Journal of MultiDisciplinary Evaluation
Volume 13, Issue 29, 2017

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

Standards of Excellence from a Leader of Excellence: Honoring Our Professional Heritage and Remembering Daniel L. Stufflebeam

Michael Quinn Patton
Utilization Focused Evaluation

Introduction

On the passing of a distinguished evaluation pioneer like Daniel Stufflebeam, it is not only appropriate to recognize his contributions, but incumbent upon us to do so. A special section of the *American Journal of Evaluation*, curated by Chris Coryn (2017), does just that. In this memorial reflection, I want to pay homage not only to the man but to celebrate the evaluation standards he fathered. The standards define what program evaluation is. After nearly 40 years it is easy to take them for granted, but to do so impoverishes the profession's historical journey and treats as inevitable a creation that almost was not, and would not have been, but for Dan Stufflebeam.

It is hard now to imagine the profession of evaluation without the standards, but I remember the time before, a time when evaluations were judged entirely by traditional research standards, essentially methodological and technical criteria. The breakthrough significance of the joint committee standards was to define the unique niche and contributions of the emergent field of program evaluation (Patton, 1994). They constituted a magnificent articulation of evaluation's potential. As such, they deserve to be savored (Joint Committee, 1994).

Read them slowly, spend some time appreciating their blend of wisdom, aspiration, idealism, pragmatism, and practicality. And ponder how they came to be. It is what in today's world is called a "back-story." In remembrance of Dan Stufflebeam, let us bring the story forward into the light, retell it for a new generation, and in so doing, honor the generation that gave us this precious and enduring legacy. So, gather round the evaluation campfire, children, and let me tell you a tale of wonder.

Once Upon a Time

Once upon a time, there were no programs, and therefore, no evaluations. Then people called policymakers, philanthropists, change agents, and their kin began creating and implementing interventions aimed at changing things and helping people in need. Those interventions were dubbed "programs." Those who funded them came to wonder about their effectiveness, so researchers did what they do – conducted research. Evaluation research. But the usual approach to academic research often failed to provide meaningful and useful answers about programs' merit, worth, and significance. Evaluation research was judged by technical quality and methodological rigor. Use was ignored. Methods decisions dominated the evaluation design process. Methodological rigor meant experimental designs, quantitative data, and sophisticated statistical analysis. Whether decision makers understood such analyses was not the researcher's problem. Validity, reliability,

measurability, and generalizability were the dimensions that received the greatest attention in judging evaluation research proposals and reports (e.g., Bernstein and Freeman 1975). Indeed, evaluators concerned about increasing a study's usefulness often called for ever more methodologically rigorous evaluations to increase the validity of findings, thereby hoping to compel decision makers to take findings seriously.

By the late 1970s, however, it was becoming clear that greater methodological rigor was not solving the use problem. Program staff and funders were becoming openly skeptical about spending scarce funds on evaluations they couldn't understand and/or found irrelevant. Evaluators were being asked to be "accountable" just as program staff was supposed to be accountable. The questions emerged with uncomfortable directness: *Who will evaluate the evaluators? How will evaluation be evaluated?*

It was in this context that Dan Stufflebeam assembled a group of academic colleagues to generate standards for evaluation. The standards were hammered out over 5 years by a 17-member committee appointed by 12 professional organizations with input from hundreds of practicing evaluation professionals. The standards published by the Joint Committee on Standards in 1981 dramatically reflected the ways in which the practice of program evaluation had matured. Just prior to publication, Dan Stufflebeam, who had chaired the Joint Committee throughout its years of labor, summarized the committee's work as follows:

> The standards that will be published essentially call for evaluations that have four features. These are utility, feasibility, propriety and accuracy. And I think it is interesting that the Joint Committee decided on that particular order. Their rationale is that an evaluation should not be done at all if there is no prospect for its being useful to some audience. Second, it should not be done if it is not feasible to conduct it in political terms, or practicality terms, or cost effectiveness terms. Third, they do not think it should be done if we cannot demonstrate that it will be conducted fairly and ethically. Finally, if we can demonstrate that an evaluation will have utility, will be feasible and will be proper in its conduct, then they said, we could turn to the difficult matters of the technical adequacy of the evaluation. (Stufflebeam, 1980, p. 90)

In 1994 and 2007, revised Standards were published following extensive reviews spanning several years. While there some wording changes in the individual standards, the overarching framework of four primary criteria remained unchanged: utility, feasibility, propriety, and accuracy. Specific standards have also been adapted to various international contexts (Russon & Russon, 2004); the overall framework has translated well cross-culturally.

Taking the standards seriously has meant looking at the world quite differently. Unlike the traditionally aloof stance of purely academic researchers, professional evaluators are challenged to take responsibility for use. No more could we play the game of blaming the resistant decision maker. If evaluations are ignored or misused, we have to look at where our own practices and processes may have been inadequate. Implementation of a utility-focused, feasibility-conscious, propriety-oriented, and accuracy-based evaluation requires situational responsiveness, methodological flexibility, multiple evaluator roles, political sophistication, and substantial doses of creativity.

Stufflebeam chaired the national Joint Committee on Standards for Educational Evaluation during its first 13 years and was the principal author of the original Program Evaluation Standards and, subsequently, the Personnel Evaluation Standards (Stufflebeam, 2000b, 2004).

The Back-Story Brought Into the Light

When Stufflebeam took on the task of leading the development of evaluation standards, he set up a committee with the help of the presidents of AERA (American Educational Research Association), APA (American Psychological Association), and NCME (National Council on Measurement in Education). The group he recruited, convened, and facilitated consisted of eminent, distinguished, busy, opinionated, and often cantankerous academics. Their identities are part of the legacy and worthy of mention here, for these were the people Stufflebeam set out to guide to consensus. The initial Joint Committee included Egon Guba, Mitch Brickell, and Lorrie Shepard, as representatives of AERA; Donald Campbell, Robert Linn, and Wendell Rivers, as representatives of APA; and George Madaus, Ronald Carver, and Stufflebeam, himself, as representatives of NCME. At the first organizational meeting, Egon Guba looked around the table and asserted that those assembled were not credible to set standards for evaluations of the work of people in the trenches like teachers and principals. He asserted that evaluations of school

programs had a history of failure due to the inadequate and misguided approaches of the very evaluation specialists in that room who were now supposed to provide guidance on evaluation excellence. He argued that those who were the subjects and targets of evaluation – in-the-trenches educational practitioners – deserved a place at the table. That opening challenge generated heated argument about the value of adding evaluation users to the committee. Those opposed to such inclusion asserted that such practitioners lacked basic knowledge of evaluation and research methods and would therefore have little to contribute; they would either slow down or otherwise impede the committee's work. After much discussion, the group agreed to enlarge the committee to ensure equal representation between evaluation specialists and users.

Stufflebeam, recounting this history in his AJE oral history interview (The Oral History Project Team, 2008), confessed that he "dreaded the consequences of this decision, because we would now be adding representatives of teachers' and administrators' unions and could become a debating society that would fail to reach agreement on standards" (p. 561). Nevertheless, the committee was expanded to include 17 members, eight from the technical side, eighty from the user side, with Stufflebeam as chairman and "de facto referee" (p. 561). He recalled what happened next:

> At the next meeting in December of 1975, I sat at the head of a conference table. The practitioners were on my left and the methodologists on my right. Immediately, the practitioners took the opportunity to lambaste the evaluation specialists for all the harm that poorly designed studies had done in the schools. In that particular year, the National Education Association had called for a national moratorium on the use of standardized testing in schools. The National Education Association representative virtually shouted at one APA representative, who had been a vice president at Educational Testing Services for all the harm that organization's tests allegedly had caused in schools. I feared that Egon had led the original committee into what now appeared to be an untenable arrangement and that we would never succeed in setting standards. However, Donald Campbell came to the group's rescue. He wrote and proposed a set of rules designed to assure that everybody's voice would be heard and that needed decisions would then be made. A key rule was that every member of the committee would hold veto power over any standard. This quieted the storm, and the committee subsequently proceeded, over the years, to debate issues and decide on standards by consensus vote. It seems amazing that through the whole standards-setting process no member ever exercised the power of veto. This diverse group of users and doers of evaluation learned to value each other's perspectives….
>
> Another rule that Don Campbell wrote into our initial set of agreements was that the committee would not develop standards for evaluations of teachers and other education personnel. We agreed and decided not to touch this sensitive area, because the teacher representatives on the committee worried that the committee would stimulate punitive ways of evaluating teachers. We faced a large enough task in developing standards for evaluations of programs and decided not to pursue standards for other types of evaluations. Following our success in publishing the 1981 edition of the standards for program evaluations, the committee's teacher representatives stated at a follow-up meeting that the committee had learned to talk to and trust each other and had succeeded in developing something they all respected. They then reversed their previous resistance to developing standards for personnel evaluations and argued that the committee should develop such standards….
>
> Overall, I learned by working with education administrators and with the diverse group of members on the Joint Committee that any evaluation should reflect interactions with the intended users. It should address their questions and help them make constructive use of sound findings. At the same time, it is essential to remember that beyond addressing the questions of users, a sound evaluation must assess as fully as practicable a program's merit, worth, and propriety….Without the practitioners, the standards likely would have included much jargon and technical language. They probably would have lacked the real world examples that seem to ring true to practitioners and evaluators alike. And probably the committee would have developed a much longer book, with extensive technical detail. The practitioners on the committee wisely counseled that the document should be concise and designed for easy use by the full range of intended users. The committee's practitioners had a decidedly positive influence on development of the evaluation standards (Stufflebeam & Miller, 2008, pp. 561-2).

Evaluation Facilitation

As noted above, evaluation research began as a methodological enterprise. The standards elevated evaluation use to primacy and, in so doing, the importance of working with an evaluation's stakeholders and intended users. That was game-

changing. Interpersonal skills, especially facilitation skills, became essential (King, Stevahn, Ghere, & Minnema, 2001). As I write this, I have just completed a book on *Facilitating Evaluation* (Patton, 2018). I have studied facilitation, analyzed and constructed case examples of successful facilitation, identified and elaborated principles for effective facilitation, and developed exercises and training for enhancing facilitation. It is with some depth of reflection and study, then, that I assert that, to the best of my knowledge, Stufflebeam's 13-year facilitation of the Joint Committee on Evaluation Standards constitutes the exemplar, the hallmark, the pinnacle, dare I say, the standard, for effective evaluation facilitation. We have only glimpses of the discussions, debates, and negotiations that took place. But think of any group process you've been part of, quadruple the challenges faced, magnify the potential landmines umpteen times, and consider the barriers to be overcome – interpersonal dynamics, monumental egos, political differences, power differentials, historical antagonisms, and huge, huge stakes. Count me as in awe of Dan Stufflebeam's facilitation acumen. There's a dissertation waiting to be done going through the committee's documentation and identifying the negotiations facilitated, the facilitation skills demonstrated, and the lessons manifest in this unprecedented evaluation undertaking.

The Personal Factor

When asked how his own experience as a teacher influenced his thinking about the standards, Stufflebeam replied:

> Well, the main experience that profoundly influenced my thinking about evaluation and the standards occurred in 1961 when I was a substitute teacher in Chicago. I had no idea that things could be as bad as they were. I quickly came to believe that because I worked in over 40 schools I was probably the only person in Chicago who knew how bad the education crisis was in the wide range of schools. Most other teachers daily went into the same classroom. They knew about deficiencies in their schools but might have thought things were OK in the other schools. Some of the schools I worked in had more than 4,000 kids, with half the faculty being substitute teachers or other persons without teaching credentials. My experiences in Chicago occurred soon after the Hungarian Revolution and there were serious language barriers for these and other immigrants. Teachers and kids could not communicate with each other. Crime was rampant in many schools and the surrounding neighborhoods. There were conflicts within the schools between neighborhood gangs. Drug dealers roamed just outside the school yards. Pairs of police patrolled the halls of schools. Schools in ghetto areas had few curricular materials, partly because broken windows had to be repaired with money from the school's book fund. I thought that city and school district officials must be ignorant of how bad things were in many of the Chicago schools. Also, I worried that they might prefer not to be reminded of the sorry situation that likely was beyond their control. It seemed that responsible parties should have been looking into the schools and delivering needed corrections.
>
> As I recall, I subsequently wrote an article for one of my classes that called for systematic approaches to assessing schools…. Before school leaders could mount needed fixes, they would have to identify needs and problems in schools across their district…. My experience in the Chicago schools, even for only one semester, definitely influenced my work in developing the Joint Committee standards, which include a context analysis standard. (Stufflebeam & Miller, 2008, pp. 561-2)

Remaining Vigilant

In 2010 revised Standards were published by a new committee. During the 7 years of systematic review leading up to the revised standards, considerable debate centered around whether to change the order of the categories, making accuracy first instead of utility. Stufflebeam wrote to the revision committee articulating why the original order should be maintained. A portion of his letter is reproduced here. It is well worth reading for both its substance and the passion Dan shows for what the standards mean to the profession. He wrote:

> *The new sequencing of the categories of standards is illogical and a counterproductive break with both the JC Standards' historic rationale and the rationale's position in mainstream evaluation thinking and literature***.**
>
> The [proposed] new sequencing of the categories of *Standards* that places Accuracy ahead of Utility would, I believe, help return evaluation practice to the days, especially in the 1960s and 1970s, when evaluators produced many technically sound reports that only gathered dust on shelves. This would be a setback for the evaluation field in terms of wasting resources for evaluations in conducting many evaluations that make no difference. Such a return to producing technically elegant but irrelevant evaluation findings would also

impair the credibility of professional evaluators.

The re-sequencing of categories of standards ignores the historic case for the original sequencing of categories of standards, as Utility, Feasibility, Propriety, and Accuracy. Originally, that sequencing was recommended by Lee Cronbach and his Stanford U. colleagues. They argued that an evaluation's potential utility should be the first concern of evaluator and constituents, because evaluators should not waste time and money in designing, conducting, and reporting studies that would not be used. Once a study's potential utility is established, then it makes sense, in turn, to assure that the study is feasible to conduct in the particular setting, to subsequently make sure it can meet requirements for propriety, and ultimately to design it to produce accurate findings. The Joint Committee endorsed this rationale, particularly because it was a step toward assuring that scarce evaluation resources would be used to conduct sound evaluations that make a positive difference. Clearly, it makes no sense to engage in extensive technical planning of an evaluation before determining that the evaluation is worth doing.

The above rationale has been well accepted by those who are clients and users of evaluation studies, as seen in the influence by JC representatives of teachers, administrators, and policy makers on the first two editions of the Standards. I would be surprised and disappointed if such representatives on the JC go along with the new sequencing of the categories of standards. . . .

If the JC leads the evaluation field back to where it was in the 1960's, with initial higher order concern for accuracy over utility, I fear we will again see thick reports—based on sophisticated, rigorous methodology—gathering dust on shelves and having only the negative outcomes of wasting effort and resources on studies that make no impact. Given the scarcity of resources for evaluation, studies should be conducted only if they will be used. Moreover, the ones that are done should be squarely focused on the intended users' intended uses and most important questions.

Addressing in order the present Standards' sequence of utility, feasibility, propriety, and accuracy has proved very functional in my evaluations and, I think, in those of others. Here is an example. In approaching the next evaluation I have been asked to conduct, I have set aside questions of feasibility (also, propriety and accuracy) until the client and I can work out questions of intended users and uses. I have convinced the client to give me a small planning grant to clarify the intended users, intended uses, and information requirements and on that basis to prepare a plan and budget for what is needed. It would make no sense for me first to ask this Foundation how much money they can afford, where the political sensitivities are, or even what procedures might not work in the given setting. Consistent with the present Standards, to me it makes eminently more sense to first reach consensus with the client about the users, intended uses, and desired value of the study. Then—given a convincing case for collecting pertinent, relevant information—the client and I will be prepared to engage in meaningful exchange, planning, and contracting concerning aspects of the projected evaluation's feasibility, propriety, and accuracy. . . .

I speak as one who has made extensive use of the *Standards* in a wide range of evaluation and metaevaluation applications. The *Standard*s has proved to be an incredibly useful guide in my evaluation work.

Daniel L. Stufflebeam
September 10, 2008
Excerpt from letter to the Joint Committee on Standards for Educational Evaluation
(quoted in Patton, 2012, pp. 389-390, with permission from Daniel Stufflebeam)

After extensive debate and review, including "the involvement of more than 400 stakeholders in national and international reviews, field trials, and national hearings" (Joint Committee, 2010), the first revision of the standards in 17 years was published, retaining the categories in their original order with utility first (Yarbrough, Shulha, Hopson, & Caruthers, 2010). What the revision did add was a new category on "Evaluation Accountability" that includes three standards highlighting the importance of metaevaluation.

Metaevaluation

Metaevaluation is evaluation of evaluation.

> Metaevaluation is a professional obligation of evaluators. Achieving and sustaining the status of the profession requires subjecting one's work to evaluation and using the findings to serve clients well and over time to strengthen services. (Stufflebeam & Shinkfield, 2007, p. 649)

Stufflebeam was prescient in anticipating the importance that metaevaluation would have as the field of evaluation matured and used the standards as the basis for metaevaluation (Stufflebeam, 2000c). He demonstrated the power of

metaevaluation for the profession by undertaking a comprehensive, exhaustive, and independent review of how 22 different evaluation approaches stacked up against the standards. He developed and applied a 100-point rating scale for measuring adherence to the standards. No one was better positioned by knowledge, experience, prestige within the profession, and commitment to the standards to undertake such a challenging endeavor. He concluded, "Of the variety of evaluation approaches that emerged during the twentieth century, nine can be identified as strongest and most promising for continued use and development." When he published his findings, I was understandably pleased to find that utilization-focused evaluation was among those nine, with the highest rating for adherence to the utility standards (Stufflebeam, 2000a, p. 80). But I was even more impressed with the rigorous process he developed and applied for using the standards as a metaevaluation framework to compare and contrast evaluation approaches.

I conducted a metaevaluation of the evaluation of implementation of the Paris Declaration on development aid. The Paris Declaration, endorsed on March 2, 2005, committed more than 100 government leaders, heads of agencies, and other senior officials to increase efforts to harmonize and align aid initiatives, and manage aid for results with a set of monitorable actions and indicators. The evaluation of the implementation of the Paris Declaration was completed and the report submitted in mid-2011. That report was aimed at the High Level Forum on Aid Effectiveness that took place in Busan, South Korea, in December 2011. The metaevaluation I conducted (Patton, 2013) was used in conjunction with the evaluation report. That metaevaluation assessed adherence to Evaluation Standards. The evaluation, supported by the metaevaluation, was awarded AEA recognition as the Outstanding Evaluation of 2012. Following the award, I had an opportunity to discuss with Dan the importance of metaevaluation and the impossibility of conducting metaevaluations without high quality and clear standards. It was the only time I got to express my personal appreciation for his work in leading development of the standards. He was, as was his style, modest about his role. I knew otherwise. Without Dan Stufflebeam's leadership, the evaluation standards would not have been produced.

Stufflebeam was ultimately disappointed that more use of the standards for metaevaluation did not occur. He commented:

> I think the evaluation field should make regular use of standards and meta-evaluation as means of ensuring that evaluations are useful, ethical, practical, and accurate. Moreover, I think that AEA and other segments of the evaluation profession should place a high priority on helping the Joint Committee obtain sufficient funds to fulfill its mission. A regular source of funding is needed to keep the joint committee standards up to date and focused on the evaluation field's needs for improving the quality and impacts of evaluations. (Stufflebeam & Miller, 2008, p. 570)

The Continuing Story

People matter. The *personal factor* makes all the difference – the commitment, dedication, and follow-through of those who have a vision, care about quality, and lead the way from what was not to what is, and in so doing, shape the future. Our future. Our shared professional future. A hallmark of any profession is that it has professional standards and ways of applying shared standards to enhance the quality of the work carried out by that profession's practitioners. Stufflebeam's leadership in developing the joint committee standards elevated evaluation to a credible profession.

The profession subsequently adopted Guiding Principles, a statement on Culturally Responsive Evaluation, and is about to endorse Essential Competencies. These are important developments in the evolution and maturing of evaluation as a profession. But it all began with the Evaluation Standards: standards of excellence from a leader of excellence. Remember.

References

Bernstein, I., & Freeman, H. E. (1975). *Academic and entrepreneurial research: Consequences of diversity in federal evaluation studies*. New York: Russell Sage.

Coryn, C. L. S. (2017). In memoriam: Daniel L. Stufflebeam (1936-2017). *American Journal of Evaluation*, forthcoming.

Joint Committee on Standards for Educational Evaluation. (1994). The program evaluation standards. Thousand Oaks, CA: Sage.

King, J., Stevahn, L., Ghere, G., & Minnema, J. (2001). Toward a taxonomy of essential program evaluator competencies. *American Journal of Evaluation*, *22*(2), 229–247.

Patton, M. Q. (1994). The program evaluation standards reviewed. *Evaluation Practice, 15*(2), 193–99.

Patton, M. Q. (2012). *Essentials of utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Patton, M. Q. (2013). Meta-evaluation: Evaluating the evaluation of the Paris Declaration. *Canadian Journal of Program Evaluation*, *27*(3), 147–171.

Patton, M. Q. (2018). *Facilitating evaluation: Principles in practice*. Los Angeles: Sage.

Russon, C., & Russon, G. (Eds.). (2004). International perspectives on evaluation standards. *New Directions for Evaluation*, *104*.

Stufflebeam, D. L. (1980). An interview with Daniel L. Stufflebeam. *Educational Evaluation and Policy Analysis, 2*(4), 90–92.

Stufflebeam, D. L. (2000). Foundational models for 21st century program evaluation. In Stufflebeam, D.L.,Madaus, G.F., & Kellaghan, T. (eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.) (pp. 33-83). Newell, MA: Kluwer.

Stufflebeam, D. L. (2000). Professional standards and principles for evaluations. In Stufflebeam, D.L., Madaus, G.F., & Kellaghan, T. (eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.) (pp. 439-455). Newell, MA: Kluwer.

Stufflebeam, D. L. (2000). The methodology of the metaevaluation. In Stufflebeam, D.L., Madaus, G.F., & Kellaghan, T. (eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (2nd ed.) (pp. 457-471). Newell, MA: Kluwer.

Stufflebeam, D. L (2004). A note on the purposes, development, and applicability of the joint committee evaluation standards. *American Journal of Evaluation 25*(1), 99–102.

Stufflebeam, D. L., & Shinkfield, A. (2007). *Evaluation theory, models, and applications*. San Francisco, CA: Jossey-Bass.

The Oral History Project Team (Miller, R. L., Coryn, C. L. S., & Schröter, D. C.). (2008). The oral history of evaluation: The professional development of Daniel L. Stufflebeam. *American Journal of Evaluation*, *29*(4), 555-571.

Yarbrough, D. B., Shulha, L. M., Hopson, R. K., & Caruthers, F.A. (2010). *The program evaluation standards: A guide for evaluators and evaluation users (3rd ed.)*. Thousand Oaks, CA: Sage.