a Center for Effective Philanthropy/Center for Evaluation Innovation national benchmarking study[12] as a strong contributing factor among foundations who have increased funding for evaluation in recent years. Additionally, the budgets that programs allocate for evaluation are reserved specifically for those purposes. As such, these funds are not directly taking away from programs' grantmaking, and they are not available for conversion to program grants—eliminating the tradeoff between program and evaluation grants, a common obstacle to increasing evaluation spending. Further, the EPG Evaluation Matching Funds (additional funds out of EPG's budget that can be used for program-funded evaluations) provide an additional incentive to spend. Still, we have not yet reached the 2% evaluation spending goal. Several factors appear to play a role in explaining why.

1. Programs often pause in funding evaluations, due to where they are in their strategy's lifecycle, staff transitions, or "evaluation fatigue." For example, while going through a longer than anticipated strategy refresh, or after completing an evaluation, programs have paused in commissioning evaluations to reflect and develop a new plan or set of evaluation questions. In a number of cases, evaluations have been postponed or suspended due to an upcoming or recent staff transition, in part related to term limits at the foundation. In a couple of instances, "evaluation fatigue" among program staff or grantees has affected the inclination to evaluate.

2. Program staff fear that the dollars or time invested may not pay off. One way staff have addressed their fear that the evaluation will not be worthwhile is to approach an evaluation in phases—to first engage an evaluator in a design phase, and if the design is solid and fit is strong, then engage in an implementation phase. This is a helpful tactic for ensuring that there is a good fit between the evaluator and program staff, but it has the additional effect of spreading out contract spending in smaller amounts over time and slowing down projected spending. It also sometimes results in discontinuing the evaluation if the evaluator is not a good fit.

3. Our program grant awards have significantly increased, so we are trying to keep pace with a large and increasing denominator. As such, our evaluation spending as a proportion of overall grant spending is lagging.

Nevertheless, we believe that our recent overall spending trends and more focused attention on what, when, and why we are commissioning evaluations by program staff are strong indicators that we are on the right track, even if we are not meeting our 2% target.

IV. Evaluation Quality

The true goal of increasing spending is to increase the quality and utility of our evaluations—in order to focus more attention on learning to inform our work and the work of our grantees. This section assesses the quality of our evaluations, using our Evaluation Principles and Practices as a guide. (For a description of other characteristics of the evaluations in the sample, including an illustration of how we choose strategically what to evaluate, see Appendix E.)

Principle or Practice	Characteristics of Strong Evaluations
Lead with Purpose	Planning and design documents with clearly articulated audience, intended use, and timing needed for results
Evaluation Questions	Precise and clearly articulated, evaluative (why/why not, how, for whom, and compared to…) questions
Rigorous and Relevant Methodology	Evaluator incorporates multiple perspectives, mixed methods and comparative reference point(s); data are well-triangulated; evaluator articulates limitations and includes data collection tools
Clear Interpretation of Findings and Recommendations (if recommendations requested)	Evaluator effectively analyzes and triangulates the data to provide a clear interpretation of findings. If requested, evaluator presents well-sounded and prioritized recommendations that are useful, actionable given context/audience, and based on the findings (note that recommendations can be counterproductive because if they are off-base or naive, they can undermine the credibility of the overall evaluation).
Engaging Grantees	Grantees are engaged in all evaluation phases: planning, implementation, and during interpretation of interim and final results
Use the Data	Findings are used as intended and fully for learning and course correction
Sharing What We Are Learning	Both internal and external sharing, including public sharing in some form

Black: Program staff primarily responsible; Red: Evaluator primarily responsible

For each quality indicator (presented in more detail below), we look at evaluations commissioned between 2009 to 2012, compared to those commissioned between 2013 and 2016, to describe whether and how quality has changed over time.[13] Where possible, we delve deeper into why a particular aspect of quality might—or might not—have changed.[14] We also highlight the areas in which evaluations are strong, and the areas in which there is room for further improvement.

Overall, we find that the quality of our evaluations has improved. More recent evaluations have a significantly clearer purpose, are more likely to have stronger evaluation questions,

are more likely to engage grantees, and are shared with more audiences. We attribute this improvement to four factors: 1) our efforts to institutionalize strong evaluation practices, including the formal adoption of a set of Evaluation Principles and Practices; 2) high quality support from the Effective Philanthropy Group, especially from the Evaluation Officer; 3) program staff who are self-reflective and eager for feedback; and 4) the increase in spending on evaluation.

CLARITY OF PURPOSE

Our Evaluation Principles and Practices describe the importance of clearly articulating evaluation purpose (including the intended use, target audience and timing of when findings are needed)—regardless of evaluation type (i.e., formative, summative, or at exit) or unit of analysis (i.e., grantee, cluster, or strategy).

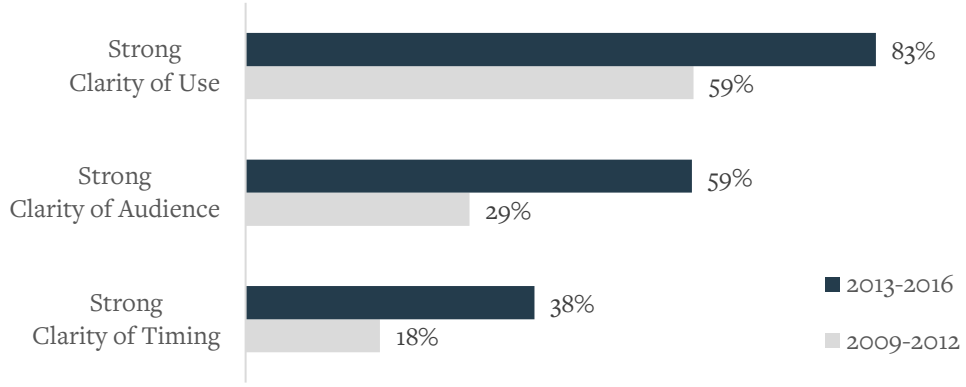
The purpose of an evaluation is central. Questions, methods, and timing all flow from a clear understanding of how the findings will be used.
–Evaluation Principles and Practices

We find that a large majority of our evaluation documents are strong in clearly articulating intended use. However, fewer are strong in documenting the intended audience and even fewer are clear on the timing of when results are needed.

In interviews, when asked what they might have done differently, staff echoed the importance of establishing purpose. In particular, they would have benefitted from clarifying who would own and act on the results, and from better planning up front to determine how the evaluations could be most useful.

Notably, we see greater clarity of purpose across all categories in our evaluations conducted between 2013 and 2016 compared to those conducted between 2009 and 2012. This is due in large part to our tendency to more clearly articulate purpose in evaluations with larger contracts, which have been more common in recent years.

Since 2013, we have become clearer on audience, use, and timing for evaluations.



EVALUATION QUESTIONS

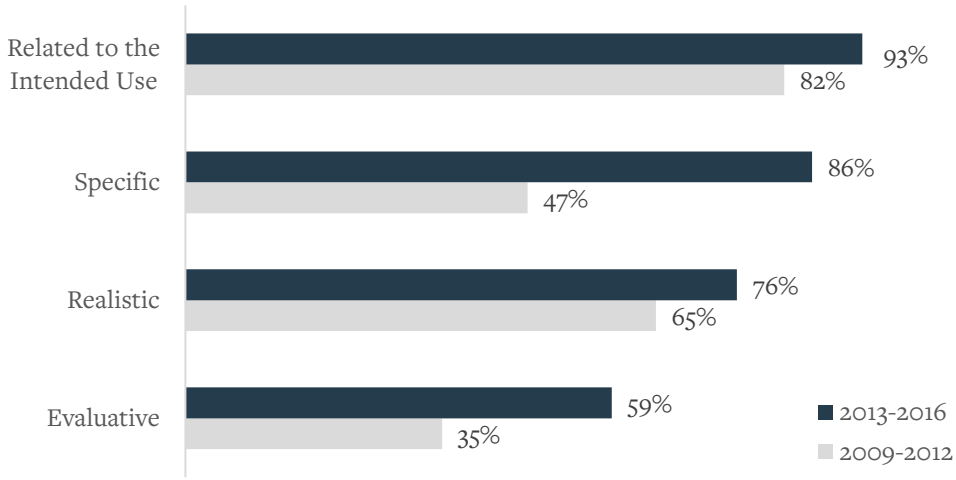
Strong evaluative questions (e.g., those that delve into answering why/why not, how, under which circumstances, for whom, and compared to what) are more likely to stimulate deeper learning and insight. These types of questions most clearly embody the foundation's principle "evaluation is a learning process."

In recent years, we see a greater proportion of evaluations guided by strong questions. Still, there is room for improvement here—more of our evaluations can begin with strong evaluative questions to best position the evaluation to provide novel insights for learning.

Crafting a short list of precise questions increases the odds of receiving helpful answers—and a useful evaluation. Well-designed questions about an initiative or program can clarify not only the expected results but also surface assumptions about its design, causality, time frame for results, and data collection possibilities.
–Evaluation Principles and Practices

Though our evaluation questions have become stronger since 2013, there is still room to improve.

Figures exceed 100% due to the possibility of demonstrating multiple categories.



The foundation's approach to strategy, Outcome-Focused Philanthropy,[15] recognizes the importance of evaluation and incorporates evaluation into its guidelines at every stage of the strategy lifecycle. The following strong evaluation questions are from evaluations commissioned both with different purposes and at different stages of a strategy lifecycle[16]:

Strategy to Apply Human-Centered Design (HCD) to Improve Family Planning and Reproductive Health Services in Sub-Saharan Africa *(Formative Assessment of Approach)*

- *How do solutions designed using HCD work? How has the HCD process contributed to their effectiveness and sustainability?*
- *What is the relative contribution and value of each of the components and design mindsets of HCD to the process of designing an effective and sustainable solution?*
- *What have been the key successes and challenges of applying HCD to increasing access and uptake, including for scalability and sustainability?*

- *What is the value of HCD-designed solutions compared with other youth reproductive health models? What is the value of the IDEO.org HCD solutions in Kenya and Zambia compared with other HCD-inspired solutions?*
- *What does it take to effectively introduce and maintain the key capacities needed for developing and sustaining HCD processes?*
- *To what extent and why does HCD hold promise for application by other donors and in other social sector fields or contexts?*
- *What factors have enabled and inhibited success?*

Quality Education in Developing Countries Citizen-Led Assessments[17] (*Formative Assessment of Process and Outcomes*)

- *How well have grantees executed on the core components of their strategies: high quality data collection and analysis (including quality of assessment tool, sampling, survey process, data entry, etc.) and communication of results?*
- *To what extent have citizen-led assessments increased awareness of learning outcomes and influenced actions to address poor learning achievement?*
- *What contributes to success? Which activities explain the impact citizen-led assessments have had on increasing awareness of learning outcomes and influencing actions to address poor learning achievement?*
- *In what contexts do these efforts have the most/least traction and why?*

Performing Arts (*Mid-point Summative Assessment*)[18]

- *In light of our short and long-term goals, what have been our key accomplishments and challenges? What do we know about the main drivers—both the enablers and inhibitors for these results?*
- *Which geographic and demographic communities have benefitted from Hewlett support and where do gaps lie?*
- *To what degree did our original assumptions about how change would happen turn out to be true?*
- *What are the shifts in the external landscape and broader Bay Area arts sector, including new research that may call for some adaptation of the Program's core strategies, our targets for change, or how we measure progress?*

Deeper Learning (*Mid-point Summative Assessment*)[19]

- *To what extent are our four clusters of activity successful to date—are they making the expected progress by this point in our Deeper Learning strategy's execution?*
- *What have been the key enabling and inhibiting factors in making progress—considering issues of concept (our ideas and assumptions about how change happens), execution, and changing context?*
- *What is the likelihood that we will achieve our 2017 goals? What gives us confidence and/or what should give us pause?*
- *What corrections might we consider making to our Deeper Learning initiative in order to maximize our chances for success? Consider issues of a)strategy—overall ambition and theory of change, including our assumptions about how change happens, our goals, and intermediate targets for progress; and b)execution—such as timing, sequencing, our grantmaking and advocacy approaches, and grantee and partner selection?*

Nuclear Security Initiative[20] (*Commissioned at Exit to Sum Up Initiative Lessons*)

- *To what extent and how did Hewlett make progress in their primary NSI investment areas? Within the investment areas, what components (e.g., advocacy, research, policy) of the strategy portfolio achieved progress?*
- *What were the primary inhibiting and enabling factors that influenced the success of the strategy? How did grantees adapt to the factors?*
- *How did Hewlett's grantmaking structure support success for the strategy (e.g., providing general operating support)?*
- *What has been effective about Hewlett's approach to the NSI work that could also be effective for other investment strategies in the future?*

RIGOR AND RELEVANCE

Nearly all the evaluations assessed use relevant, multiple, complementary methods that include a range of perspectives (e.g., grantees, field experts); our evaluations were rigorous and relevant before 2013, and continue to be.[21]

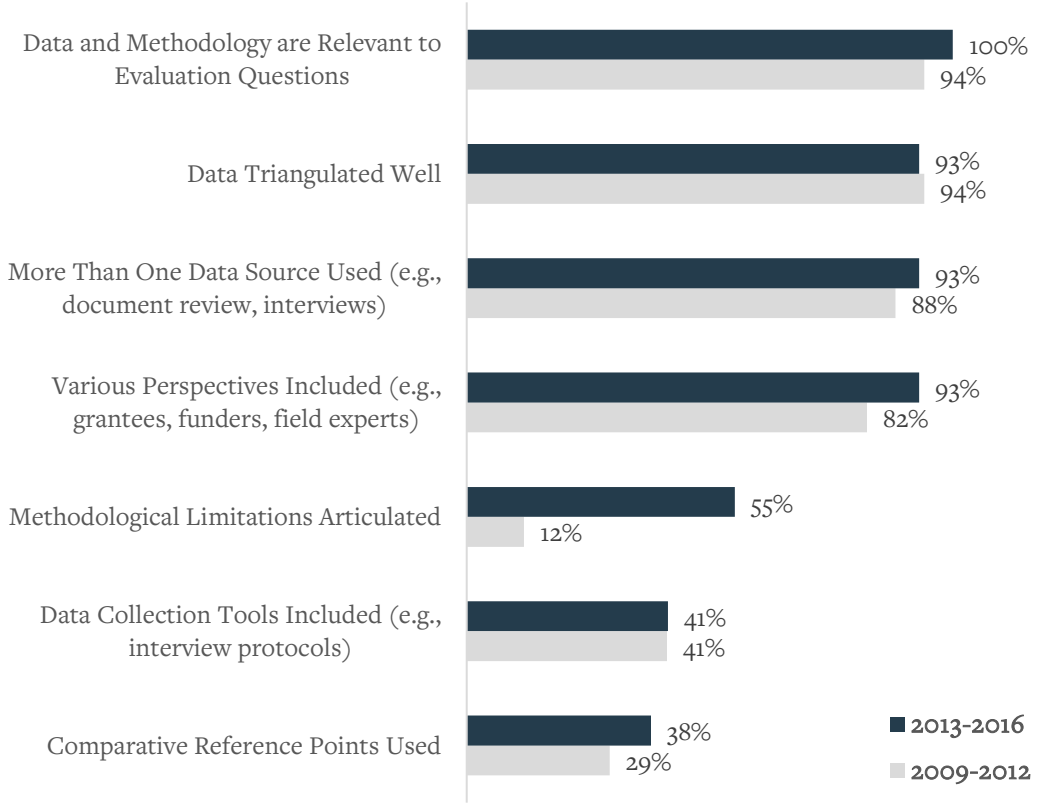
However, there is ample room for us to improve our evaluations' methodology in three key areas. First, only roughly one-third of evaluations are contextualizing their data against comparative reference points/benchmarks to provide an evaluative assessment—for instance, comparisons over time, against stated objectives or goals, relative to similar efforts, or based on agreed upon markers of quality or progress. Without such comparisons, the evaluation findings and insights are often not compelling, and they leave out important information that can help interpret the value of strategies or grants. Second, fewer than half of evaluations demonstrate further evidence of rigor in the project documents by including data collection protocols. Finally, only just over half provide a discussion of methodological limitations.

Most strong evaluations use multiple methods to collect and analyze data. This process of triangulation allows one method to complement the weaknesses of another…It is ideal to select methods that match evaluation questions.

The essence of good evaluation involves some comparison—against expectations, over time, and across types of interventions, organizations, populations, or regions.
–Evaluation Principles and Practices

Though our evaluations are strong in some areas, there is room to improve rigor and relevance—especially by using more comparative reference points.[22]

Figures exceed 100% due to the possibility of demonstrating multiple categories.



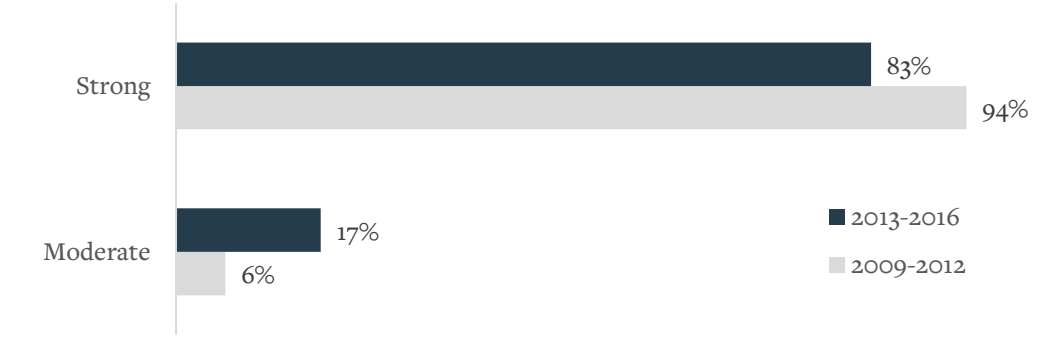
As noted above, to strengthen their methodology, more evaluations can integrate more comparative references. One challenge is that baseline assessments were not regularly conducted by the evaluations in the sample, hindering the possibility of comparison. A few of the more recent evaluations (post-2013) have been building in stronger data collection systems and gathering data earlier in implementation, to be able to provide a comparative reference point in the future.

EPG's Knowledge for Better Philanthropy strategy is a prime example of an evaluation that faced—and dealt with—the challenges of ensuring methodological rigor and establishing a comparative reference point. In 2013, when the first evaluation of the Knowledge for Better Philanthropy strategy was commissioned after 13 years of implementation, Program Officer Lindsay Louie wanted to assess grantees' progress in producing high-quality knowledge for the philanthropic sector. But she and her evaluator soon realized that little useful information was available to identify what progress had been made. To address this issue, the evaluator gathered information on a sample of products generated by the grantees and created a rubric to rate the products' quality.[23] While the evaluation was helpful in answering questions about quality and channels of dissemination, and "reach" of

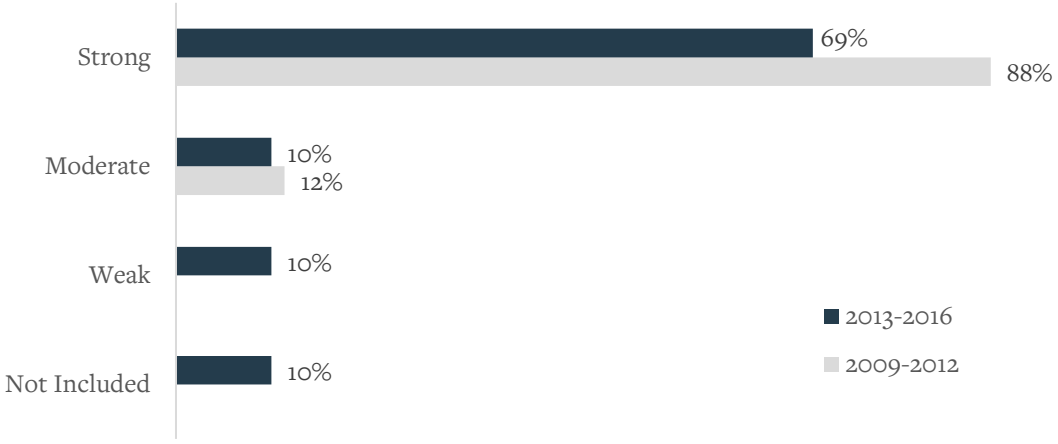
knowledge, it did not yield information about use of the knowledge products. To tackle this challenge, Lindsay convened the grantees to discuss the evaluation's findings and limitations; together, they decided to create a shared data collection system allowing for a more systematic assessment over time. She also commissioned a new evaluation to determine to what extent target audiences were using the knowledge produced by grantees.[24] As a result of these evaluation efforts, Lindsay and the grantees now have a "baseline" in place, which informs realistic targets and makes comparisons possible going forward.[25]

We include interpretation of data and recommendations in this section, since they are deeply connected—and can result from—rigorous and relevant methodology. A large majority of evaluations in both time periods receive high ratings for effectively analyzing and triangulating the data—providing clear interpretation of data and findings.

A majority of evaluations receive high ratings for interpretation of data and findings.



The decrease in strong recommendations is related to an increase in reports that do not include recommendations.



A shift down in the proportion of evaluations with "strong" recommendations over time is related to an increase in evaluation reports that do not include recommendations. In one case, the EPG Organizational Effectiveness staff specifically requested that the evaluator not provide recommendations since they felt that the evaluator was not in a position to truly understand the context within which the OE program would be making strategic decisions. Instead, the evaluator provided a set of questions to reflect on.[26] Sometimes, in fact, EPG has found that recommendations can be counterproductive because if they are off-base or naive, they can undermine the credibility of the overall evaluation.

Sometimes asking an evaluator NOT to provide recommendations works best, since the evaluator may not be in a position to truly understand the context within which program staff will be making strategic decisions.

That said, several program staff indicate that they prefer to have a clear set of recommendations from the evaluator, even if they end up not acting on all of them, because they are helpful for sparking conversation and ideas. To the extent that recommendations are useful at all, staff described more valuable recommendations as those that are more "*actionable*" and "*on target*"—and those less valuable as "*G-rated, too broad, not prioritized*" or "*soft-pedaled.*" [27] To increase the value of the recommendations, some evaluations build in time for the evaluators to co-create recommendations along with program staff, grantees, and/or evaluation advisory committees; this can be helpful given that evaluators are not necessarily aware of all the constraints or factors that shape future actions or affect their implementation.

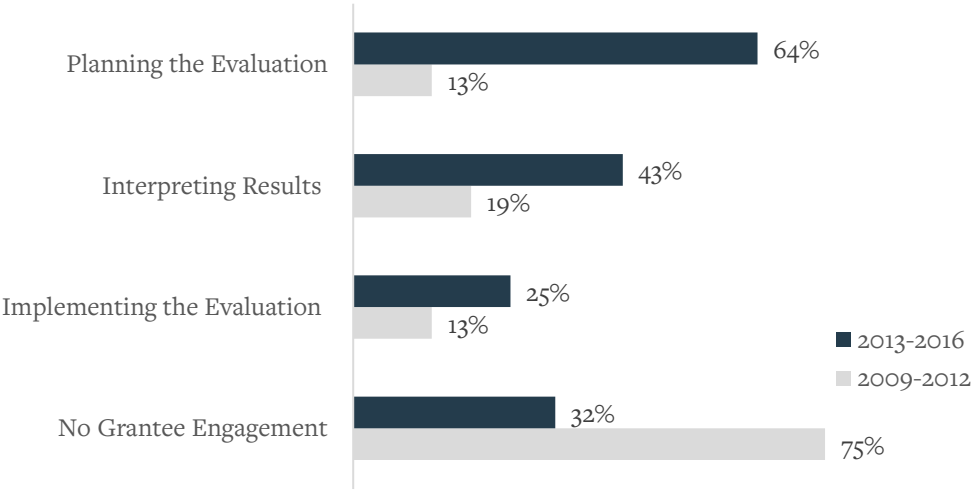
GRANTEE ENGAGEMENT

We have greatly increased grantee engagement since 2013 overall and in all evaluation phases—planning, implementing, and interpreting results—yet there is room to further reap the benefits. One third of our evaluations conducted between 2013 and 2016 still do not engage grantees at all. We heard from program staff that the time it takes to get grantees' input and apply it can be challenging, and that engaging grantees can hamper the ability to complete an evaluation quickly.

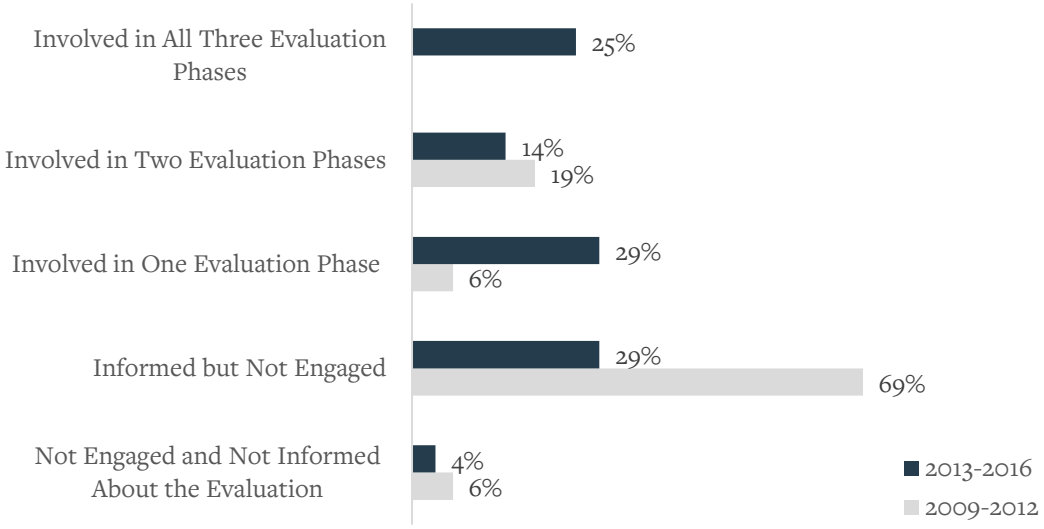
[Engaged grantees] will be: (1) more supportive with respect to data collection; (2) more likely to learn something that will improve their work; (3) less likely to dismiss the evaluation; and (4) better able to help strengthen the evaluation design, especially if engaged early.
-Evaluation Principles and Practices

We have increased grantee engagement in all phases since 2013, though there is still room for improvement—especially in interpreting the results and implementing the evaluation.

Figures exceed 100% due to the possibility of demonstrating multiple categories.



A large percent of grantees are informed about an evaluation, but not engaged in the process.



The evaluation of the Early Learning Innovation cluster of grantees from the Global Development and Population program provides an excellent example of grantee engagement across all three evaluation phases. To kick off their process of involving grantees, the program staff (Pat Scheid and Dana Schmidt) prepared a draft document outlining the evaluation purpose (audience, intended use, and timing) and initial set of evaluation questions. They shared the document with the grantees, who were also intended users of the evaluation findings for their own learning and improvement. Pat and Dana then met with each of the grantees, who provided helpful input about what would be most

valuable for them to learn, and added evaluation questions of their own. As the evaluation launched, the grantees provided ideas for strengthening data collection processes. Finally, as the evaluator drafted the findings and recommendations, the grantees helped with the interpretation, particularly in relation to their own processes and work with sub-grantees, as well as on the broader lessons from the evaluation. The evaluator produced several reports, including one that focused on findings for funders, and others that focused on individual grantees. Involving grantees helped maximize the value of this evaluation in terms of its quality, practicality, dissemination, and use.[28]

EVALUATION USE

If we look at "use" as a proxy for "value," we are deriving value from our evaluations. Evaluations are used frequently for program staffs' grant/grantee-level decisions, including shaping grant renewals, closing a grant, switching from project grants to general operating support, or identifying potential areas for organizational effectiveness support. Evaluations are used for strategy-level decisions, such as making course corrections, testing assumptions, strengthening a strategy, or exiting aspects of a strategy. Program staff also commission smaller evaluations that they plan to use to ultimately "roll up" into a larger summative, to improve future data collection, or to engage other funders.

The majority of our evaluations seek to inform the decisions and practices of the Hewlett Foundation and our grantees—to support our ongoing learning, adjustment, and improvement.
-Evaluation Principles and Practices

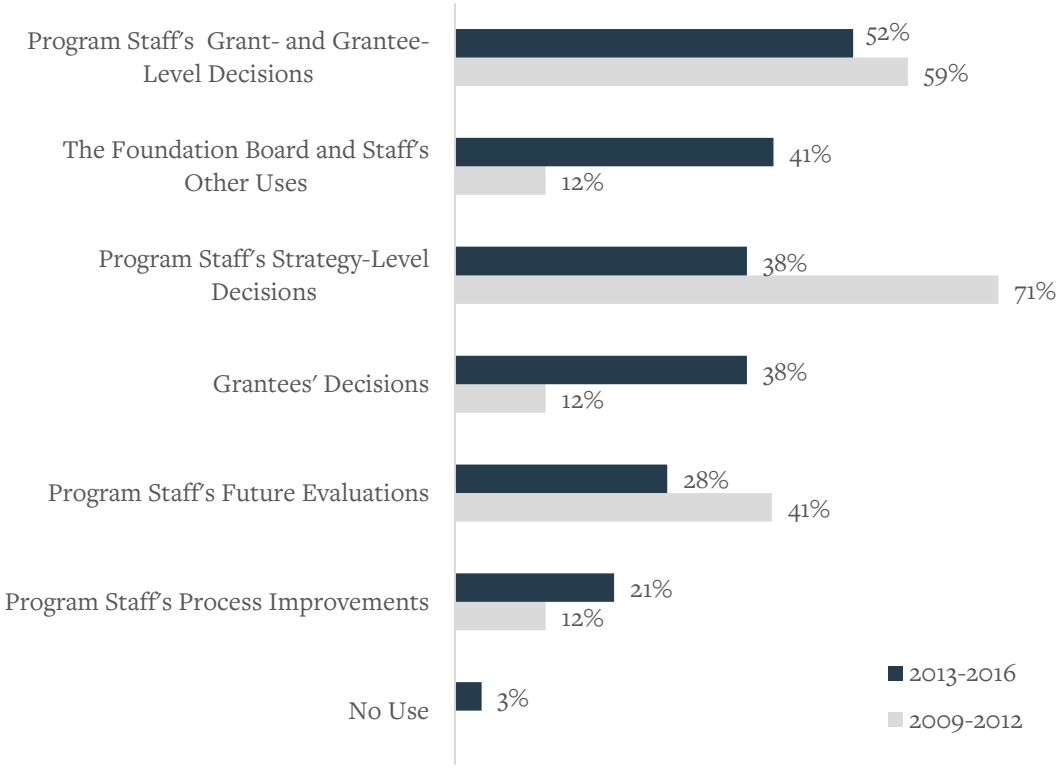
The amount evaluations are used has not changed much since 2013—it was a relatively high priority before 2013 and has remained so. That said, we do see some differences in the types of uses. For example, post-2013, more evaluations were commissioned at exit, used to "tell the story" about a strategy, to cultivate interest among other funders, to promote field learning, and/or to inform similar types of initiatives in the future (collectively referred to as "The Foundation Board and Staff's Other Uses" in the graph below). Relatedly, we find that a smaller proportion post-2013 have been used for strategy-level decisions. We do not find this to be alarming, since it can be explained in part because post-2013 we have commissioned more exit evaluations and smaller evaluations of sub-strategies, grant clusters, or individual grantees. Exit evaluations, in particular, can have broad application. For example, the exit evaluation of the Nuclear Security Initiative gathered important information on how we handled the exit, and those lessons have been incorporated into our Outcome-Focused Philanthropy guidance on preparing for exit.

A relatively small proportion of evaluations are actively or directly used by grantees, although this type of use has increased considerably since 2013. These evaluations typically treat grantees as co-learners. In these instances, the evaluator will often prepare individual reports comparing a grantee's data to others', and programs convene conversation with grantees and the evaluator to discuss findings and implications. As expected, the more

grantees are engaged in evaluation planning, implementation, and/or interpreting results, the more they use the evaluations to inform decisions.

Nearly all of our evaluations are being used, often in multiple ways.

Figures exceed 100% due to the possibility of demonstrating multiple categories.



For the most part, staff say that the timing of results has not been a key barrier for evaluation use. This finding stands in contrast to many complaints about evaluation in the broader field, where results are often received too late to make a difference.

> **One criticism of evaluation is that results often come too late to act upon. But that is in our control! There are trade-offs to keep in mind, but it is important to NOT sacrifice relevance by having evaluation findings be delivered too late to matter. -Evaluation Principles and Practices**

Yet, timing may have been a barrier to evaluation relevance or usefulness in other ways. For example, staff noted that the exit evaluations they commissioned had limited use for their own strategies, because findings come too late in the life of their strategy; that said, as noted above, these exit evaluations are valuable for other reasons. One staff regrets rushing the evaluation and decreasing the extent of data collection to meet what they understood as the "foundation's deadline for expending administrative funds in a given year." However, the foundation is actually quite flexible in spending guidelines, supportive of rolling over unused funds, and generally accommodative of the purpose of the work.

Numerous evaluations were described as "very useful." We highlight a few specific examples of use to give a flavor for the variety of uses program staff reference.

- In 2013, the Education program commissioned an evaluation of the policy research, design, and outreach portfolio in its Deeper Learning strategy. The Education staff described the evaluation,[29] conducted by ORS Impact, as providing "news you can use" about a "cluster" of policy grantees working together to better align their activities to address measurable, collective goals. Its findings were particularly valuable for informing Hewlett's grantmaking decisions. Indeed, an earlier evaluation of the full Deeper Learning strategy indicated the need to "pivot to practice" from an initial emphasis on the drivers of education systems—the bookends of policy and testing. Knowing that the budget for funding policy work would be decreasing the next year in order to make room for complementary grants focused on improving the education system's practice capacity, the program team commissioned the ORS Impact evaluation to assess progress to date and inform their next steps. Together, the evaluator and program staff visualized evaluation results inside a four-quadrant grid along a vertical axis of "grantee capacity to impact the policy field" and a horizontal axis of "alignment with the deeper learning strategy."[30] Staff used the findings to identify the grantees best positioned to successfully advance Deeper Learning-aligned policies—those with both strong capacity for policy impact and high alignment with Deeper Learning goals; for organizations in this quadrant, funding was shifted toward longer, larger, and more general operating support grants. Comparing the next two quadrants, the program staff ultimately decided to invest more in the high capacity/lower alignment quadrant than in the high alignment/low capacity quadrant—focusing on increasing strategic alignment and developing an effective coalition among "heavy hitters in the field" as opposed to trying to help low influence organizations become bigger, more impactful players. Lastly, the evaluation undergirded the staff's decision to provide "tie-off" grants to those in the low capacity/low alignment quadrant, so that the foundation could respectfully exit from relationships with grantees who, despite having other strengths and well-earned reputations, had proven less essential to the deeper learning strategy's initial success.

- In 2014, the Madison Initiative—which aims to help create the conditions in which Congress and its members can deliberate, negotiate, and compromise in ways that work for more Americans—commissioned a 3-year developmental evaluation of the entire strategy. The Madison team described the evaluation, undertaken by the Center for Evaluation Innovation, as very useful at different levels and for key decision points. First, at a broad level, the evaluator played a valuable role as "critical friend," helping the team to "sharpen and fine tune" their thinking and strategy. Early in the evaluation, for example, the evaluators worked with the team to "pressure test" with external

stakeholders some of the initially-defined framing aspects of the Madison Initiative—which led the team to shift the overarching goal from an emphasis on reversing polarization to a frame of helping political institutions cope with and adapt to polarization. Second, the evaluators also developed a systems map.[31] This map helped the Madison team establish a common understanding around what the initiative was trying to do, and the key variables both inside and outside of the Madison Initiative's funding areas. The map also provided a mechanism for sharing that information with the field, and was useful in grounding discussions with other funders. Seeing and working with the map helped the team to recast their thinking because they could see holes and gaps in different ways, leading the team to change the grantmaking avenues they pursued. Third, as the evaluators turned to investigating specific clusters of grantees working in different avenues, staff came away with more information about the relative merits and contributions of these, helping guide strategic pivots and grantmaking decisions. As one example, by early 2016, the Madison Initiative team had been investing in campaign finance data grantees for several years and "began wrestling with whether, in light of technological advances and all the talk about the big data revolution, foundation support for basic campaign finance data collection and curation was still necessary." Findings provided by the evaluator[32] firmed up the team's convictions that these grantees in fact play an important role in reform; the team the made the decision to support large, long-term grants to them. Finally, as the Madison team prepared a request to the board for renewal and expansion of the initiative in 2016, they were able to use the evaluation findings from over the course of the initiative's first three years as one important source and reference for suggesting the path forward.

- In 2015, the Performing Arts program conducted a midpoint summative evaluation of their overall strategic framework—which aims to sustain artistic expression and encourage public engagement in the arts in the Bay Area, through three strategies: Arts Education, Infrastructure, and Continuity and Engagement.[33] The team found the midpoint assessment of their overall strategy very valuable; it helped them understand that the strategy was basically on target for meeting its intended outcomes, but could be even more successful with some "fine tuning." For instance, because the evaluators, a partnership between Olive Grove and Informing Change, had compared the arts grantmaking data to others in the region, the team could see where they had strengths and weaknesses, and where they could diversify their portfolio to better reach the participants and artists they hoped to reach. Staff also learned that they needed to communicate more regularly with grantees and provide more opportunities for convening to cultivate new relationships and to spark ideas for working together.

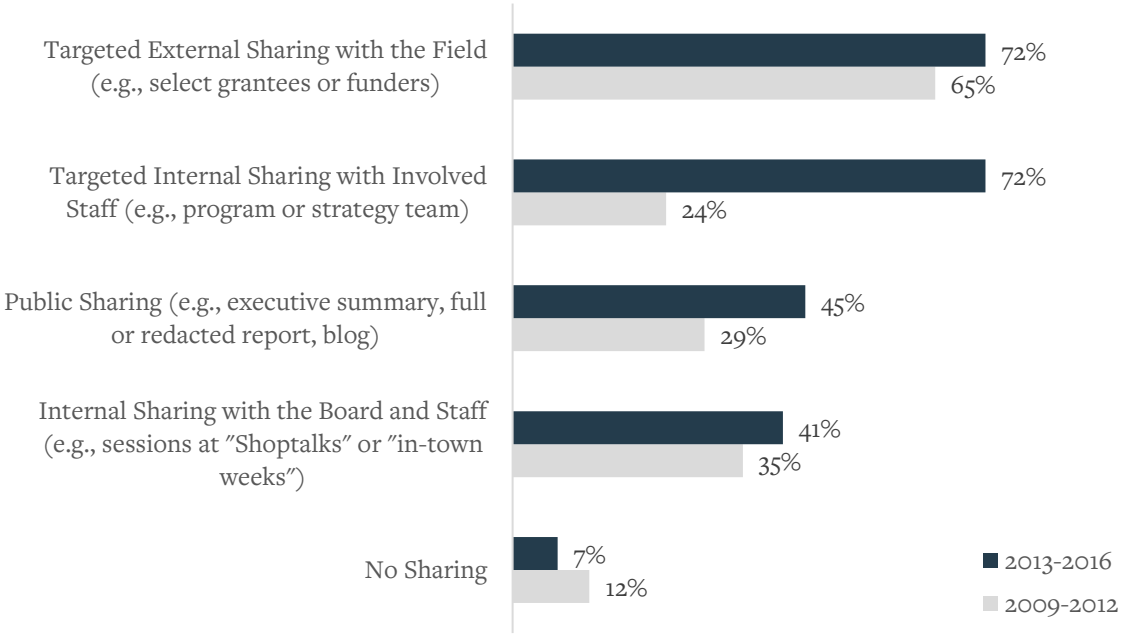
SHARING

Overall, we are sharing evaluation findings more than we used to—particularly with targeted internal audiences.[34] That said, there is still a lot of opportunity for improvement, as only 45% of our evaluations since 2013 have been shared broadly with the public.[35] This result is in line with other foundations included in the CEP/CEI foundation benchmarking study,[36] in which only 14% of respondents reportedly share evaluations with the public quite a bit or a lot, and to a Foundation Review survey, wherein 40% indicated sharing their evaluations on their websites. However, we think we can do a better job of aligning our actions with our foundation's principle of openness.

We presumptively share the results of our evaluations so that others may learn from our successes and failures. We will make principled exceptions on a case-by-case basis, with care given to issues of confidentiality and support for an organization's improvement.
-Evaluation Principles and Practices

While we are sharing our evaluations with targeted audiences, we can improve our public sharing.[37]

Figures exceed 100% due to the possibility of demonstrating multiple categories.



We have learned that higher quality evaluations are more likely to be more broadly shared. In addition, when we have spent more, we have tended to use the results in more ways, to share the results with more audiences, and to share a public version.

In order to improve our sharing, we must more deeply understand the obstacles—perceived or real—to sharing. In this assessment, we heard quite a bit that effective sharing requires a substantial investment of staff time, money, and up-front planning. A primary barrier to sharing is the time it takes program staff to think through how to frame evaluation findings such that they do not cause unintended harm (either to grantees or to the effectiveness of a particular strategy), or so that they are of interest to broader audiences. Another obstacle is

the evaluation products themselves. On the one hand, sometimes evaluation reports as written by evaluators can be lengthy and cumbersome to read. Other times, the evaluator provided a slide deck or PowerPoint, rather than a report. While these products might have been helpful and informative for the program staff, they were not then "ready" (either too detailed or too sparse) to be shared publicly or to offer meaningful content for external audiences. In some cases, program staff used additional resources to work with the evaluator or other external parties to develop a report for public sharing. In cases where staff did not share publicly at all, we heard that: 1) they had not planned to share, so there were no products developed appropriate for sharing, for example, without breaking confidentiality or scrubbing for sensitivities; 2) they felt that the sensitivities of the strategy couldn't lend themselves to a final product that could be shared.

Obstacles to sharing evaluations include: upfront investment in staff time and money; adequate planning; intentional framing of evaluation findings to not do harm and to be more accessible; and sensitivities of the strategy that limit sharing.

These challenges—real and perceived—can be discussed with your Communications Officer.

V. Other Ways to Improve the Value of Our Evaluations

Our basic assumption is that by spending a little more on evaluation—in proportional terms, 2% of program grant dollars—we can learn a lot about how we can increase the effectiveness of our grantmaking. We anticipated that by increasing spending on our evaluations we might also increase their quality and value. Overall, we found this to be true: we have increased spending and we see a corresponding increase in the overall quality of our evaluations. However, in alignment with what others in the field have noted[38] and our own Evaluation Principles and Practices guidance, we recognize that spending more money is not the only thing that matters for getting the most from the evaluations we commission.

This leads us to the question: what else matters most to foster evaluation quality and value?

Below, we summarize specific findings which provide useful insights on other factors that matter for improving evaluation quality and value. [39]

- **Program staff time.** Evaluations that involve greater amounts of program staff time tend to be *used* in more ways, and are *shared* in significantly more ways. When staff invest sufficient time—particularly during key junctures, such as when structuring the evaluation plan, engaging in deep discussion around data interpretation and insights, and adapting reports for specific audiences—there is a payoff in terms of the overall evaluation experience and utility.

- **Competitive selection process.** Although a formal Request for Proposal (RFP) is not a necessity, when program staff use a competitive process (even with just a few candidates), evaluation quality tends to be stronger. Evaluations initiated through a competitive process tend to more clearly articulate purpose, have much stronger evaluation questions, and are used and shared in significantly more ways than sole source evaluations. Of course, an RFP or other competitive process in itself does not *cause* quality. But, it is a signal of greater intentionality in the evaluation set up, of the selection of an evaluator with qualifications and an approach appropriate for the specific project, and of larger contracts, all of which are associated with quality.

- **Evaluator "fit."** Regardless of who is selected, it is critical that the program staff commissioning the evaluation ensure a strong fit with the evaluator and the overall approach to carrying out the evaluation activities; presentation style (e.g., tables and graphs, or narrative style reporting) is one area in which program staff expressed differences in what products were of most value to them.

- **Duration.** Evaluations with longer timelines usually cost more and tend to be of higher quality, as evidenced by sharper evaluative questions, more grantee engagement, and much more sharing. Several program staff regretted trying to shorten the evaluation timeframe because it caused them to shortchange the time needed to collect and reflect on the data, and also limited their use and dissemination. Ideally, the intended evaluation uses, audiences, and timing for decisions should drive duration, with cushion built in to address any challenges along the way.

- **Engaging grantees and advisory groups.** Evaluations which engage grantees at different stages of an evaluation—for example, by using evaluation advisory groups—are also associated with higher quality. In particular, evaluations that engage grantees are more likely to clarify purpose, develop strong questions, be shared more widely, and inform grantees' decisions. The small number of evaluations (six) that have involved advisory committees[40] tend to engage grantees to a greater extent, are used in more ways, and are significantly more likely to be shared. Involving relevant parties also reflects other foundation values, including transparency and collaboration.

VI. Recommendations

Given what we have learned through this assessment, we make the following set of recommendations, with the goal of further increasing the quality, practicality, and value of our evaluations, which in turn leads to better grantmaking.

1. **The foundation should use a benchmark of 1.5% to 2% spending on evaluation as a proportion of grant awards—recognizing that rates of spending on evaluation by specific programs and strategies will fluctuate, depending on where they are in a strategy lifecycle.** This analysis showed that increased spending did improve certain aspects of quality. Aiming for a 1.5 to 2% benchmark should be a helpful target for improving the quality of our evaluations in the areas in which we are not yet strong. We suggest that every strategy or sub-strategy begin at least one evaluation within a 3-year time period.

2. **We should focus on increasing the quality of our evaluations in two key areas—engaging grantees and sharing the findings.** Both the Evaluation Principles and Practices and the Hewlett Foundation Guiding Principles[41] stress the importance of these practices. For grantee engagement, we recommend building in the time it takes to involve grantees during the planning, implementing, and interpreting results phases of an evaluation. For sharing, we recommend working closely with Communications staff from the beginning of an evaluation to consider plans, especially for public distribution.

3. Finally, similar to our recommendation in the last board memo, we believe it is important to **continue to track evaluation quality, spending, and value, to ensure we are learning and adapting, and report back to the board in five years—in 2022.** We will also revise our Evaluation Principles and Practices paper based on the lessons we have learned from this analysis.

Appendices

A. Instrument: Quality Rubric

B. Data Gathered Through Evaluation Quality Assessment

C. Instrument: Interview Questions

D. Instrument: Pre-Interview Survey

E. Characteristics of Evaluations in Sample

A. Instrument: Quality Rubric

PRINCIPLE/ PRACTICE	1: NOT AT ALL	2: BARELY	3: SOMEWHAT	4: MOSTLY	5: EXCEPTIONAL
LEAD WITH PURPOSE: CLEAR AUDIENCE	Audience/ intended users are not mentioned.		Audience is mentioned but vague/unclear.		Audience is clearly articulated.
LEAD WITH PURPOSE: CLEAR USE	Use is not mentioned.		Use is mentioned but vague/unclear.		Use is clearly articulated.
LEAD WITH PURPOSE: TIMING/ WHEN FINDINGS NEEDED IS CLEAR	Timing of need for results is not mentioned.		Timing of need for results is vaguely mentioned (ie, list timetable but not rationale)		Timing of need for results for is well-articulated (ie, there is a timetable or specific points at which feedback from evaluation are noted, includes rationale for timetable)
STRENGTH OF EVALUATION QUESTIONS	No indication of the questions investigated.	Questions are mentioned but are not: ☐ Evaluative (how well did we do, why/why not/ how/to what extent/which ways)	Questions are mentioned and are: ☐ Evaluative (in a limited way) And one or two of the following: ☐Realistic (we can know within time frame and budget) ☐ Specific (spells out what we're interested in) ☐Aligned with the purpose (the questions reflect what we want to know and consider our audience)	Questions are mentioned and are: ☐ Evaluative-Moderate (how well did we do, why/why not/ how/to what extent/which ways) And one or two of the following: ☐Realistic (we can know within time frame and budget) ☐ Specific (spells out what we're interested in) ☐Aligned with the purpose (the questions reflect what we want to know and consider our audience)	Questions addressed are: ☐ Evaluative-Strong (how well did we do, why/why not/ how/to what extent/which ways) ☐Realistic (we can know within time frame and budget) ☐ Specific (spells out what we're interested in) ☐Aligned with the purpose (the questions reflect what we want to know and consider our audience)
RIGOROUS AND RELEVANT METHODOLOGY 1) good data quality and	0 of the following exist (all are unchecked):	1-3 of the following exist: ☐Complete and representative data (ie, sample	4 of the following exist: ☐Complete and representative data (ie, sample	5 of the following are true: ☐Complete and representative	All of the following exist: ☐Complete and representative data (ie, sample

PRINCIPLE/ PRACTICE	1: NOT AT ALL	2: BARELY	3: SOMEWHAT	4: MOSTLY	5: EXCEPTIONAL
adequate sample given questions; 2) more than one source of data, more than one viewpoint, and triangulation; 3) comparative data	☐Complete and representative data (ie, sample includes various perspectives, appropriate response rate) ☐ Data collection tools (interview tools, etc) are noted ☐More than one data source ☐Data are well-triangulated ☐ Data and methodology are relevant to questions (match evaluation purpose) ☐Evaluator articulates limitations ☐Comparative data (compared to another data source) ☐Comparative reference point (compared to an ideal or hypothetical benchmark)	includes various perspectives, appropriate response rate) ☐ Data collection tools (interview tools, etc) are noted ☐More than one data source ☐Data are well-triangulated ☐ Data and methodology are relevant to questions (match evaluation purpose) ☐Evaluator articulates limitations ☐Comparative data (compared to another data source) ☐Comparative reference point (compared to an ideal or hypothetical benchmark)	includes various perspectives, appropriate response rate) ☐ Data collection tools (interview tools, etc) are noted ☐More than one data source ☐Data are well-triangulated ☐ Data and methodology are relevant to questions (match evaluation purpose) ☐Evaluator articulates limitations ☐Comparative data (compared to another data source) ☐Comparative reference point (compared to an ideal or hypothetical benchmark)	data (ie, sample includes various perspectives, appropriate response rate) ☐ Data collection tools (interview tools, etc) are noted ☐More than one data source ☐Data are well-triangulated ☐ Data and methodology are relevant to questions (match evaluation purpose) ☐Evaluator articulates limitations AND (at least one of the following) ☐Comparative data (compared to another data source) ☐Comparative reference point (compared to an ideal or hypothetical benchmark)	includes various perspectives, appropriate response rate) ☐ Data collection tools (interview tools, etc) are noted ☐More than one data source ☐Data are well-triangulated ☐ Data and methodology are relevant to questions (match evaluation purpose) ☐Evaluator articulates limitations AND (both of the following): ☐Comparative data (compared to another data source) ☐Comparative reference point (compared to an ideal or hypothetical benchmark)
CLEAR INTERPRETATION OF FINDINGS FROM THE EVALUATOR	Evaluator does not translate or apply meaning to data.		Evaluator analyzed data to determine meaning, but insights were lacking/not very informative.		Evaluator effectively analyzed data and determined key insights.
IMPLICATIONS/RECOMMENDA-TIONS ARE REALISTIC AND BASED ON THE FINDINGS (if included)	Evaluator presented implications or recommendations that were not usable/not based on		Evaluator presented implications or recommendations that were not all usable or all based on		Evaluator presented well-sounded implications or recommendations that are useful,

PRINCIPLE/ PRACTICE	1: NOT AT ALL	2: BARELY	3: SOMEWHAT	4: MOSTLY	5: EXCEPTIONAL
	findings.		findings.		actionable given context/audience and are based on the findings, and if necessary prioritized.
ENGAGEMENT WITH GRANTEES	Grantees not informed about evaluation.	No grantee involvement in any of the following stages: ☐ Planning the evaluation (Review RFP, add questions, help select evaluator) ☐ Implementation of the evaluation (review methodology, etc) ☐ Results (findings shared, discussed, involved in interpretation)	Grantee engagement in only one of the following stages: ☐ Planning the evaluation (Review RFP, add questions, help select evaluator) ☐ Implementation of the evaluation (review methodology, etc) ☐ Results (findings shared, discussed, involved in interpretation)	Grantees involved in at least 2 stages: ☐ Planning the evaluation (Review RFP, add questions, help select evaluator) ☐ Implementation of the evaluation (review methodology, etc) ☐ Results (findings shared, discussed, involved in interpretation)	Grantees involved in all of the following stages: ☐ Planning the evaluation (Review RFP, add questions, help select evaluator) ☐ Implementation of the evaluation (review methodology, etc) ☐ Results (findings shared, discussed, involved in interpretation)
USE OF THE DATA (FOR LEARNING/ COURSE CORRECTION)	Results are not used at all		Findings used as intended but limited ☐ The foundation's improvement to strategy ☐ The foundation's improvement to process ☐ The foundation's grant-level decision making ☐ The foundation's learning and engagement (including Board and staff) ☐ The foundation's future evaluations (e.g., pulse		Findings used as intended and fully ☐ The foundation's improvement to strategy ☐ The foundation's improvement to process ☐ The foundation's grant-level decision making ☐ The foundation's learning and engagement (including Board and staff) ☐ The foundation's future evaluations (e.g., pulse

PRINCIPLE/ PRACTICE	1: NOT AT ALL	2: BARELY	3: SOMEWHAT	4: MOSTLY	5: EXCEPTIONAL
			taking, baseline setting) ☐ Grantees' learning and decisions ☐ Other		taking, baseline setting) ☐ Grantees' learning and decisions ☐ Other
SHARING WHAT WE ARE LEARNING (BOTH INTERNALLY AND EXTERNALLY)	No sharing	Limited sharing At least one of the following ☐Targeted internal sharing with involved staff (e.g., strategy or program team) ☐Broad internal sharing with the Board or staff (e.g., sessions at Shoptalks or in-town weeks) ☐Targeted external sharing with the field (e.g., e-mail, webinar, conferences)	Some internal and external sharing ☐Targeted internal sharing with involved staff (e.g., strategy or program team) ☐Broad internal sharing with the Board or staff (e.g., sessions at Shoptalks or in-town weeks) ☐Targeted external sharing with grantee(s) involved ☐Targeted external sharing with the field (e.g., e-mail, webinar, conferences) ☐Broad external sharing with the public (e.g., executive summary, full or redacted report, blog)	Both targeted and broad internal and external sharing ☐Targeted internal sharing with involved staff (e.g., strategy or program team) ☐Broad internal sharing with the Board or staff (e.g., sessions at Shoptalks or in-town weeks) ☐Targeted external sharing with grantee(s) involved ☐Targeted external sharing with the field (e.g., e-mail, webinar, conferences) ☐Broad external sharing with the public (e.g., executive summary, full or redacted report, blog)	All of the following ☐Targeted internal sharing with involved staff (e.g., strategy or program team) ☐Broad internal sharing with the Board or staff (e.g., sessions at Shoptalks or in-town weeks) ☐Targeted external sharing with grantee(s) involved ☐Targeted external sharing with the field (e.g., e-mail, webinar, conferences) ☐Broad external sharing with the public (e.g., executive summary, full or redacted report, blog)

B. Data Gathered Through Evaluation Quality Assessment

In addition to scoring the evaluations according to the evaluation quality rubric, the following information was collected from the evaluation contracts, the documents reviewed, and/or the interviews and survey of the program staff representing each evaluation.

- Program area (e.g., Environment, Education, etc.)
- The name of program staff with whom we spoke
- Ownership of the evaluation (i.e., were the staff who managed/received the results the ones who commissioned the evaluation or was it someone else?)
- Evaluator selection via competitive or sole source process
- Duration of the evaluation (i.e., date contract signed to date contract ended)
- Years of strategy/intervention covered by the evaluation
- Contract amount (including amendments)
- Evaluation type (i.e., formative, summative or exit)
- Types of measurement/information covered by the evaluation questions (e.g., process, outcome, impact)
- Involvement of an evaluation advisory committee
- Unit of evaluation (i.e., strategy, grant cluster, individual grantee)
- Approximate number of grantees included in the evaluation
- Internal or domestic grantees included
- Involvement of other funders
- Other funders' financial contribution, if relevant
- Evaluator name
- Evaluator type (i.e., larger firm, smaller firm or individual consultant)
- The program staff's likelihood of recommending the evaluator (and why or why not)
- The program staff's perception of whether the timing of the evaluation mattered or was a barrier for the results/use
- The percent of time the program staff who commissioned the evaluation spent on the evaluation
- The program staff's perception of whether the budget influenced the quality of the evaluation
- The program staff's perception of what else would have been helpful for the evaluation

C. Instrument: Interview Questions

We are compiling information from a review of evaluation documents and interviews with program staff into an assessment of the quality, practicality and usefulness of our evaluations over the last 5 years. This is part of our foundation-wide effort to track our evaluation spending. Our goal with this interview is to learn from you more about the _____ evaluation. We are particularly interested in learning about the usefulness of the evaluation, and your experience and insights. We've read the _____ documents, so this is an opportunity to go beyond what is captured in them.

1) *Time spent on evaluation:*
   i. Once the contract was underway, about what percent of your time did you devote to managing the evaluation? How did this vary across the phases of the evaluation?
2) *Clarity of evaluation purpose:*
   i. Do you feel that you had a clear sense of the audience, use, and plan for sharing before you commissioned the evaluation?
   ii. If not, did the evaluators help you with this? How?
3) *Engagement with grantees:*
   i. Were grantees involved in evaluation planning or use? *(yes/no)*
   ii. If yes, at which points: during the development of the evaluation questions? RFP? discussion of interim findings? discussion of findings after the evaluation was complete? Any other ways?
4) *Overall quality (rigor, methods):*
   i. On a scale of 0 to 10 (not meaningful to most meaningful), how would you rate the overall interpretation of data?
   ii. On a scale of 0 to 10 (not useful to most useful), how would you rate the recommendations that the evaluator presented?
5) *Use of results for learning/course correction:*
   i. How and when did you use the findings from this evaluation? (consider decisions/changes to grantmaking, strategy, capacity building efforts, data collection, influencing others, informing refresh, as an input for other evaluation, improving likelihood of outcomes, etc.)
   ii. What changes did the grantees make as a result, if you are aware of any?
   iii. If you didn't use the evaluation findings (or use was limited), why was that?
6) *Timing:*
   i. Was timing a factor at all for use, for this evaluation? That is, did evaluation findings come in time to provide useful insights?
   ii. If not, was the delay in planning or implementation—i.e., was the evaluation planned early enough to allow for timely findings? Was the delay due to logistical/implementation delays after the consultant was selected? Please describe.
7) *Sharing findings both internally and externally:*
   i. With whom and how did you share the findings internally? (team meeting, cross-program session, shoptalk, board, strategy memo, etc.?)
   ii. With whom and how did you share externally? (blog, webinar, one on one with grantee, etc.?)
   iii. What did you share? (i.e., exec summary, full report, redacted, video, PowerPoint?)
   iv. Did you plan/talk with the evaluator about sharing?
8) What else would have been helpful for you if you had to do the evaluation again? What changes would you have made?

D. Instrument: Pre-Interview Survey

Name of Evaluation:

1) *Ownership:* Did you decide to commission this evaluation? If not, who did?

2) *Staff transition:* Did any staff transitions at Hewlett affect the quality and use of this evaluation? If so, how?

3) *Quality of Evaluator:* On a scale of 0-10, how likely would you be to recommend this evaluator? Why or why not?

4) *Quality/budget relationship:* How much of a role do you think budget played in the ultimate quality of this evaluation?

5) *Timing (Within strategy):* Where does this evaluation fall in your strategy's lifecycle? Please name the strategy, and note when this **and** other evaluations (if any) took place.

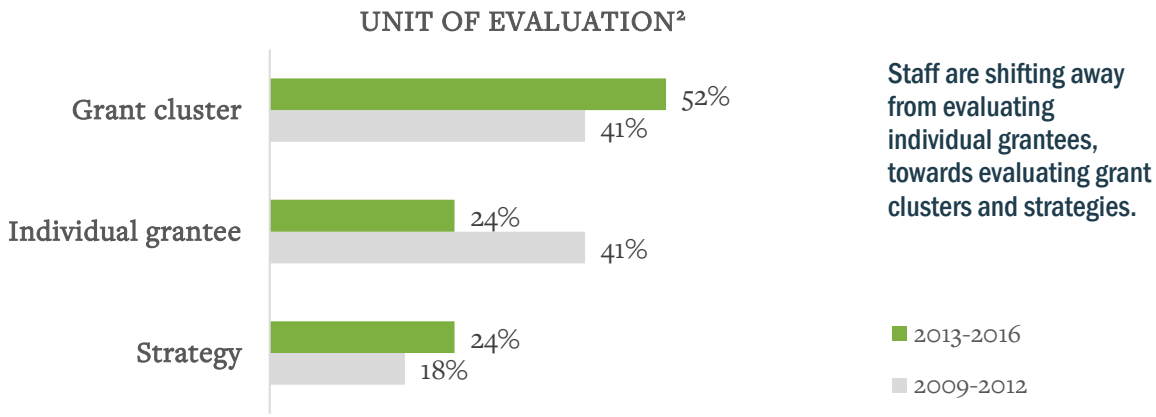
Strategy and Approximate Year Grantmaking Began: _____
Year/Name of Evaluation:_____
Year/Name of Evaluation:_____
Year/Name of Evaluation:_____

E. Characteristics of Evaluations in Sample

WHAT IS BEING EVALUATED?

Choosing what to evaluate is a selective process, as staff cannot (and need not) evaluate everything; one of the foundation's seven evaluation principles, "we choose strategically what to evaluate," urges staff to think carefully about where to spend their time and resources. The foundation's guidance suggests "several criteria to guide decisions about where to put our evaluation dollars, including the opportunity for learning; any urgency to make course corrections or future funding decisions; the size/dollar amount of our investment as a proxy for importance, the potential for strategic or reputational risk."

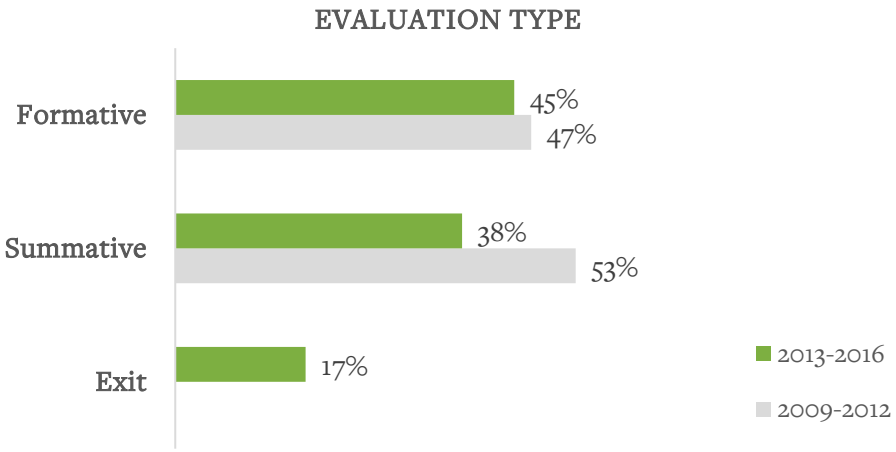
With these factors in mind, a comparison of evaluations commissioned 2009 through 2012 to those commissioned 2013 to 2016 indicate that program staff have begun to shift away from evaluating individual grantees (41% to 24%), mostly in favor of evaluating grant clusters (41% to 52%), which are typically two to ten grants that share key characteristics (often, grants within a sub-strategy). Strategy evaluations (typically more than ten or twenty grantees) account for the remaining proportion (18% to 24%). Evaluations of strategies have the largest contracts on average ($166K),[1] while grantee and grant cluster evaluations have lower average contract amounts ($120K and $109K, respectively). The individual grantees selected for evaluation tend to be network conveners, research hubs or other types of intermediaries, which have usually received multiple foundation grants over time—and which often operate essentially as a substrategy.



[1] Of note, evaluations of strategies in the sample do not use individual consultants, and instead opt for larger evaluation organizations—in part explaining the higher contract amounts.

[2] Initiative-level evaluations (e.g., the evaluation of the Nuclear Security Initiative) are categorized as "strategy" evaluations, based on the foundation's Outcome-Focused Philanthropy guidance.

We tend to commission three types of evaluations: formative (ongoing, more real-time evaluation, intended to inform adjustment), summative (at the conclusion or inflection point in a strategy to inform grant decisions or strategy refresh), and exit (to "tell the story" at close-out). There are differences in the amount of each type we commissioned in each time period. The proportion of summative evaluations is lower in the 2013 to 2016 period, likely paralleling the maturity of strategy or grants. The proportion of formative evaluations is basically the same in both time periods.[3] Overall, we have tended to spend more on exit evaluations ($156K) in comparison to what we have spent on formative or summative evaluations ($132K and $123K, respectively).[4]



Regardless of evaluation type, there is a strong interest in measuring outcomes, including: what changes have occurred, to what extent, why, how, and with what contribution from the grantees and the foundation. Fewer evaluations include process measures, addressing how strategies or grants have been implemented and how well; the evaluations that include such measures are typically formative. Evaluations that include methodology to assess impact (e.g., the extent to which changes can be attributed to grantees or the foundation) are rare, and generally fall under the exit category.

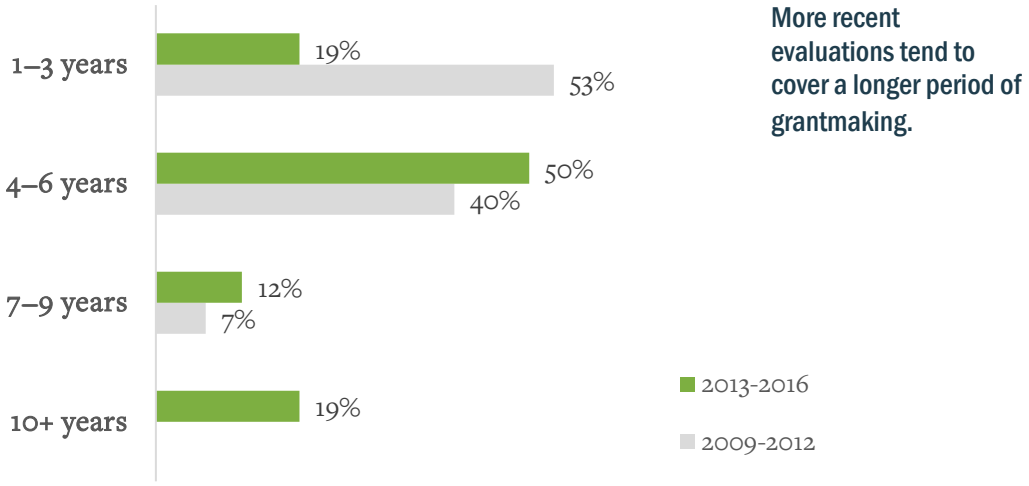
Evaluations can cover a broad time period, from one year to more than a decade. Evaluations conducted 2013 to 2016 cover longer periods than those commissioned between 2009 and 2012 (6 years and 3.5 years on average, respectively). This difference is partly due to the fact that the latter group of evaluations includes several exit evaluations, which tend

[3] The formative category includes baseline assessments and a developmental evaluation.

[4] Three of the exit evaluations are from the same initiative's exit, but in different countries, using different evaluators. These are treated as three individual contracts when looking at quality but are summed up here into one contract amount to calculate the average cost of exit evaluations.

to cover about twice as many years as summative or formative evaluations, and 2009 to 2012 did not include any.[5]

YEARS OF STRATEGY OR GRANTS COVERED IN EVALUATIONS

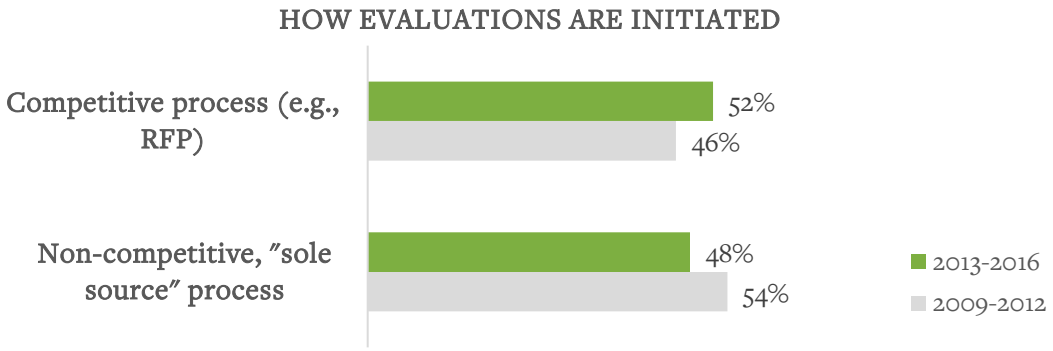


HOW DO WE INITIATE EVALUATIONS?

About half of the evaluations commission evaluators through a more formalized, competitive process, typically through a Request for Proposal (RFP). Competitive selection processes are more likely to be used for strategy evaluations (e.g., prior to a strategy refresh). Evaluations initiated through a competitive selection process also tend to have much higher contracts on average ($178K) than those commissioned via sole source ($69K).

The other half are initiated through a non-competitive, "sole source" process, meaning that program staff typically identify the evaluator they want to work with based on their own or their colleagues' experiences or recommendations. Sole source processes are used more frequently in the Environment and Education programs, particularly with "repeat evaluators" who have conducted previous evaluations with those programs.

[5] We note that two publicized evaluations commissioned at exit--one of our Neighborhood Improvement Initiative and the other of our Conflict Resolution initiative--took place earlier than the time period included in this assessment.



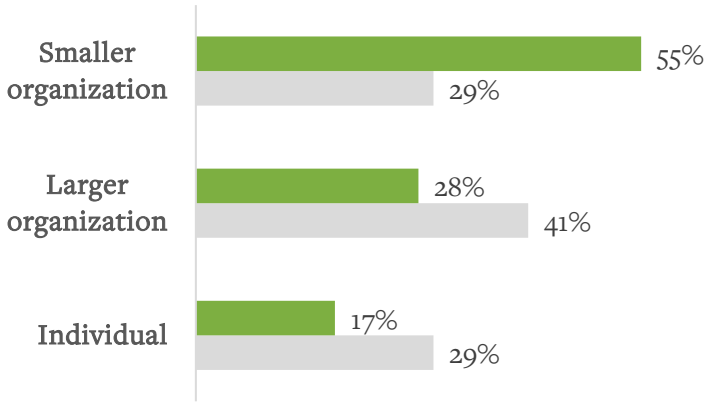
WHO CONDUCTS THE EVALUATIONS?

When program staff decide to commission an evaluation, they have the opportunity to make choices which will ultimately affect the evaluation experience and its usefulness. One important choice is who to select as an evaluator: an individual consultant, a smaller "boutique" organization, or a larger, national or international organization. Tradeoffs exist with each choice; moving forward with an individual consultant may allow program staff to choose someone who specializes in the program's content or who is more agile to work under a short timeline, while an evaluation organization may offer the benefit of additional staff capacity or a broader methods skill set. Evaluator selection often depends on the size and nature of the project, as well as staffs' personal experiences and preferences.

Evaluations conducted by larger organizations cost the most (on average $199K), followed by smaller organizations ($111K). When we have more ambition for an evaluation (e.g., engage more grantees, plan to share the results), we tend to put more money into the contract—and want to make sure we are getting the evaluation team with the best capacities to meet our needs. The average contract amount is much smaller for individual evaluation consultants ($28K). The costs associated with different types of evaluators do not appear to be driving the choices; staff generally say that the available evaluation resources enable them to hire their desired evaluator. Program staff are selecting organizations to conduct most of the evaluations. Over time, staff are increasingly selecting smaller organizations rather than individual consultants or larger organizations.[6]

[6] In a few cases, multiple individuals or independent consultants collaborate on an evaluation. These cases are considered "smaller organizations" for this graph because the evaluator configuration is similar to project teams at a smaller organization.

TYPE OF EVALUATOR



Staff are increasingly selecting smaller organizations to conduct their evaluations.

In total, 26 evaluators have conducted the 46 evaluations in the "quality sample;" 62% of the evaluators have conducted just one evaluation for the foundation, 27% have conducted two evaluations and 11% have conducted three or more evaluations. When the same evaluator is chosen for multiple projects, it is often because program staff have received a recommendation from their program peers or because they worked well with the evaluator and opted to use them again. Staff often want to continue working with someone with whom they have had a positive experience in the past. Upon completion of the evaluation, lead staff from the majority (66%) of the evaluations say they would recommend their evaluator. Another 15% of staff might recommend the evaluator and 19% would not recommend.[7]

The ideal evaluator is strong technically, has subject matter expertise, is pragmatic, and communicates well, both verbally and in writing. Often in our work, cultural awareness and sensitivity to the context in which nonprofits are operating are also very important. -Evaluation Principles and Practices

Ultimately, staffs' choice about who to work with, and their perceptions about the evaluation experience, is often related to "fit." For instance, program staff have different preferences and communication styles (e.g., regarding deliverables: some staff enjoy seeing tables and charts in an evaluation report, while others prefer high-level findings). While there is some room to tailor interactions during the evaluation, being aware of these preferences up front will help staff select an appropriate evaluator and maximize the value of their evaluation.

In line with the foundation's notion of "ideal" evaluators, most program staff are seeking—and finding—strong evaluators who:

[7] This variable captures the likelihood of recommending an evaluator, not the actual proportion of recommendations made. A few staff say their recommendation may be specific to the type of project, methods or field. The willingness to recommend is based on the lead staff's experience from each individual evaluation. Evaluators who conducted multiple foundation evaluations may have received multiple/duplicative "would recommend" responses.

- Ask good questions and are good listeners
- Engage different groups well (e.g., connected, culturally competent, tactful)[8]
- Have relevant expertise (e.g., nonprofits, strategy development) and/or experience working with the foundation
- Are timely and meet deadlines
- Manage the process well, including staffing the project appropriately and managing relationships with foundation staff
- Are rigorous and thorough, sharing in-depth findings and honest, critical feedback
- Use custom and tailored approaches
- Offer useful perspectives, new insights and thought partnership
- Are credible and held in high regard by trusted sources inside and outside of the foundation[9]

Evaluators that do not earn recommendations or positive reviews are generally noted as lacking in such qualities as noted above (e.g., they are poor listeners, do not fully understand the work/context, lack creativity, or need considerable administrative assistance). Of note, criticisms are generally not related to evaluators' technical skills.

HOW ARE THE EVALUATIONS MANAGED?

For about three-quarters of the evaluations, for both time periods, the program staff who commission the evaluations oversee them for the duration; for the remaining one-quarter, the responsibility for oversight and eventual use of the findings shifts to other staff.[10] The shift in ownership is primarily due to program staff leaving the foundation. When staff reflect on ways the evaluation experience could be improved, some recommend better managing staff transitions, and being more intentional about which staff are involved in the evaluation and when.

Many staff estimate that they spend somewhere between 5% to 20% of their time managing an active evaluation, though the actual time commitment varies considerably depending on the evaluation duration, phase of work, and type of project. The proportion of staff time spent on evaluations is comparable pre- and post-2013. Managing the evaluation entails communicating with the evaluator, tracking budget and timeline, helping engage grantees

[8] One staff member says it was particularly valuable having an evaluator local to where the evaluation was taking place, as they knew who to talk to and what questions to ask given the context.

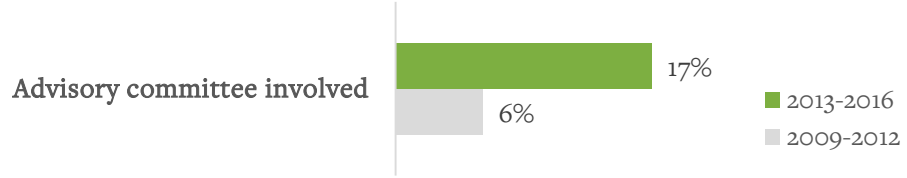
[9] The qualities of strong evaluators are based on staffs' experience with evaluators in the "quality sample."

[10] In a few cases, (the idea for the) evaluations are initiated by someone other than (or in addition to) the Program Officer who manages the evaluation, such as Program Directors, other team members, the foundation President or Board.

and other informants, reflecting with the evaluator around the findings, providing direction and input on draft documents, and more. As stated in the Evaluation Principles and Practices paper: "Active management is essential…exchanges [with the evaluator] can be useful forcing functions to keep an evaluation on track and to start troubleshooting early." Some program staff suggest that the time they spend working with the evaluator—to discuss what they are finding, to help provide context for the findings and to consider different interpretations and insights—can be even more valuable than an evaluation report.

Program staff have involved an evaluation advisory committee on only a few evaluations, though the use of such groups has increased over time. Evaluations that involve advisory committees tend to have much more expensive contracts on average ($226K) than those which do not involve these entities ($108K). The Global Development & Population program's more frequent use of advisory committees for international evaluations (which are also typically more expensive than domestic evaluations) may be one factor in explaining their higher average cost. The Effective Philanthropy Group has also used advisory committees for all of their evaluations, including of their Knowledge and Organizational Effectiveness work.

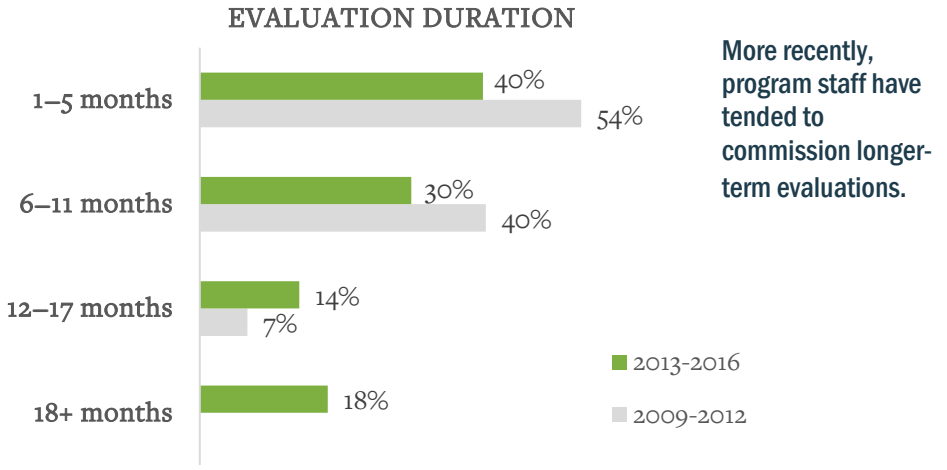
INVOLVEMENT OF EVALUATION ADVISORY COMMITTEE



Advisory committees have been used to: promote rigor and relevance (e.g., provide input on the evaluation design and data collection instruments, include diverse perspectives); increase evaluation buy-in (e.g., by including grantee representatives on the committee); broaden the communication of results (e.g., by sharing early results in meetings); and, cultivate interest from potential funders (e.g., by including other current or potential funders on the committee). While managing an advisory committee takes additional time and resources to engage and coordinate, our analyses suggest that these investments typically pay off as intended.

HOW LONG ARE THE EVALUATIONS?

Evaluation duration varies widely, with the shortest evaluation taking 1 month and the longest taking 32 months.[11] Evaluations commissioned since 2013 typically last longer than pre-2013 evaluations (an average of 9 months and 6 months, respectively). Evaluation duration and contract amount also have a strong positive correlation—longer evaluations and larger contracts tend to go hand in hand. Although there is no one "right" length for an evaluation, some program staff who commissioned short evaluations wished they had allowed more time for interviewing informants, obtaining the most valuable data, or considering the findings and potential implications with intended audiences.



[11] The median value is used for all foundation evaluations from 2009 to 2016 for which data are available (n=66). The contract start and end dates were used for this calculation. A 2009 and a 2010 evaluation are missing specific contract start and/or end dates, so their duration could not be calculated.

Endnotes

[1] See Benchmarks for Spending on Evaluation, a memo written for the Hewlett Foundation board in Fall 2013 and available on the foundation website at http://www.hewlett.org/library/benchmarks-for-spending-on-evaluation/

[2] In CEP/CEI's Benchmarking Foundation Evaluation Practices (2016), the authors estimate that, in general, the median foundation spends approximately $1 on evaluation for every $100 program dollars, or 1%. When they break that down by smaller and larger foundations, the proportion spent on evaluation for those at over $200 million in giving is 2.75% (median $5.5 million). Yet, we also know that foundations struggle with estimating their spending on evaluation. As noted in the report: "Evaluation spending is notoriously difficult for foundations to estimate, however, and only 35 percent of respondents were quite or extremely confident in the accuracy of their estimates." Importantly, many foundations include large-scale impact evaluations for the field in their total—which significantly drives up the spending estimates. Taking this range, and the fact that we are excluding large-scale impact evaluations for the field, the 2% target continues to make sense for the foundation.

[3] See the Hewlett Foundation Guiding Principles at https://www.hewlett.org/about-us/values-and-policies/

[4] See endnote 1.

[5] The 2% recommendation is based on an informal survey the foundation conducted in 2013 of the evaluation spending of 10 large/peer foundations, along with an analysis of what we might spend realistically, given the size of our program grants and the relatively large grant portfolios handled by lean program staffing.

[6] For simplicity, we use the term "quality" throughout this report to refer broadly to evaluation practicality, usefulness, value, rigor, relevance and more.

[7] See: Benchmarking Foundation Evaluation Practices (2016) from the Center for Effective Philanthropy (CEP) and the Center for Evaluation Innovation (CEI); and, How Do You Measure Up? Finding Fit Between Foundations and Their Evaluation Functions (2016) from CEI.

[8] The sample size for each variable in the report's "quality sample" is ≤ 46 evaluations; please use caution when interpreting results since the sample size for sub-groups may be relatively small. We used statistical tests to help identify meaningful relationships and differences.

[9] In CEP/CEI's Benchmarking Foundation Evaluation Practices (2016), the authors estimate that, in general, the median foundation spends approximately $1 on evaluation for every $100 program dollars, or 1%. When they break that down by smaller and larger foundations, the proportion spent on evaluation for those at over $200 million in giving is 2.75% (median $5.5 million). Yet, we also know that foundations struggle with estimating their spending on evaluation. As noted in the report: "Evaluation spending is notoriously difficult for foundations to estimate, however, and only 35 percent of respondents were quite or extremely confident in the accuracy of their estimates." Importantly, many foundations include large-scale impact evaluations for the field in their total—which significantly drives up the spending estimates. Taking this range, and the fact that we are excluding large-scale impact evaluations for the field, the 2% target continues to make sense for the foundation.

[10] Although we did not track annual evaluation spending prior to 2013, analysis of the contracts we gathered for the years 2009 to 2016 indicate an increase in mean contract amount from pre-to post-2013 ($126K to $156K).

[11] Under IRS rules, DCA expenses count as part of a foundation's qualifying distributions and must be reported annually on IRS Form 990-PF. For evaluation contracts, DCA funds are typically used for evaluations that provide feedback to and help inform improvements for grantees, or where findings are shared publicly.

[12] http://effectivephilanthropy.org/wp-content/uploads/2016/09/Benchmarking-Foundation-Evaluation-Practices.pdf

[13] For any given variable, the sample size for the "quality sample" is ≤46; there are ≤29 evaluations in the post-2013 group and ≤17 evaluations in the pre-2013 group. Graphs containing quality ratings are based on a 5-point scale in the quality rubric (a rating of 4 or 5 is "strong," a 3 is "moderate," and a 1 or 2 is "weak").

[14] We also looked at key characteristics of our evaluations (e.g., what was evaluated, by whom) and examined how these characteristics have changed, as context for understanding the quality and spending analyses. The findings from those analyses are included in Appendix E.

[15] See Outcome-Focused Philanthropy at http://www.hewlett.org/wp-content/uploads/2016/12/OFP-Guidebook.pdf
[16] Two of these evaluations are included in the spending analysis sample, but not the quality analysis sample, because we did not have a final report by August 2016.
[17] http://www.hewlett.org/library/bringing-learning-to-light-the-role-of-citizen-led-assessments-in-shifting-the-education-agenda/
[18] http://www.hewlett.org/taking-stock-of-our-performing-arts-grantmaking/
[19] http://www.hewlett.org/deeper-learning-six-years-later/
[20] http://www.hewlett.org/library/the-william-and-flora-hewlett-foundations-nuclear-security-initiative-findings-from-a-summative-evaluation/
[21] Although almost all evaluations have data and methodology relevant to evaluation questions, this does not necessarily mean that the methodology or evaluation questions are of high quality.
[22] One evaluation noted as using comparative reference points also used comparative data (i.e., study and control groups with random assignment).
[23] http://www.hewlett.org/evaluation-philanthropy-knowledge-creation/
[24] http://www.hewlett.org/peer-to-peer-at-the-heart-of-influencing-more-effective-philanthropy/
[25] One grantee praised this effort to authentically engage grantees in shaping an evaluation that would benefit both the foundation and its grantees.
[26] http://www.hewlett.org/wp-content/uploads/2016/10/Evaluation-of-OE-Program-November-2015.pdf
[27] Evaluators' interpretations and recommendations that appear to be appropriate and useful in evaluation documents may not always align with the perceptions of staff, who have more context and experience with the strategy, grantee work and evaluation.
[28] http://www.hewlett.org/supporting-innovations-in-learning-hits-misses-and-advice-for-funders/
[29] http://www.hewlett.org/library/deeper-learning-advocacy-cluster-evaluation/
[30] http://www.hewlett.org/wp-content/uploads/2016/08/Grantmaking%20Trends%20Memo_Education_2014.pdf
[31] http://www.hewlett.org/the-madison-initiatives-view-of-the-world-version-1-0/
[32] http://www.hewlett.org/funding-campaign-finance-data-critical/
[33] http://www.hewlett.org/taking-stock-of-our-performing-arts-grantmaking/
[34] Staff are generally sharing findings directly with the grantees involved/assessed in the evaluation. The few evaluations that were not shared with grantees were typically summative evaluations without any grantee engagement. Of note, sharing evaluation results does not equate with engaging grantees in the evaluation or using the results, as discussed in earlier sections of this report.
[35] Among those evaluations shared publicly, there are differences pre- and post-2013 in what has been shared. Of the 29% pre-2013 shared publicly, 20% shared the full report, 20% an edited or redacted report and 60% shared an executive summary only. Of the 45% post 2013 the proportions are 54% full, 46% edited or redacted, and 0% executive summary only.
[36] http://effectivephilanthropy.org/wp-content/uploads/2016/09/Benchmarking-Foundation-Evaluation-Practices.pdf
[37] The graph excludes sharing with the grantees involved. Three evaluations in the "No sharing" category shared only with grantees involved; one of the evaluations did not share at all.
[38] http://www.betterevaluation.org/en
[39] We examined a few other factors that we thought would be related to evaluation quality: Foundation program area, what is evaluated (i.e., grant, cluster, strategy) and type of evaluation (formative, summative, exit). None of these factors are meaningfully related to quality, in part because some of the variables' sub-groups are too small to assess significant differences. We did find that where there are quality variations, different programs shine in different ways.
[40] These advisory groups tended to be comprised of other funders, regional or content area experts, and grantee representatives.
[41] See the Hewlett Foundation Guiding Principles at https://www.hewlett.org/about-us/values-and-policies/