



HHS Public Access

Author manuscript

New Dir Eval. Author manuscript; available in PMC 2023 August 03.

Published in final edited form as:

New Dir Eval. 2022 ; 2022(174): 11–20. doi:10.1002/ev.20506.

Understanding the context and appreciating the complexity of evaluating the Diversity Program Consortium

Lourdes R. Guerrero,

Teresa Seeman,

Heather McCreath,

Nicole M.G. Maccalla,

Keith C. Norris

University of California San Diego, La Jolla, California, USA

Abstract

The National Institutes of Health (NIH) made a sizeable investment in developing a scientific approach to understanding how to best increase diversity in the NIH-funded workforce by fostering inclusive excellence at a national scale through the Diversity Program Consortium (DPC). This chapter provides an overview of the context in which the consortium-wide evaluation study has taken place to provide readers with an understanding of its level of complexity. This evaluation effort is the first large-scale, national, systemic, longitudinal evaluation of harmonized interventions focused on undergraduate biomedical research training programs in the history of the NIH and the National Institute of General Medical Sciences.

INTRODUCTION

New approaches to the evaluation of programs and initiatives are needed to meet the demands of large-scale biomedical diversity efforts (Davidson et al., 2017; Valantine et al., 2016). The National Institutes of Health (NIH) made a sizeable investment in developing a scientific approach to understanding how best to increase diversity in the NIH-funded workforce by fostering inclusive excellence at a national scale through the Diversity Program Consortium (DPC). The DPC is a national collaborative research project in which the NIH works together with institutions to advance the overarching goal of developing, implementing, assessing and disseminating innovative, effective approaches to engaging, training and mentoring students, enhancing faculty development, and strengthening institutional research and research training infrastructure. Through this national research project, grantees implement approaches intended to improve research training, mentoring, faculty development, and institutional capacity building and conduct local evaluations of site-level processes and outcomes. By contrast, the national evaluation of this effort analyzes outcomes based on multisite implementation of promising interventions designed to enhance

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Correspondence: Guerrero, Lourdes, University of California San Diego, La Jolla, CA, USA. loguerrero@health.ucsd.edu.

outcomes for students from backgrounds traditionally underrepresented in biomedical sciences and disseminating lessons learned to the broader extramural scientific community (Hurtado et al., 2017).

The role and function of the Coordination and Evaluation Center (CEC) has been to design, collect, and report evidence regarding the outcomes of this federally funded intervention across multiple study sites (McCreath et al., 2017). This integrated DPC evaluation, known as the Enhance Diversity Study (EDS), is significant because it is the first large-scale, systemic, national, longitudinal evaluation of harmonized interventions targeting undergraduate biomedical research training programs in the history of the NIH and the National Institute of General Medical Sciences (Hurtado et al., 2017). The size and scope of the EDS is fertile ground for equity evaluation. It allows us to look at the wider impact of the proposed initiatives, make evident inequities, and help us understand how to address different groups in society that are affected by the interventions (Carden, 2017). Linking site-specific DPC interventions and practices through evaluation research can facilitate the identification of exemplars that others can adopt or avoid to advance student equity in biomedical science careers and, ultimately, address challenging health problems in an increasingly diverse nation.

BACKGROUND

The CEC was supported as part of a series of funding opportunity announcements from NIH's Common Fund. The three related notices are RFA-RM-13-015, the NIH CEC for Enhancing the Diversity of the NIH-Funded Workforce Program (U54); RFA-RM-13-016, the NIH BUILD Infrastructure Leading to Diversity (BUILD) Initiative (U54); and RFA-RM-13-017, the NIH National Research Mentoring Network (NRMN; U54). The NIH created the DPC as a prospective initiative to implement and evaluate several novel interventions designed to enhance outcomes for students from backgrounds traditionally underrepresented in the biomedical sciences within and across a variety of US academic institutions and their key partner institutions. The CEC was funded simultaneously with the other two initiatives, BUILD and NRMN. This meant that the consortium-wide evaluation plan was developed as programs were initiating their start-up phases and was based on emerging interventions and program activities that were not yet fully developed.

Because the NIH had issued P20 planning grants (RFA-RM-13-002 and RFA-13-002) for the BUILD and NRMN initiatives, some (but not all) awardees had planning time prior to the full launch of the consortium. Moreover, these planning award grantees were expected to plan and evaluate interventions and activities between BUILD and NRMN, meaning some sites were more aware and integrated with consortium-wide activities than others. Thus, from the very beginning, there were multiple implementation and evaluation efforts underway simultaneously.

Given the multiplicity of efforts that were in place at the outset of the DPC, the consortium-wide evaluation plan was developed with a focus on examining key transition points, termed *Hallmarks of Success*. These transition points exist along the trajectory from undergraduate student to independently funded investigator, which is inclusive of both BUILD and NRMN

populations. To optimize potential survey responders' awareness of the consortium-wide evaluation survey, it was eventually renamed and branded as the EDS in hopes that the title would engender more interest and relevance leading to higher response rates (see Chapter 3 by Ramirez et al. for more detail).

The BUILD initiative is designed to provide prospective, multisite, evidence-based practices around several key predictors of undergraduate biomedical students' degree completion and transition to biomedical careers. These predictors are viewed as critical to overcoming explicit and implicit barriers to excellence for underrepresented students, thereby increasing diversity in the biomedical research workforce. The external evaluation of the BUILD programs is broadly framed as equity-focused impact evaluation and uses a longitudinal, quasi-experimental, multimethod design, collecting data through a variety of approaches, including surveys and case studies. The main aims of the longitudinal BUILD evaluation were based on logic models designed at the student, faculty, and institutional levels, given the interventions and programmatic activities designed for these groups. Yet each site has implemented varying strategies and activities for these groups (see Chapter 5 by Maccalla et al. for more details on how activities were categorized). Moreover, many BUILD sites were required to develop partnerships with top-tier research institutions and with pipeline institutions (e.g., community colleges) to implement their interventions and programmatic activities. Just as program designs varied, the nature of these cross-institutional partnerships varied. Some built on existing relationships and others leveraged BUILD funding to engage in new partnerships.

During the first 5 years of the DPC, the CEC was also involved in evaluating NRMN. The NRMN evaluation focused on addressing questions related to mentor and mentee skills, research self-efficacy, trainee experiences, and early-career research outcomes. A common set of shared measures was identified for use by all program components so that data could inform both short- and long-term evaluation questions. Some BUILD activities overlapped with NRMN activities, but NRMN efforts did not entirely fit into the logic models designed for BUILD. Thus, although the NRMN evaluation was part of the consortium-wide evaluation plan, these data collection efforts were separate from those of BUILD, with independent logic models and outcomes that will not be discussed extensively here.

The CEC employs a context-sensitive consortium-wide evaluation approach to navigate the complexity of the DPC and engage with BUILD/NRMN partners. This approach allows for flexibility, accommodating both common and unique program features and working styles across the DPC. Given that programs naturally evolve as learning occurs and environments change, a highly systematic yet responsive and adaptive approach to evaluation over the 10-year study is employed. In designing, conducting, and describing our work, we pull from context-sensitive, theory-driven, utilization-focused, participatory, and equity evaluation frameworks (Alkin & Vo, 2017; Carden, 2017; Chen, 1990; Cousins & Earl, 1992; Hood et al., 2015; House & Howe, 1999; Marra & Forss, 2017; Mathison, 1994; Mertens, 2009; Patton, 1986). While the CEC is part of the DPC, it is external to BUILD program implementation and approaches the role of evaluator as a partnership (Mathison, 1994).

APPRECIATING THE COMPLEXITY

In several ways, the DPC, individual BUILD programs, NRMN, and the EDS are distinct from other diversity and STEM-related consortium work (e.g., Aspire Alliance, Project Kaleidoscope, Undergraduate STEM Education Initiative) that mainly work as a network of partners to support change but are not designed to systematically examine outcomes from structured interventions (AAC&U, 2022; AAU, 2017; NSF Aspire Alliance, n.d.). First, the national evaluation involves the assessment of multiple program initiatives, multisite implementation, and examines multilevel effects (i.e., student, faculty, institutional). Second, there is a collection of a large volume of primary data sources, including use of standardized evaluation tools for outcome and process measures. Third, the study duration and set of evaluation resources have been extensive. Although distinct in these ways, the CEC can relate to common challenges with conducting multisite, multiprogram evaluations, including variability across sites in local contexts, intervention designs, evaluation approaches, data management, and site-level resources (Lewis et al., 2009; Stainbrook et al., 2015). As noted in the literature, a high level of heterogeneity can make it challenging to standardize research protocols and generalize findings (Constantine & Cagampang, 1998; Rog & Ponirakis, 2002; Stainbrook et al., 2015).

Next, we outline some of the notable complexities encountered with conducting the rigorous assessment of DPC interventions and describe the approaches we took to managing them.

APPROACH TO THE DPC EVALUATION

As more programs are funded to address the growing awareness of and interest in social justice, evaluators may find themselves struggling to fully implement evaluation approaches that have equity at their center (e.g., culturally responsive evaluation, deliberative democratic evaluation, inclusive evaluation, Indigenous evaluation, transformative evaluation) (Hood et al., 2015; House & Howe, 2000; LaFrance & Nichols, 2008; Mertens, 1999; Mertens, 2009). Programs that can be viewed through the lens of a DEI framework (i.e., changing conditions for historically marginalized groups) often face challenges implementing equity evaluation by still operating within contexts that have strong preferences for measuring expected changes in outcomes, expected use of evaluation theory, and preferences for methodological approaches and products that establish or approximate causality (Carden, 2017). The tensions that Christie and Guerrero described in the editors' notes in this issue can be difficult to navigate in large-scale, high-profile evaluation studies such as the EDS.

While the evaluation field continues to partner with funders and programs to create conditions conducive to the application of social justice approaches to evaluation, equity-minded evaluators can play a vital role in ushering in incremental change, one program evaluation at a time. Doing our part to bring about a fair and just society requires that we infuse principles of equity evaluation into more traditional evaluation studies, embracing what Christie and Wright have described as “the role of the evaluator as an advocate for social change” (Chapter 6, this issue). Various chapters in this issue address how this was accomplished and what opportunities remain.

SCOPE OF THE DPC EVALUATION

Originally, the CEC was charged broadly with evaluating all components of the DPC, but it quickly became obvious that a more focused definition of the evaluation scope was required to avoid duplication of site-level and NRMN internal evaluation efforts. Thus, several key decisions were made. First, CEC survey activity for a given BUILD program would focus only on students and faculty at the program's primary BUILD institution. The nature of partnerships with other institutions varied widely between programs, making the identification of appropriate comparison groups difficult; the evaluations conducted by the BUILD programs incorporated assessment of the partnerships. In addition to logistical complexity, conducting surveys at these additional institutions would have been very costly and results would have been difficult to interpret. Thus, we thought our efforts would be better used to gather more program-specific data. These institutional partnerships were explored in our qualitative inquiry through the BUILD case studies, which are described by Cobian et al. in Chapter 2.

Second, rather than duplicate the survey activity of NRMN participants, NRMN and CEC leadership determined that a sequential, collaborative evaluation would be the best approach. We agreed that the NRMN Professional Development Core would conduct the short-term evaluation of their programs (up to 18 months post-participation, depending on the activity) and the CEC would conduct a longitudinal evaluation. Reaching consensus on these scope refinements was a demanding process that necessitated much negotiation; it required balancing competing expectations and initial grant-proposed scope-of-evaluation work that varied across BUILD and NRMN consortia members. BUILD programs varied in the completeness of their local site evaluations at the time of funding, so some still had significant work to refine measures and design approaches during the first year of funding.

Many groups were also not expecting to work so extensively with the CEC on consortium-wide evaluation efforts, so discussions took place to clarify expectations in an effort to avoid overlapping with local evaluation work-scope efforts and to harmonize assessments across institutions. The CEC also provided technical assistance to sites as needed, since evaluation capacity and experience differed between programs and the NIH specifically requested that the CEC review each program's logic model in the first year. Unfortunately, programs did not perceive this review as assistance but as oversight; as such, it established distrust and created a perceived hierarchical relationship between the CEC and the BUILD evaluation teams that took years to overcome.

Another challenge with the scope of the evaluation was the length of time needed to reach meaningful outcomes. Identification of significant predictors for important intermediate Hallmarks of Success can help to inform recommendations for program effectiveness, but graduation rates and completion of advanced degrees are the primary outcomes of interest to NIH. Undergraduates at most institutions implementing BUILD interventions take, on average, approximately 6 years to acquire their bachelor's degrees (Davidson et al., 2017); students in graduate programs typically require 2–8 years to complete. Thus, it is important to collect data from students for at least 6–8 years, and ideally longer, to better assess graduation and transition to graduate school. As of now only five cohorts of students are

enrolled in the DPC evaluation limiting the number who have transitioned to graduate school.

By providing a total of 10 years of funding, the NIH acknowledged the inherent challenge of needing robust numbers of BUILD participants going to and then graduating from graduate school to produce a robust evaluation. However, evaluating these interventions will require ongoing follow-up beyond the present 10-year funding period. More time is needed to track the longer-term outcomes for students who engaged in the BUILD programs as undergraduates (many started as freshmen). Doing so will enable assessment of the relative extent to which those students (compared to non-BUILD peers) pursue careers in biomedical research and are tracking to become NIH-funded researchers.

HETEROGENEOUS INTERVENTIONS

DPC interventions varied widely, and some were modified slightly over time based on feedback from the local evaluations conducted by programs, creating an adaptive-style intervention. This resulted in a set of challenges for the DPC evaluation, including describing the commonalities of programs while maintaining variability in context and process. For BUILD programs, this challenge and the accompanying solutions are discussed in Chapter 5. Another challenge was determining the appropriate comparison group. Some programs identified incoming freshmen for programs, so it was important to collect information from students as they enrolled at the institution. Other programs identified student participants as they engaged with the coursework in their selected majors, so it was important to have a sample of students who had navigated the early college years in a similar way to BUILD student participants. Analysis approaches are discussed by Crespi and Cobian in Chapter 4.

DATA COLLECTION APPROACHES AND CHALLENGES

Central to the CEC's evaluation of the DPC is a design based on implementation of annual surveys for students and faculty at each BUILD institution to track outcomes of interest. In order to understand whether or not BUILD exposure (i.e., engagement with one or more of the BUILD intervention activities) resulted in improved outcomes for students and faculty, the surveys were administered to all those engaged with BUILD as well as other students and faculty at the BUILD institutions. Depending on the size of the institution, either *all* other students or a *sample* of other students were invited to participate. For faculty, the focus was on those in biomedical fields (natural sciences, social sciences, and engineering). Thus, we invited all faculty involved with the BUILD program and enough who were not involved with the BUILD program to get a sample of 50–100, depending on the number available at the institution. The core evaluation entails comparisons of those with and without BUILD exposure. We used surveys administered nationally at colleges and universities by the Higher Education Research Institute (HERI) for a portion of our data collection.¹

¹For more information about HERI, see <https://heri.ucla.edu/>

Institutions choose to participate in HERI surveys, but the content is stable across all institutions in the EDS. HERI data have been collected for over 50 years, so they provide a rich context for understanding the experiences of undergraduate students and faculty on a wide variety of campuses. Surveys available for students include The Freshman Survey (administered in the beginning weeks of each fall term to entering students), The College Senior Survey (administered to graduating seniors during winter/spring terms), and The HERI Faculty Survey (administered every 3 years). These surveys collect demographic data and key attributes around the main constructs being assessed (e.g., teaching, learning and mentoring experiences, research activities, scientific productivity, perceptions of institutional climate).

The hope was to use these surveys for longitudinal comparisons for intervention campuses (prior to and during the program) as well as for comparisons with peer campuses not involved in the intervention. Practically, however, several campuses with BUILD programs did not historically use these surveys, so longitudinal comparisons have not always been possible. It has also been difficult to identify comparable campuses not involved in the intervention for all years of the evaluation. Another challenge with using HERI surveys relates to the need to collect data over the course of undergraduates' college experiences. One possible survey, The Diverse Learning Survey, did not include enough of the measures identified as DPC Hallmarks of Success to match the developed logic model. Therefore, the CEC needed to develop in-house surveys to collect data between the first and final college years so that we could build a longitudinal dataset with yearly data points. Additionally, we needed to survey students who left school (graduated or not) and moved on to the biomedical workforce or training in advanced degree programs. In-house, CEC-tailored surveys were designed to address these needs as well.

One additional challenge early in the development of the DPC evaluation involved a requirement for Office of Management and Budget (OMB) clearance. Federal OMB clearance was required given the large scale of the DPC evaluation. Assembling the OMB proposal package and moving through the approval process took nearly 14 months, even with NIH assistance to expedite. This delayed the CEC primary data collection efforts during the first two years of funding.

Fortunately, our partners at BUILD institutions understood the importance of obtaining these data and they worked with the NIH and HERI to launch surveys during those first 18 months. This was a significant unanticipated burden to groups already working to implement their own wide-ranging intervention programs. In retrospect, as with development of the DPC scope, allowing a planning year prior to the intervention launch would have allowed enough time to complete the OMB approval process. This should be considered for any future NIH studies that may require OMB approval.

In addition to the local BUILD evaluations being conducted, the BUILD sites are part of institutions engaged in formative change. They are therefore collecting institution-wide data from students and/or faculty to inform their actions. Thus, every year the CEC must negotiate with institutional offices administering an increasing array of non-BUILD

campus wide surveys as well as with BUILD program evaluators to determine a survey administration timeline that does not compete with or overburden students or faculty.

ENGAGING STUDENTS AND FACULTY

While the DPC evaluation was intended to be complementary to individual program evaluation, it was critical and necessary for each program to assist the CEC in regularly engaging students and faculty. The CEC had no direct contact with students or faculty at these institutions, so local BUILD team members were needed to help identify campus “influencers” (i.e., known institutional representatives who could reach out to students and faculty to explain the BUILD study, why it was important to their institution, and request their participation). As noted earlier, details of engagement processes and challenges are described by Ramirez et al. in Chapter 3.

LIMITATIONS AND OUTSIDE FORCES

In addition to the complexity outlined above, this evaluation study had several limitations. First, the design is quasi-experimental, which necessitates the ability to “control” for confounding effects, which, within the national evaluation, proves circuitous. This is because each site has different (but similar) interventions and different supplemental student biomedical science enhancement efforts that could lead to spillover with the BUILD intervention. Similarly, comparison students within a BUILD site as well as comparison institutions may be exposed to a variety of programs and services designed to assist their career paths. And finally, the BUILD interventions are robust and can have their own spillover across BUILD campuses, not only at the student level but also at the faculty and institutional leadership levels.

Moreover, several outside forces may have impacted this evaluation study. First, the DPC and consortium-wide evaluation activities were well underway when the 2016 presidential election led to a change in leadership at the White House—a change that included open support for groups promoting white supremacy ideologies and for similar narratives and policies, such as efforts to eliminate equity efforts for non-*cis*-White males. This peaked with Executive Order 13950 of September 22, 2020, which banned diversity training on systemic racism and critical race theory among federal agencies.

Second, the COVID-19 pandemic began in early 2020 and had a massive direct impact on education, with school closures, remote learning, and/or campus limitations on in-person activities to ensure the health of students, faculty, and staff. Further, the pandemic highlighted the endemic yet subtle nature of structural racism and the nation’s caste system, with disproportionately high numbers of hospitalizations and deaths among minoritized communities. This further levied burdens upon many students and faculty at BUILD institutions (Krieger, 2020). Despite effective vaccines, the COVID-19 pandemic continues to impact the world due to emerging variants, suboptimal vaccine uptake (Robinson, Jones, Lesser, & Daly, 2021), and global maldistribution of vaccines, which helps to sustain coronavirus variant development (Schaefer, Leland, & Emanuel, 2021). Outside forces

disrupted consistency in program delivery and data collection procedures and may have introduced threats to internal validity based on historical events.

LESSONS LEARNED

The complexity and limitations experienced in the development and implementation of this evaluation have facilitated the learning of many lessons. First, we cannot emphasize enough the importance of setting expectations among all involved for actual “consortium” partnerships at the beginning of such efforts. The level of required interaction and partnership was not clear in the initial Request for Applications for this initiative. Thus, we have been struggling to ensure seamless and supportive collaboration for the consortium-wide evaluation plan. It is unclear whether or not, or to what degree, a true consortium has been created. Second, the EDS is not the foremost survey being completed by DPC participants. Local BUILD and NRMN program surveys take precedence among participants. Managing competing surveys has been an ongoing effort within the consortium. A lack of coordination has diminished the overall sample size and associated statistical power (as a result of low response rates) of this national evaluation.

The CEC has also lacked the ability to connect and engage with prospective survey participants before they become BUILD participants. Some BUILD programs have become better at promoting the importance of the national evaluation study, but future efforts should ensure that institutions themselves (not just programs) engage in this effort, especially if they are attempting to implement an experimental or quasi-experimental design. Like many others implementing survey-based research, this evaluation has been fighting the general trend toward lower participation in surveys. The main solution to this issue seems to be offering higher and higher monetary incentives, but this is not always a sustainable option.

Last, we underscore the importance of having an adequate amount of time at the beginning of such a large effort to build relationships and community amongst evaluators so that trust and understanding are established. Taking the time to ensure collaboration across sites and with the CEC is a key to success that we could not undertake early on, and the long-term consequences of this are notable.

CONCLUSION

The role and function of the CEC has been to design, collect, and report evidence regarding the outcomes of DPC initiatives. This consortium-wide, national evaluation was both complex and promising from the beginning. This chapter has outlined some of the complexities, challenges, and opportunities that have impacted this evaluation effort. Despite these, we have gained many insights and learned many lessons. We are confident that our efforts will indeed provide an understanding of site-specific DPC interventions and practices that advance student excellence and equity in biomedical science careers and, ultimately, assist others in evaluating such efforts in the future.

The Diversity Program Consortium Coordination and Evaluation Center at UCLA is supported by the Office of the Director of the National Institutes of Health/National Institutes of General Medical Sciences under award number U54GM119024.

Biographies

Dr. Lourdes Guerrero was associate adjunct professor at the David Geffen School of Medicine at UCLA and served as an associate lead for the Evaluation Core of the Coordination and Evaluation Center for the Diversity Program Consortium. She is currently a researcher at UC San Diego.

Dr. Teresa Seeman is a professor of medicine and epidemiology, and associate chief for equity, diversity and inclusion at UCLA's David Geffen School of Medicine. She serves as multiple principal investigator for the Coordination and Evaluation Center and co-director of the Data Coordination Core for the Diversity Program Consortium.

Dr. Heather McCreath is an adjunct professor in the Division of Geriatrics in the David Geffen School of Medicine at UCLA and serves as co-director of the Data Coordination Core in the Coordination and Evaluation Center for the Diversity Program Consortium.

Dr. Nicole Maccalla is housed in UCLA's School of Education and Information Studies and serves as a lead investigator for the Coordination and Evaluation Center's Enhance Diversity Study, as well as the BUILD evaluation study coordinator and chair of the Evaluation Implementation Working Group.

Dr. Keith Norris is a professor and executive vice chair for equity, diversity and inclusion for the Department of Medicine at UCLA and serves as multiple principal investigator for the Coordination and Evaluation Center and director of the Administrative Core for the Diversity Program Consortium.

REFERENCES

- Alkin MC, & Vo AT (2017). *Evaluation essentials: From A to Z* (2nd ed.). Guilford Press.
- American Association of Colleges & Universities. (2022). Project Kaleidoscope (PKAL). Retrieved from <https://www.aacu.org/initiatives/project-kaleidoscope>
- Association of American Universities. (2017). Progress towards achieving systemic change: A five-year status report on the AAU Undergraduate STEM Education Initiative. Retrieved from <https://www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/STEM-Status-Report.pdf>
- Carden F (2017). Building evaluation capacity to address problems of equity. *New Directions for Evaluation*, 154, 115–125. 10.1002/ev.20245
- Chen HT (1990). *Theory-driven evaluations*. Sage Publications.
- Constantine N, & Cagampang H (1998). Improving local evaluation utility within multi-site risk behavior prevention program evaluations [Paper presentation]. 1998 Annual Conference of the American Psychological Association, San Francisco, CA, United States.
- Cousins JB, & Earl LM (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis*, 14(4), 397–418. 10.3102/01623737014004397
- Davidson PL, Maccalla NMG, Afifi AA, Guerrero L, Nakazono TT, Zhong S, & Wallace SP (2017). A participatory approach to evaluating a national training and institutional change initiative: The BUILD longitudinal evaluation. *BMC Proceedings*, 11(Suppl. 12), 15. 10.1186/s12919-017-0082-9 [PubMed: 29375658]
- Hood S, Hopson R, & Kirkhart K (2015). Culturally responsive evaluation. In Newcomer KE, Hatry HP, & Wholey JS (Eds.), *Handbook of practical program evaluation* (4th ed., pp. 228–317). John Wiley & Sons.
- House ER, & Howe KR (1999). *Values in evaluation and social research*. Sage Publications.

- House ER, & Howe KR (2000). Deliberative democratic evaluation. *New Directions for Evaluation*, 85, 3–12. 10.1002/ev.1157
- Hurtado S, White-Lewis D, & Norris KC (2017). Advancing inclusive science and systemic change: The convergence of national aims and institutional goals in implementing and assessing biomedical science training. *BMC Proceedings*, 11((Suppl.)12), Article 17. 10.1186/s12919-017-0086-5
- Krieger N (2020). ENOUGH: COVID-19, structural racism, police brutality, plutocracy, climate change—and time for health justice, democratic governance, and an equitable, sustainable future. *American Journal of Public Health*, 110(11), 1620–1623. 10.2105/AJPH.2020.305886 [PubMed: 32816556]
- LaFrance J, & Nichols R (2008). Reframing evaluation: Defining an Indigenous evaluation framework. *The Canadian Journal of Program Evaluation*, 23(2), 13–31.
- Lewis A, Brubaker SJ, Karph AS, & Ambrose B (2009). The Virginia Abstinence Education Initiative evaluation structure: A lesson in how to successfully overcome the challenges of multi-site program evaluation. *Journal of Youth Development*, 4(3), 43–51. 10.5195/JYD.2009.251
- Marra M, & Forss K (2017). Thinking about equity: From philosophy to social science. In Forss M (Ed.), *Speaking justice to power* (pp. 21–33). Routledge.
- Mathison S (1994). Rethinking the evaluator role: Partnerships between organizations and evaluators. *Evaluation and Program Planning*, 17(3), 299–304. 10.1016/0149-7189(94)90009-4
- McCreath HE, Norris KC, Calderón NE, Purnell DL, Maccalla NMG, & Seeman TE (2017). Evaluating efforts to diversify the biomedical workforce: The role and function of the Coordination and Evaluation Center of the Diversity Program Consortium. *BMC Proceedings*, 11(Suppl. 12), Article 27. 10.1186/s12919-017-0087-4
- Mertens DM (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation*, 20(1), 1–14. 10.1177/109821409902000102
- Mertens DM (2009). *Transformative research and evaluation*. Guilford Press.
- National Science Foundation INCLUDES Aspire Alliance. (n.d.). *Aspire: The National Alliance for Inclusive & Diverse STEM Faculty*. Retrieved from <https://www.aspirealliance.org/>
- Patton MQ (1986). *Utilization-focused evaluation* (2nd ed.). Sage Publications.
- Robinson E, Jones A, Lesser I, & Daly M (2021). International estimates of intended uptake and refusal of COVID-19 vaccines: A rapid systematic review and meta-analysis of large nationally representative samples. *Vaccine*, 39(15), 2024–2034. 10.1016/j.vaccine.2021.02.005 [PubMed: 33722411]
- Rog D, & Ponirakis A (2002). Cross-site evaluation of the CMHS/CSAT homeless family initiative: Challenges and opportunities [Conference paper]. 2002 Annual Meeting of the American Public Health Association, Philadelphia, PA, United States.
- Schaefer GO, Leland RJ, & Emanuel EJ (2021). Making vaccines available to other countries before offering domestic booster vaccinations. *JAMA*, 326(10), 903–904. 10.1001/jama.2021.13226 [PubMed: 34382999]
- Stainbrook K, Penney D, & Elwyn L (2015). The opportunities and challenges of multi-site evaluations: Lessons from the jail diversion and trauma recovery national cross-site evaluation. *Evaluation and Program Planning*, 50, 26–35. 10.1016/j.evalprogplan.2015.01.005 [PubMed: 25723338]
- Valantine HA, Lund PK, & Gammie AE (2016). From the NIH: A systems approach to increasing the diversity of the biomedical research workforce. *CBE Life Sciences Education*, 15(3), fe4. [PubMed: 27587850]