Summative Evaluation Report on STEM Evaluation Community

Submitted by
Alexis Kaminsky, Ph.D
Kaminsky Consulting, LLC

June 11, 2020

This work was made possible by NSF grant: OIA-1650215. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

BACKGROUND

In 2017, the Education Development Center (EDC) received a grant from the National Science Foundation (NSF) to bring together PIs of STEM Program Resource Centers (PRC) funded under NSF's Education and Human Resources (EHR), evaluators of NSF STEM projects and programs, and evaluation advisors to address concerns about the quality and consistency of STEM evaluations. According to the grant proposal, the goal was "to increase the capacity of evaluators to produce high quality, conceptually sound, methodologically appropriate evaluations of NSF programs and projects, specifically in the area of STEM education and outreach" (EDC Grant Proposal, 2017). Primary activities undertaken to this end included three 2-day meetings with members of what became the STEM Evaluation Community (STEM EC) and production of two studies (a landscape study of the STEM Evaluation resources and an evaluator survey) between December, 2017 and July, 2019.[1]

The evaluation of the STEM EC project was designed as a Responsive Evaluation (Abma, 2006; Greene and Abma, 2002; and Stake, 2003). As a responsive evaluator, I sought to be a critical friend, asking questions and identifying tensions that could affect the value created through the project through the reports produced and in conversations with the EDC team.

I applied Wenger, deLaat, and Traynor's 2011 conceptual framework for assessing value creation in Communities of Practice (CoP) to characterize the value that accrued to members and the STEM EC community as a whole over the course of the project. This conceptual framework lays out five cycles of value creation possible within communities of practice and networks: (1) Immediate value or the activities and interactions members engage in; (2) Potential value or knowledge capital; (3) Applied value or changes in practice; (4) Realized value or performance improvement; and finally, (5) Reframed value or redefining what is valued and what counts as success. These cycles of value creation are neither linear nor hierarchical, and project success is not defined by achieving reframed value. Rather, different types of value that derive from a CoP depend on where it is in its maturity, the nature of the problem it's been formed to address, its membership, and the activities and interactions undertaken.

In this final evaluation report, I directly address the two questions guiding the evaluation:

1. What types of value creation do the STEM EC activities and strategies yield? and
2. What evidence do we have that strategies employed by the STEM EC have built capacity of the community to conduct methodologically appropriate and conceptually sound evaluations?

[1] Additional activities included: presenting at AEA (think tank, November, 2018), securing funding to host a third convening (September, 2018), participating in the AEA 365 blog series during STEM TIG week (February, 2019), receiving funding for AGEP proposal that involves another group of evaluators and conferencing around capacity building (February, 2019), and leveraging personal and professional relationships to extend STEM evaluation capacity building efforts (ongoing).

METHODS AND ANALYSIS

Methods used to produce this report include participant observation at the three convenings (December, 2017; July, 2018; July, 2019); document review of meeting materials and studies produced; and participant written feedback on reflection questions following the second two convenings. Fieldnotes were taken real-time on a laptop with a focus on capturing content. Fieldnotes were supplemented by notes taken by the EDC team during the meeting, meeting agendas, and slide decks.

Data (fieldnotes and reflections) were coded in Excel v.16.10 for analysis. Codes included constructs important to the project (i.e., community building, evaluation quality, and capacity building), communities of practice and networks value frameworks from Wenger et al (2011), emergent themes (e.g., evaluation use, context of practice), and change targets or where capacity building efforts should focus (i.e., evaluation practice, clients, funders, resources).

FINDINGS

I begin with the first convening because origin stories reveal a lot about how groups evolve, and the convening in December 2017 was the first time that members of the STEM EC were brought together face-to-face. At this first meeting key questions about target audience and the nature of the problem were raised that sent the STEM EC down a path that hadn't been planned for and served as an illustrative example of how activities organized around "people as resources" creates value.

The First Convening

The STEM EC first came together in December, 2017. The meeting was made possible through a grant the EDC received from NSF. Although the convening was the first time the group came together, the ground that the group would cover was initially defined in the EDC team's proposal. The proposal framed the problem, identified key stakeholder groups (at least at the outset), and laid out a plan for the work. The presenting need was framed as one of evaluator capacity building, specifically to conduct methodologically rigorous and conceptually sound evaluations of NSF STEM outreach and education programs and projects (Proposal narrative, 2017). Key stakeholder groups included NSF PRCs (PIs, staff and evaluators), evaluation advisors (individuals working in the field of evaluation in a variety of roles and with a range of perspectives), evaluation representatives (individuals who had conducted evaluations for multiple NSF programs), and the funder, NSF. The plan was organized around a 4-phase theory of change: (1) identify resources, needs, and opportunities; (2) convene representatives from the key stakeholder groups to discuss, reflect, and plan; (3) share the work with a wider audience; and (4) plan for sustainability. The EDC team proposed using communities of practice and responsive evaluation to support the process.

Twenty-five people attended the 2-day Kick-off convening. Attendees included nine people from PRCs (PIs, staff, and evaluators), six evaluation advisors, four evaluation representatives, the four EDC team members, the NSF Program Officer, and me, the external evaluator. Based on the bios that individuals had written, the majority of STEM EC members were mid-to-late career professionals with substantive experience in evaluation. A handful of members were

retired. Many had worked with NSF in roles ranging from PRC PIs to program officers to evaluators of NSF STEM projects and programs. STEM EC members worked in academia, research/evaluation organizations (from large corporations to small evaluation companies), industry, and government.

The goals of that first convening were to:

- Share vision for project and community of practice
- Get to know each other and begin to develop community
- Share learning from the Landscape Study
- Solicit ideas and plans for developing community infrastructure
- Begin process of collaboration and planning for community convening
- Enjoy being together as people who are committed to evaluation capacity building! (PPT slides STEM EC Kick off meeting, Dec 7-8, 2017)

STEM EC attendees made it clear how much they looked forward to the meeting and the opportunity to come together at the get-go as each person responded to two ice-breaker questions: "What are you most interested in about this meeting?" and "What is one exciting thing about the work you do (related or not related to what brings you here)?" Among the things that people stated that they were interested in getting from the meeting were "learning others' perspectives," "meeting new people and hearing about new initiatives," "hav[ing] the energy and passion to participate, to engage, to build our capacity," "Sharing expertise, generative conversations," and "Collaborat[ing] around work, our work" (field notes).[2] A few remarked how valuable other Communities of Practice experiences had been for them. One person noted an interest in learning about orientations to evaluation held by people in the group. Another attendee hoped that people would speak their minds and be open.

Participants' commitment to bettering evaluation was also evident in responses to the ice breaker questions. About a third of the attendees indicated that they were most interested in learning more about evaluation at NSF and getting a better understanding of the landscape of STEM Resources. One person stated they were most interested in "Leveraging whatever we can to build capacity—tools, resources, practical knowledge." Another person expressed interest in "shared measures" while a third wondered if "successful collaboration [would] yield a repository of useful and used materials."

After the getting-to-know-you part of the morning, the EDC team presented an overview of the project and planned work together, and I shared some information about Wenger et al's definition of Communities of Practice, types of value created by CoPs and networks, and stages

[2] Unless otherwise indicated, quotations denote material taken directly from fieldnotes, not necessarily direct quotes from individuals. As noted in the methods, the fieldnotes taken focused on communicated content, and I tried to stay true to the language used by the STEM EC members. Often my fingers were not fast enough to capture everything but based on comparing my fieldnotes with those taken by the EDC group, they were reasonably complete and true to what was said.

that CoPs move through.[3] These activities were meant to get members on the same page and to develop some shared language and common understandings. For instance, I suggested that Wenger et al's definition of communities of practice could be a helpful way to think about a STEM EC CoP. As they define it, CoPs are made up of a *domain* or shared identity/area to come together around, a c*ommunity* or members and connections between them, and a *Practice* or set of tools, concepts, language, frameworks, and stories (Wenger et al, 2011).

On the heels of these introductory remarks snapped questions such as "Capacity building for whom?" and "What's the need we are trying to address?" The following snippet from my fieldnotes[4] captures some of the concerns:

- What is the need? We haven't talked about that.
- What is valuable for people? Where's the sustainability plan? What's the value to the people here? And what's the value to the larger community?
- Are we already a community?
- What's the problem?
- What's the lifecycle?
- Do we practice evaluation? Who are we?
- How do people in our world/profession define our practice? Are we similar or different from lawyers? Doctors? Other professionals?
- Is CoP just a phrase with a short life span?
- Value of your time. Who can support this initiative post-funding, without some sort of institutional support?
- Rapidly changing landscape poses a real need
- Need in general / hypothetical – or informed by data? Which way to go now?
- Tension between content knowledge and evaluation knowledge
- Performance need v instrumental need
- CoP as one of many ways to build evaluation capacity
- Can we think about need in relation to the market (in this case, NSF is the market)?
- Building capacity to what end and for what purpose?
- Difference between felt need and expressed need.
- For whom?
- Is this a solution in search of a problem?

[3] See Appendix A for a figure of Stages of Communities of Practice and a list of the value CoPs create.
[4] Each bullet represents one person's comment.

- This dialog is a means to finding solutions to whatever problem defined

The discussion above marks to me an important interrogation of assumptions embedded in the approach and solution proposed by the EDC team. Who was this work for? Who were we as a group? Why a Community of Practice approach? Building capacity for whom? Is this a solution in search of a problem?

We dived into the Landscape Study next. It served multiple useful functions directly and indirectly through the conversation it engendered. First, the study itself provided a baseline of NSF's STEM evaluation resources, particularly those that existed across EHR. It showed that there were ample resources for the early stages of evaluation (i.e., evaluation planning, proposal writing, instrument repositories) but few on the "messy middle," a term coined by one of the EDC team members that was picked up by the rest of the STEM EC and referred to throughout subsequent convenings.

The study found that NSF STEM evaluation resources were housed in different places and for different purposes, raising questions about duplication and specialization. Insights from STEM EC members, particularly evaluators who had used the resources and PRC PIs and staff were invaluable, creating more awareness about the audiences PRCs serve and how the PRCs might better meet the needs of project evaluators under the programs they support. For instance, in response to the reflection questions following the third convening, one STEM EC member wrote that the conversation about the Landscape Study "has made me more sensitive to the wide variety of resources that may be lacking in the way NSF support centers think about developing and disseminating evaluation resources." Another STEM EC member wrote, "Hearing the continuity of messages such as the 'messy middle' being the area that evaluators and consumers of same most struggle with, has helped me think about what my own project/resource center might do to build capacity and understanding in the informal sector" (Participant Reflection, Convening 3).

Second, the Landscape Study showed the variability of evaluation expectations and requirements across NSF programs. For instance, only 23% of programs require an external evaluator while others are shifting from evaluations to external advisory boards (Landscape Study, February, 2019). These findings resonated with the experiences of evaluation representatives who had conducted evaluations at the program and project levels for different NSF programs under EHR. Sources of this variation became clearer over the three convenings as STEM EC members learned more about NSF from those who had worked with the Foundation as program officers and PIs of PRCs, from evaluators of NSF programs and projects, and from the vantage point of other initiatives and agencies at the federal level.[5]

[5] EDC team incorporated learnings from discussion and conducted some additional analysis that were presented at the third convening and in the final Landscape Study report (February, 2019).

Third, the Landscape Study highlighted several challenges to how STEM evaluation resources had accumulated at NSF. For instance, although a wealth of resources had been accumulated, the resources had not necessarily been well-managed. One STEM EC member likened this to the early days of the internet, noting that we had a "landscape of links but not generative. The resources from the landscape study feels like that." The conversation that ensued raised more concerns about the tendency for programs to create repositories and for projects to build more websites. As one STEM EC member observed at the second convening, repositories had been unhelpful at best. Worse, they became "the dumping ground of instruments… [and this] bred laziness and bad practice."

Lastly, the discussion of the Landscape study once again intimated that orientations about evaluation practice within this group were many. Assumptions about practice derive from an individual's training and experience, and many of us had witnessed significant changes in the field over the last decade. For instance, there's been a decline in the number of academic evaluators and a corresponding rise of practitioners, many who lacked formal evaluation training. Indeed, some of the diversity amongst the STEM EC membership reflected these changes as individuals worked in a wide variety of settings, mostly outside academia.

The first day concluded with a presentation by Cynthia Philips, the project's program officer. She offered STEM EC members valuable context about evaluation at NSF. Philips spoke in particular about the Office of Integrated Activities' Evaluation and Assessment Capabilities which was charged with building evaluation capacity, changing the Foundation's learning culture, and making better use of data. She brought up the *NSF 17-111 Discovery! Dear Colleagues* solicitation that targeted STEM researchers, inviting them to propose education and outreach activities as part of their own research agendas, as well as HB 4147 which would eventually be passed into federal legislation as the *Foundations for Evidence-Based Policymaking Act of 2018* that required all federal agencies to develop a plan to facilitate data use for policy making.

In response to all that had been surfaced the first day, the EDC team scrapped the 4-phase plan and used much of the second day to identify next steps to take together. As one of the team members said, we decided we should "open up rather than drive" to the broader convening. When asked what the STEM EC was funded to do, the response was "The grant has funded a process. A community of practice might be what emerges. Might be something else." The next morning, the EDC team presented a new agenda for the day that included reflecting on the previous day's work, talking about what problem(s) the group wanted to address, what needs they saw, and what the STEM EC might do about it.

The opening up started with reflecting on the previous day's work. STEM EC attendees raised a slew of relevant questions about the emerging community's boundaries. One person noted, "STEM community isn't solely relegated to NSF. NIH isn't open to that kind of communication. NASA too…Needs and problems haven't been sufficiently explored. Look at solutions as well." Another asked, "What are we really trying to affect? STEM education. Put NSF aside. We know a lot about good STEM education. The world is changing fast. What does that mean for STEM

education? Knowledge accumulation v knowledge production. Evaluators should be part of that."

Other comments pertained to evaluation at NSF. One person observed, that "NSF [was a] churn of personnel and politics of experience. Really impacts evaluation use and how it's conceptualized. Where's the problem? Haven't designed the programs in ways that are amenable to the kinds of evaluation and impacts that we want. NSF blames evaluation but it is [also] the culture of the Foundation. Not all of the Foundation is ready for that. Some are." Another person asked about "the role that NSF plays in communicating expectations to projects and evaluators. Evaluation often considered synonymous with compliance. Other roles? Can NSF educate? NSF needs education about evaluation. Very different perspectives depending on who you ask." Some of the structure and policies at NSF made it difficult to influence evaluative thinking at the Foundation. For example, "Evaluators don't get direct access to program officers, making it difficult to directly influence evaluative thinking at the foundation." Finally, one person acknowledged the reality that "NSF [is] messy and that includes understandings/perspectives on science, evaluation, etc. Not consistent for good or for bad."

Focusing back in on evaluation itself, some STEM EC members returned to the observation that evaluation practice means a lot of different things to people within the STEM EC. People practice evaluation differently. Some evaluators are very engaged with stakeholders as part of doing evaluation. They provide technical assistance, offer informal feedback on a regular basis, and work closely with project staff to make sense of collected data and how it can be used for program improvement. Others keep themselves at a distance and are hands off when it comes to direct program/project improvement. Evaluation reports are generated and given to PIs and funders with the assumption that they will make use of the findings. One STEM EC member observed that "Evaluation use is more than instrumental use. We engage stakeholders throughout the process. Educative function that happens through conversation, on the fly. Often this type of use doesn't rise to the level of being documented. Are we looking at use too narrowly?" Musing on the changing context of evaluation exemplified in policies such as the Evidence Act, changes in evaluation requirements evident in solicitations examined as part of the Landscape study, and the concerns about quality and value that gave rise to this project in the first place, one STEM EC member stated, "We [evaluators] don't want to go the way of the dinosaur."

After much conversation and some corralling, STEM EC members agreed on four problem/need areas that affect evaluation quality to explore during the "unconferencing" portion of the convening: (1) evaluator pathways and training, (2) evaluators' social networks, (3) a STEM EC theory of change (ToC), and (4) evaluation use. STEM EC members sorted themselves into small problem/need groups to unpack related issues and strategies that might be employed to address them.

In my fieldnotes, I write, "In groups, conversations all touch on evaluation quality. It's a problem. Lots of ways to tackle it, depending on how one frames it."

- The *evaluator pathways group* connected the problem of quality with how people become evaluators. One person in that group reported out that we could "try to meet people earlier on their pathways, build awareness of evaluation as a career so [they] have clearer understanding of how to get there." Capacity building efforts would help "individuals develop the competencies that they need depending on where they come from."
- The *social networking group* posited that evaluation quality could be improved through better use of social and knowledge capital distributed across the network of STEM evaluators. To this end, there could be value in mapping out relationships to identify points of access to human resources that could be leveraged for capacity building.
- The *theory of change group* acknowledged that numerous strategies could be employed to improve the quality of STEM evaluation depending on where one focused in the model. For instance, the group emphasized that quality evaluations should account for the dynamic contexts in which evaluation is practiced. A quality evaluation would build evaluative thinking capacity across stakeholders and contribute to a long-term vision of making the world a more just place.[6]
- The *evaluation use group* connected evaluation quality with usability. They talked about expectations for evaluation that funders hold and communicate to PIs (often being evaluation as compliance), the role that funders play in promoting use, challenges associated with sharing critical findings, curbing misuse, and the need for "safe spaces" for learning.

The first convening concluded with a round-robin where each person stated one thing that they'd like to work on. Pathways, social networks, evaluative thinking, use, and resource development for the messy middle each were listed by 2-5 people. Other areas individuals wanted to work on were sustainable communities, bridging the left and right sides of the ToC that was created, helping PIs see evaluation as "cutting edge," and finding innovative ways to provide training.

The STEM EC's first convening concluded with a box lunch and coordinating rides to the airport and elsewhere.

…

The STEM EC was brought together two more times, once six months after the first convening and again one year after that. Participants at the first two convenings were mostly the same. Fewer people from the first two convenings attended the third one. Several new individuals also attended.[7] The structure of these convenings was the same as the first one. The group met

[6] The theory of change model created at the first convening and revisited at the third one can be viewed in Appendix B.

[7] There are several possible explanations for why the third convening's attendance differed from the two previous ones. A third convening was not in the original plan and was a result of securing supplemental funds. Every one of the STEM EC members had many commitments so it may simply have been an issue of time and timing. Staffing changes and internships brought a few new people to the STEM EC and took others away. Lastly, it's also possible

in the afternoon the first day, had dinner together that evening, and then met again the next morning. Other than that, there was very little coordinated activity of the STEM EC as a whole outside of the convenings.

The STEM EC and Value Creation

Three sets of strategies affected the value created by the STEM EC: (1) selecting members and subsequent members; (2) building community; and (3) supporting learning. Neither independent nor mutually exclusive, these strategies interacted and overlapped as much as they served specific functions. For instance, selecting members was done based on what individuals could contribute to the community be it expertise, perspectives on practice, or role at NSF. Community building required attending to what members brought to the community, and as we saw in the first convening described above, being responsive to them. Similarly, specific activities such as producing data-based resources in the form of the Landscape Study and the Evaluator Survey contributed to the knowledge base as well as established shared reference points about STEM evaluation resources at NSF. Below I unpack each major strategy, linking activities and decisions to the value created for individuals and beyond as evidenced in fieldnotes, member reflections, and products generated.

Membership

Wenger et al (2011) identify *community* or *membership* as one of three key elements in communities of practice.[8] The value created in any CoP depends on the composition of the membership and its fit to the needs of the task(s) the CoP has been assembled to address. Stakeholder groups represented in the STEM EC included: evaluation representatives; evaluation advisors; PRC PIs and staff; and NSF. The evaluation representatives brought expertise and experience conducting STEM evaluations in multiple NSF EHR programs. PRC PIs and NSF-affiliated persons brought insight and understanding of the NSF context to the community as well as possible places to implement capacity building efforts. Evaluation advisors rounded out the community, bringing experiences and perspectives from the wider evaluation community. Members of the EDC team also had significant experience doing evaluation of STEM outreach and education initiatives at NSF as well as a PI strategically situated as American Evaluation Association president and past NSF rotating program officer. Given the original task of the STEM EC to build capacity of STEM evaluators, the composition of the membership made sense.

Coming together for the convenings created immediate value for the people who attended. After the third convening, one person wrote, “I loved the opportunity to have this conversation with such an amazing group, we had some really generative conversations” (Participant Reflection, Convening 3). Another wrote, “I was equally impressed and amazed at the collective wisdom that resides within the community members” (Participant Reflection, Convening 3).

that some of that early energy evidenced in the first two meetings had dissipated over the intervening year between the second and third convening, and people found themselves focusing on other things.

[8] The other two elements are domain and practice.

The STEM EC membership also yielded potential value for participants. Potential value in the form of knowledge capital accrued through dialogue and reflection and the unstructured time together afforded by the face-to-face gathering. This was evident at all of the convenings in individuals' responses to the ice breaker and check in questions that opened up each meeting, in the second morning's ah-ha's, and in participant reflections following the second and third convenings. For instance, one person reflected, "The opportunity to meet face-to-face with STEM Evaluation Community members to discuss issues related to STEM evaluation is critical as a mechanism for reflecting, collaboratively engaging in discussions, and sharing experiences and expertise with other evaluators" (Participant reflection, Convening 3). Another wrote, "A lot of what I am learning/discovering through this project will be folded into my own work in terms of building networks of colleagues" (Participant reflection, Convening 2).

Finally, the STEM EC convenings created applied value for individual members. Applied value was evidenced in changes in practice and new collaborations. STEM EC members shared these experiences in response to the question, "Did anything happen that was an inspiration or connection as a result of the last meeting?" that opened the second and third convenings. The following are examples of applied value created:

- ATE did a webinar with EvalFest
- AISL working with evaluation advisor to provide culturally responsive evaluation training;
- An evaluation representative recruited several readers for a book she was writing with colleagues;
- Several people made note that they have put renewed attention to collaboration and building communities of practice within their own project evaluations.

Although individuals experienced personal and professional benefit from participating in the STEM EC, with a few exceptions as illustrated under applied value in the table above, value created beyond the group was unclear, at least as far as membership went. This frustration was expressed by a number of STEM EC members who shared a commitment to better evaluation. For instance, one person wrote, "Face-to-face opportunities are expensive, and our group has been extremely fortunate to have the privilege to participate. However, technology may provide other, more cost effective, means of providing a similar opportunity to a larger audience" (Participant reflection, Convening 2). Along similar lines another member wrote, "[Although] it is extremely enjoyable, and professionally rewarding, [my concern is that] it doesn't lend itself to any sustainable or impactful outcome that will influence or modify future direction" (Participant reflection, Convening 2). After the third convening another person lauded the personal and professional benefit they'd experienced, continuing, "I think the process is interesting and it's valuable to me, personally as an evaluator. It's invigorating and the relationships are professional [sic] valuable. I do wish that we could latch onto some concrete output of the work" (Participant reflection, Convening 3).

One challenge to value creation beyond STEM EC members themselves emerged when the original 4-phase plan was jettisoned without something to replace it. Rather than having defined task-based groups working toward a convening with a broader set of stakeholders and organizing those groups around individuals' interests and skill sets as would likely have been the case, STEM EC membership roles ended up being largely undifferentiated. With the exception of the EDC team and myself, everyone was considered a valued member of the STEM EC and participated the same way as everyone else did. What the STEM EC members ended up participated in was an extended interrogation of problems of (NSF) (STEM) evaluation practice, some of which could be addressed with the people in the room, others not so much (particularly when it came to evaluation use and impacting NSF evaluation requirements).

A second challenge was the diversity within the group itself. Three types of diversity appeared to be particularly relevant: professional working context; perspectives on evaluation practice; and cognitive diversity. As noted earlier, STEM EC members differed in terms of where they worked (academia, large research/evaluation organizations, small evaluation companies, industry, and government). They also differed in where they were in their professional careers (mid-late career, retired). These differences impacted what individuals needed from participating. For example, several people involved in PRCs noted that armed with a better understanding of key issues facing STEM evaluators, they could respond to those needs more effectively. For others, stimulating conversation was all they needed from participating in the STEM EC. Yet others, particularly those in research firms and small consulting companies, needed something tangible from their participation, such as access to new streams of funding be they sought collaboratively or competitively.[9]

The diversity of perspectives on evaluation practice had positive and not so positive consequences for the work of the STEM EC. On the positive side, it was very good for unpacking evaluation problems of practice from multiple perspectives. On the less positive side, it made it difficult to come to concerted action. Reviewing the fieldnotes from the three convenings along with participant reflections, I was able to identify at least three frameworks on evaluation brought to the STEM EC: instrumental, educative, and emancipatory.[10] The frameworks we carry with us to practice go on to define what we think evaluator roles should be, our relationships with stakeholders (or clients or evaluand or "evaluatives"[11] or participants), and the ways that evaluation can be used (see Table 1).

[9] Competition for funds emerged as a confounding factor in community building. We might want to work together but we have to compete with each other for access to funds, curtailing incentive to share knowledge and collaborate. Members of the STEM EC referred to this as the "collapetitive" evaluation environment.

[10] I used a similar framework in my master's thesis to organize a literature review of evaluation practice. See Kaminsky, A. (1993). *Participatory evaluation in hierarchies: Practical dilemmas of implementation (A case study)*. (Unpublished master's thesis). Cornell University, Ithaca, New York.

[11] B. Parson and colleagues in their book, Visionary Evaluation (2019, Information Age Publishing), use the term "evaluatives" to disrupt power differentials implied in the language of evaluand and evaluator and to highlight shared commitment to making the world a better place.

Table 1: Perspectives on Evaluation Practice

Perspective	*Role*	*Relationships*	*Use*
Instrumental	Technician	Formal, distance	Compliance. Did the program meet its goals? Was our money well spent?
Educative	Facilitator	Engaged	Learning/program improvement
Emancipatory	Advocate	Engaged	Disruption/breaking down systems of oppression

These different perspectives on evaluation practice, unsurprisingly, lead to divergent conceptualizations of problems and how they might be addressed. A question that kept coming up in convenings without being directly addressed was what to do with the different beliefs about evaluation practice. Did STEM EC members need to arrive at agreement about what evaluation was? What if the community embraced the multiplicity? How about people outside of the community, like evaluation consumers? How did their expectations about evaluation affect demand for any particular approach to evaluation? Consensus about practice would make it easier to come to coordinated action but risked ostracizing people who held different beliefs. That said, accepting that evaluation is a pluralistic field would make it more likely that a wide variety of evaluators would find a place in the community.

Finally, cognitive diversity affected STEM EC members' comfort level with the open-ended process. One STEM EC member brought attention to this type of diversity during the second convening when individuals were tasked with coming up with a vision for the STEM EC and three tangible steps to get there in small groups. The member observed that the room had people who were strong *reactors* (critical thinkers, tendency to want to excavate assumptions) and others who were better *creators* (visionary, creative). Both types of people made valuable contributions to the STEM EC but thrived under different conditions with tasks suited to their ways of thinking.

STEM EC members acknowledged positive and negative consequences of the level and types of diversity within the group and how they were addressed (or not) as illustrated in the following passages excerpted from reflections:

- "It seems we could have gotten a lot accomplished rather than just sort of brainstorming. I guess it is a good way to handle such a large group of diverse opinions. Everyone gets to say whatever they want and everything is accepted" (Participant Reflection, Convening 3).
- "It's a very mixed bag in my view. It seems like there is a lot of wisdom in the group that is simply not getting surfaced" (Participant Reflection, Convening 3).

- "It may be useful to do some high-level affinity and consensus building with the group to prioritize going forward" (Participant Reflection, Convening 2).
- "It makes me increasingly attentive to how important it is to find that right balance of knowing what those in a STEM CoP (as well as the designers) want to accomplish and what the range of ideas and resources they bring to the table for engagement within the CoP" (Participant Reflection, Convening 3).

Community building strategies

The EDC team demonstrated a commitment to building the community aspect of the STEM EC in multiple ways over the course of the project. First, there were the "getting to know you" activities such as ice breakers, check ins, sharing biographies, meeting face to face, and sharing meals. Face-to-face meetings, in particular, afforded people opportunities to have unplanned conversations and make connections due to the combination of structured and unstructured time together. The EDC team organized convenings to take advantage of the time together to include work and play (in this case, eating). The informal time opened up space for individuals to explore shared interests and potential collaborations. One person wrote, "What this group should be focusing on [is] bringing together evaluators, whether they be STEM, Women, or general. Form those groups, and get the conversation flowing which leads to networking, and stronger evaluations" (Participant Reflection, Convening 2). Indeed, a number of STEM EC members drew on the new connections they made, turning potential valuing from the STEM EC convenings into applied value beyond the group itself as noted under membership.

A second way that the EDC team demonstrated a commitment to attending to the people making up the STEM EC was by being responsive to where people were and being willing to let go of the original 4-phase plan in light of concerns raised at that first convening. As stated that first afternoon, "We decided to open it up rather than drive to a larger convening." The rest of the first convening was spent identifying and agreeing upon four areas for further investigation: evaluator pathways, social networks, theory of change for the work, and evaluation use. The Evaluator Survey was one mechanism implemented to investigate pathways, networks, and issues around use. After the second convening, one person wrote, "Starting with the mindset of, 'we don't know how or if this is going to work, but let's figure it out together' has been motivating and inspiring and I think this approach has the potential to transcend all levels of evaluation."

A third set of strategies to build community focused on fostering shared language and reference points amongst community members. Definitions for concepts such as communities of practice were shared as part of PPT presentations to the group. Reports and articles such as the Landscape Study and Schwandt's article, "Evaluative Thinking as a Collaborative Social Practice: The Case of Boundary Judgment Making" (2018) were distributed to STEM EC members before convenings and provided common reference points for group members. Invited presentations during the convenings served a similar function.

Stories shared at the convenings also supported the development of a shared language and reference points. One could see which ones were particularly valuable to people's thinking

when they were referenced later in the day or at subsequent meetings. For example, one story shared with the group at the second convening was about General Electric creating the transistor radio. The story had been shared as part of characterizing innovative thinking and creating demand before people even knew that they wanted something. The analogy was later appropriated in small group work during the second convening. In response to the task to come up with an innovative vision for the STEM EC and three tangible steps to get there, one group suggested creating "an evaluator app" that "puts the evaluation community in your pocket." Similarly, "messy middle," a phrase coined by one of the EDC team members was returned to again and again over the course of the three convenings. Thinking about problems with evaluation from a "supply and demand" lens was another perspective that found purchase for many STEM EC members.

As the foregoing paragraphs intimate, shared language and common reference points were built through dialogue and reflection on the information brought to the group and through the conversation process itself. In short, it happened through learning.

Learning focus

One core function of any community of practice is learning around a domain of knowledge. This learning, and the potential value it creates, can be applied to alter practice, improve performance, and in some cases, reframe how we think about and approach our practice. Indeed, an essential premise of the STEM EC was that learning is a social endeavor and that we can learn a lot when a lot of smart, experienced, knowledgeable people come together to share what they know. This orientation created potential value for STEM EC members, applied value as learnings were shared with groups beyond the STEM EC, and, depending on one's beliefs about evaluation practice, some reframed value.

The data-based resources developed by the EDC team stimulated rich discussion within the STEM EC. Recall, the Landscape Study conducted by the EDC team provided a baseline of NSF's STEM Evaluation resources, particularly those that existed across EHR, and showed gaps and variability in how evaluation has been cast by NSF over time. Key learnings from the study were also shared at a think tank session at the 2018 American Evaluation Association (AEA) annual meeting and through AEA's 365 Blog posts during STEM week.

The Evaluator Survey also spurred conversation both in terms of what was found and the difficulties fielding the survey in the first place. A major impediment to administering the Evaluator Survey was the fact that comprehensive lists of people who had evaluated STEM projects and programs were not maintained at NSF. Some of the PRCs had evaluator lists but these were limited to the projects under the programs the PRCs were set up to support. This challenge resulted in a response set skewed towards mid-late career evaluators. With those caveats in mind, the Evaluator Survey did afford the group some insight into the pathways that people travel to get to STEM evaluation and led to conversation about what unique skills and knowledge that evaluators bring to any particular study. It confirmed that a good proportion of people were not formally trained in evaluation (i.e., extended courses of study vs just in time trainings). It was also informative insofar as it showed that evaluators go to their professional

networks to navigate messy-middle type problems of scope creep, reporting negative findings, and ethical dilemmas.

In the responses to the questions sent out after the third convening, most participants confirmed that the data-based products produced and discussions about them at convenings yielded better understanding of the STEM evaluation context, particularly at NSF. One person wrote, "I have become more clear about some of the ways that [our PRC] can support [program] projects and evaluation. I also have a better sense of key issues facing STEM evaluators" (Participant Reflection, Convening 3). Another wrote, "Hearing the continuity of messages such as the 'messy middle' being the area that evaluators and consumers of same most struggle with, has helped me think about what my own project/resource center might do to build capacity and understanding in the informal sector" (Participant Reflection, Convening 3). A third person reflected, "It has made me more sensitive to the wide variety of resources that may be lacking in the way NSF support centers think about developing and disseminating evaluation resources, although it also confirmed my suspicion … that each center is sufficiently unique in its products/services that there is not much to be gained by trying to synthesize these" (Participant Reflection, Convening 3). As these quotes illustrate, the documents produced and discussions that ensued created quite a bit of potential value for STEM EC members that could be applied to individual evaluation practice as well as used by PRCs to better support evaluators of their programs and projects. That said, "How this effects our individual work/projects vs. the overall project is a bit unclear" (Participant Reflection, Convening 3).

The conversations about the products developed also helped identify unintended consequences of previous efforts to build evaluation capacity, particularly in the realm of measurement and instrumentation. For instance, the proliferation of instrument databases has resulted in uncritical, unthoughtful adoption of instruments (i.e., using surveys for populations that they were not validated for or to measure constructs not relevant to the project being evaluated). To address these types of concerns, STEM EC members started talking about how to create better resource bases. When the group went back to what could be done in response to the problems identified through the Landscape Study, one person recommended creating living resources that were "easily accessible through digital form, with opportunities to modify or input how it worked and under what conditions…for new and seasoned evaluators…I see these resources being generated as simple, two-page strategy descriptions and messages, or something like that." Another person suggested "develop[ing] products to guide practice [such as] straightforward guidance for all stakeholders about how to pick from the available instruments - or equally, why they might not. We could influence the big-picture conversations among evaluators, funders, and grantees. That would do us all some good." This person emphasized that the products "wouldn't be like the 'how-to' stuff that's linked in the NSF solicitations and elsewhere" and was "*not* …another repository of instruments." (Participant Reflection, Convening 3).

In another example, one STEM EC member mused that they may have unintentionally contributed to a perception that doing good evaluation is a matter of taking a few evaluation

workshops. At the second convening, one member reflected on the evaluation materials put out by their PRC, saying, "[There] may have had unintended consequences of making people think that they are evaluators since they've taken a couple of webinars. We were never set up that way. We made the trainings accessible and practical but maybe that's done us a disservice." Learnings generated through conversation contributed to the expanded conceptualization of capacity building where attention shifted away from the more technical aspects of evaluation practice to more intentional consideration about how relationships in evaluation affect quality and the importance of developing evaluation habits of mind or evaluative thinking.

The invited presentations at the convenings also furthered learning within the group. Five presentations were shared with the STEM EC over the three convenings: one on building communities of practice in the Computer Science Impact Network (CSIN); three related to the wider context of STEM evaluation at the US federal government level; and one on convening cross-stakeholder meetings based on the work of the Evaluation Roundtable for Philanthropy.

The communities of practice presentation and cross-stakeholder presentations provided STEM EC members with analogous situations against which to compare and contrast their own efforts in the STEM EC. For example, STEM EC members Jason Ravitch and Tom McKlin shared their experiences building the CSIN at the second convening. The presentation hit on a number of factors that, in retrospect, could account for some of the differences in how the STEM EC has evolved compared to the CSIN. I compare the CSIN and the STEM EC in the table below.

Table 2: CSIN and STEM EC Compared.

	CSIN	*STEM EC*
Stage of development[a]	Stewardship: In development for 8 years	Potential: In development for 2 years
Domain	Computer Science Education	Evaluation of STEM outreach and education programs (or evaluation writ large or STEM writ large or NSF)
Community	Mostly CS evaluators	Mix of NSF STEM evaluators, evaluation advisors from different areas, and NSF people (program officer, PRCs)
Practice	Identified need based on early needs assessment. Found that repository of instruments focused on classroom implementation and effectiveness was needed	Multiple approaches to evaluation practice; STEM evaluation has long history, combines many disparate areas under one acronym; different beliefs about evaluation influence what one considers a problem and how it should be addressed.
Process	Structured facilitation using Fetterman's Empowerment process for "deep collaboration around visioning"	Open ended, iterative, problem investigation, and learning focused.

Roles	Differentiated; some funding for continuity in leadership and facilitation; task-focused subcommittees	Undifferentiated with the exception of the EDC team and external evaluator
Communication	Regular meetings, tinyURL, live site that's always updated	Sporadic, face to face during meetings; email otherwise
Incentive to participate	Learning, solving problems, opportunities to collaborate /seek funding; dissemination of ideas	Learning, solving problems, opportunities to collaborate /seek funding; dissemination of ideas
Sponsorship	Seeking institutional champion	NSF (maybe?)

a. Wenger et al (2002) posit that Communities of Practice go through 5 stages of development: potential, coalescing, maturing, stewardship and transformation (see Appendix A for figure).

As the foregoing table illustrates, the CSIN and the STEM EC differ in many ways. The domains within which each operate differ greatly. CS education is a relatively new field, compared to STEM, an amorphous conglomerate of science, engineering, mathematics, and technical disciplines. In terms of community, CSIN members are interested in CS for the most part, whereas a number of STEM EC members described their relationship with STEM in terms of broadening participation, not in terms of STEM substance. Practice-wise, members of the CSIN seem to have more in common than those in the STEM EC. CSIN is made up mostly of evaluators who, through a facilitated process, collectively recognized a need for better evaluation of classroom implementation and their effectiveness. As discussed under *diversity*, members of the STEM EC held different (at times divergent) beliefs about evaluation and evaluation practice that made it difficult to come to agreement about problems of practice and where capacity building could have the biggest bang for its buck. Lastly, the CSIN and STEM EC were in different places with regard to where they were in their life cycles. CSIN, now in its 8th year is in the stewardship stage where the community is focused on ownership and openness and can account for the focus on seeking an organizational champion and finding ways to sustain the community; the STEM EC, on the other hand, is still in the Potential stage, feeling out what it is and what it could do in the world.

Presentations about federal initiatives related to evaluation and STEM also contributed to learning. These presentations afforded STEM EC members opportunities to reflect on how actions at the federal level might inform the actions taken by the STEM EC. For example, at the first convening, Cynthia Philips provided the STEM EC with an overview of her office of Integrated Activities Evaluation and Assessment Capabilities at NSF. This presentation highlighted some of the actions NSF was taking around evaluation as well as current STEM-related initiatives, including ways that the Foundation's efforts to stimulate STEM researcher involvement in education and outreach activities. Nick Hart's presentation on evaluation initiatives at the federal level at the second convening, like Philips' at the previous convening, helped to clarify the larger federal context of evaluation. The just-in-time presentation related to STEM evaluation in other federal agencies and the Evidence Act given by two STEM EC members at the last convening further contributed to learning how the work of the STEM EC fit within a larger framework. As one person wrote in their reflections following the third convening, "The final meeting took an even larger perspective by considering the

social/organizational context in which evaluators operate and how the community might move the field forward by looking at outside forces rather than focusing on internal needs such as capacity building" (Participant Reflection, Convening 3).

The last talk presented to the STEM EC was the work done as part of the Evaluation Roundtable for Philanthropy by Katrina Bledsoe. The presentation was particularly timely since the STEM EC members were considering where they thought the group could have the most influence given three possibilities put forth by the EDC team (NSF, the field of evaluation, or a cross-stakeholder convening). The presentation and the discussion that ensued highlighted the things that make cross-stakeholder learning opportunities possible. Numerous people observed that these conversations need to happen in low-stakes settings where people feel safe in that there will be no retribution for honest efforts to engage with hard topics, to acknowledge mistakes, and to offer critical feedback. Indeed, this is one significant challenge to evaluation practice. A compliance mentality is anathema to being open and honest about mistakes made and learning from them. Beldsoe shared a couple of comments that had been made at the Roundtable where she thought, "Do you really want to go there? Do you really want to have that job tomorrow?" In response, one STEM EC member stated, "I keep hearing it Isn't safe to be wrong. NSF is the poster child for this. I cannot understate the importance of creating an environment where it's safe to learn."

DISCUSSION

So then, to return to the questions that guided this evaluation, what have we learned? Specifically, what have we learned about:

1. What types of value creation do the STEM EC activities and strategies yield? and
2. What evidence do we have that strategies employed by the STEM EC have built capacity of the community to conduct methodologically appropriate and conceptually sound evaluations?

First, the question of value creation. For STEM EC members, participation in the community yielded immediate from participation and potential value in the form of knowledge production, and in some cases, applied value defined as changes in practice and reuse of community-generated materials with new groups or in different ways. Immediate value as evident from the enthusiasm members brought to the convenings and the reflective comments they made at convenings and in reflections. Potential value in the form of learning and knowledge production was created through the products generated by the EDC team, the invited presentations and other resources shared with the STEM EC, and the dialogue and reflection on them.

The EDC team sought to foster a responsive and inclusive environment, demonstrated most clearly in the willingness to let go of the original plan in light of the questions, comments, and concerns raised by participants at the first convening about the purpose of the group and the need that they were tasked with addressing. The willingness to change course came with a cost because there wasn't something to replace the original plan with; as a result, much of the STEM EC convening time was consumed with open-ended processes where STEM EC members tried

to identify what problem(s) contributed to inconsistent evaluation quality and where capacity building should/could happen.

Coming up with a new plan collaboratively was challenged by the diversity that existed in the group. While diversity in terms of beliefs about evaluation practice may have been valuable to thorough interrogation of presenting problem and capacity building needs, it was not particularly useful for coming up with action steps outside of the STEM EC itself. Although the EDC team suggested that the plurality might yield different actions to address different capacity building needs, it did not seem sticky enough to stay at the forefront of people's minds. Indeed, after the second convening, one participant wrote, "[The] diversity [in this group] probably takes 10x the amount of effort than anyone realizes" (Participant Reflection, Convening 2).

It's difficult to figure out how to skillfully work with the level of diversity present in the STEM EC. As noted at several junctures in this report, while the diversity in the group was acknowledged, it was not really engaged with, and some participants did not feel their perspectives were fully embraced by the community. Following the third convening, one person reflected, "It's a very mixed bag in my view. It seems like there is a lot of wisdom in the group that is simply not getting surfaced."

Beyond the immediate and potential value, the work of the STEM EC yielded applied value for some of its members and beyond. Several new connections were formed, leading to opportunities such as new webinars at PRCs and collaboration on grant proposals. The Landscape Study as well as broad learnings from the STEM EC project were shared with people outside the project through AEA and informally through participants' social networks. Several STEM EC members shared how they had taken what they had learned through participating in the STEM EC to their own evaluation practice such as attending to building community across project evaluators working under the same program. For the most part, applied value was created out of individual actions and initiatives with the exception of the AEA think tank and the 365 Blogs.

The second question guiding this evaluation, "What evidence do we have that strategies employed by the STEM EC have built capacity of the community to conduct methodologically appropriate and conceptually sound evaluations?" ends up not being particularly useful at this point in the STEM EC's evolution. As noted repeatedly, the community's energy that might have been applied to capacity building was side-tracked by an examination the assumptions embedded in how the problem and needs for capacity building were framed, questions about what the STEM EC's domain was (i.e., NSF, evaluation of all types of STEM including research, evaluation writ large), and a recognition that the needs identified depended on where the person situated themselves vis a vis evaluation practice.

That said, it can be argued that the open-ended process yielded better understanding of problems related to evaluation of STEM outreach and education initiatives, particularly those out of NSF. The conversations served to push the STEM EC to think beyond the usual capacity

building responses of setting up more instrument repositories and writing more how-to guides. The benefits people experienced participating in the STEM EC showed how investing in people and making spaces for cross-fertilization have the potential to deeply influence practice. Remember from the Evaluator Survey, it was the networks and relationships with trusted colleagues that seasoned evaluators turned to for advice on messy middle type problems related to ethics, evaluation use, and relationships, all aspects of evaluation that affect quality.

The evidence we have that the STEM EC has enhanced the community's capacity to conduct better evaluations, perhaps, is found in the reframing of capacity building needs from things to people. Better questions might be "How do we demonstrate a commitment to 'people as resources' and what does that mean for investments in STEM evaluation?" With this in mind, capacity building might become more about creating opportunities for people to come together to work on specific problems be they through the PRCs or new solicitations that push the field of STEM evaluation forward through cross-fertilization between projects, programs, and stakeholder perspectives. This also means providing tangible resources that make it possible for people to participate such as institutional investment and support.

REFERENCES

Abma, T.A. (Ed.). (1999).*Telling tales: On evaluation and narrative. Advances in program evaluation*. Stamford, CT: JAI Press, Inc.

Abma, T.A. (2006). The practice and politics of responsive evaluation. *American Journal of Evaluation 27*(1): 31-43.

Greene, J. and Abma, T.A. (Eds.). (2001). *Responsive evaluation: New directions for evaluation 92*. San Francisco: Jossey-Bass.

Kaminsky, A. (1993). Participatory evaluation in hierarchies: Practical dilemmas of implementation (A case study). (Unpublished master's thesis). Cornell University, Ithaca, New York.

Schwandt, T. (2018). Evaluative thinking as a collaborative social practice: The case of boundary judgment making. In A. T. Vo & T. Archibald (Eds.), Evaluative Thinking. New *Directions for Evaluation. 158*, 125–137.

Stake, R.E. (2003). Responsive evaluation. In T. Kellaghan and D.L. Stufflebeam (Eds.), *International handbook of Educational Evaluation,* (pp. 63-68). Dordrecht: Kluwer Academic Publishers.

Wenger, E., McDermott, R., Snyder, W.M. (2002). *Cultivating communities of practice.* Boston: Harvard Business Review Press.

Wenger, E., Trayner, B., and de Laat, M. (2011). *Promoting and assessing value creation in communities and networks: A conceptual framework.* The Netherlands: Ruud de Moor Centrum.

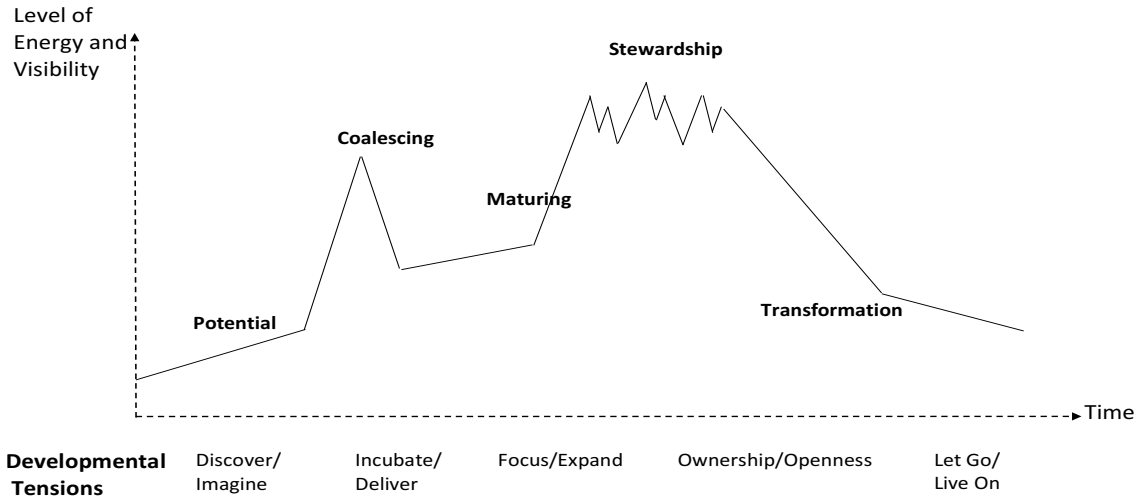
Wenger-Traynor, E., Fenton-O'Creevy, M., Hutchinson, S., Kubiak, C., and Wenger-Trayner, B. (2015). *Learning in landscapes of practice: Boundaries, identity, and knowledgeability in practice-based learning*. London: Routledge Publishers.

Appendix A: Value Created and Stages of Development in Communities of Practice

Cycles of value creation in communities of practice (based on Wenger et al's 2011 typology)

Cycle	Focus	Indicators/metrics
Immediate Value	Activities and interactions	Levels of participation and engagement; quality of interactions; value of participation; opportunities for collaboration; new connections made; frequency of interaction
Potential Value	Knowledge capital (includes: human capital, social capital, tangible capital, reputational, and learning capital)	Skills developed; information received; resources developed; perspective changes; changes in inspiration and confidence; quality of relationships; changes in networks that make up community of practice; value to individual and to community
Applied Value	Changes in practice	Implementation of new knowledge and practices; innovation in practice, theory and ways of thinking about practice; reuse/repeated use of new knowledge, practices, resources; new collaborations and connections (i.e., changes in network); expansion or transfer of learning to wider community; changes in institutional or systemic context (i.e., cross-silo pollination); dissemination of learning
Realized Value	Performance improvement	Interaction across PRCs reduce duplication of efforts; shared information across PRCs yields more robust understanding of appropriate evaluation design for given project; new evaluators access resource base; sustainable infrastructure for supporting resource base and sharing realized
Reframed Value	Redefining success	Aspirational narratives; new assessment metrics/processes; stakeholder relationships; institutional changes (i.e., less duplication, more collaboration and coordination, common as well as unique metrics for assessment); new frameworks

Communities of Practice: Stages of Development



Source: Wenger, E., McDermott, R., & Snyder, W.M. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Boston: Harvard Business Review Press.

Appendix B: Original Theory of Change Model Developed