Journal of MultiDisciplinary Evaluation
Volume 21, Issue 50, 2025

JMDE
Journal of MultiDisciplinary Evaluation

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

Evaluation, INTEVAL, and Two National Audit Offices (U.S. GAO & Netherlands Court of Audit): Ray Rist's Contributions and Leadership in the Early Years of INTEVAL

Frans L. Leeuw
Maastricht University

Background: Background: This paper describes Ray Rist's intellectual, research-focused work in the early years of INTEVAL (largely from the mid-1980s to around mid-1990s). It sets out the position and roles of the (then) U.S. General Accounting Office (GAO) and the Netherlands Court of Audit (NCA) with respect to evaluation.

Purpose: Differences and similarities between the GAO and NCA experience are described, with the GAO as a "first-wave" organization in the evaluation profession and the NCA as a "second-wave" organization, as it learned from the United States' experiences. The labeling of first- and second-wave developments in evaluation was formulated by Derlien (1989). There is a particular focus on the role of leadership in shaping practice.

Setting: Setting: During the period from the mid-1980s to mid-1990s, Rist and the author of this article were working at, respectively, GAO and NCA in the field of policy evaluation / program evaluation and methodology. The setting was such that they worked together on a number of occasions and products, amongst others dealing with how (performance) audits and evaluation developed; the introduction (in the NCA) of government-wide (comparative) evaluations of tools of government such as subsidies, information campaigns, and inspections (covering all ministries); how evaluation was related to the learning capability of governments; and how INTEVAL as a group of evaluators, auditors, and other social scientists developed.

Intervention: Not applicable.

Research Design: Not applicable.

Data Collection and Analysis: Not applicable.

Findings: A number of lessons are set out about the role of leadership in evaluation.

Keywords: *Ray C. Rist; GAO; NCA, INTEVAL; evaluation.*

Working at Leiden University (Netherlands) in the mid-1980s as an associate professor in the Institute for Social Policy Research, I planned a trip to the United States to visit several organizations that did program and policy evaluations. During my Fulbright scholarship in 1979–80 at the University of Chapel Hill (United States) I started to read papers and books on evaluation of policies and programs in the United States. Middle-aged, white men like Duncan MacRae, William Dunn, Mark van de Vall, and (a theoretical sociologist) William E. Snizek were among the ones I met. In those years, in some of the Dutch social sciences faculties, the abbreviation "IAG" was well known: "In Amerika geweest…," meaning "Having been in America…"). That was seen as an important component of one's (future) career and helped, a bit later in life, to go back to the United States on research trips (of course, only if you had enough peer-reviewed publications and invitations). In 1987 I made such a trip. U.S. GAO, and in particular its Program Evaluation and Methodology Division (PEMD) was on the agenda, along with several other institutes. Marie-Louise Bemelmans-Videc (who sadly passed away a few years ago), also working at Leiden University in the law faculty, was a member of the then very young INTEVAL group and suggested I make an appointment with Dr. Ray C. Rist, who was the chair of INTEVAL and the director of operations of PEMD. The meeting took place and led to a long-lasting collaboration and friendship.

In this article I will describe Ray's intellectual, research-focused work in the very early years of INTEVAL (largely between the mid-1980s and around 2000). Most of those years, we were both working at national audit offices in the field of policy evaluation / program evaluation and methodology—Ray at GAO as the director of operations of PEMD, and I as head of the division for doelmatigheidsonderzoek (performance audits, comparative audits, and evaluations) of the Netherlands Court of Audit.

Institutional Contexts: US GAO and the Netherlands Court of Audit in the 1980s and 1990s

My focus here is on the period from the 1980s until the early 2000s, which means that our work institutions (GAO and the Netherlands Court of Audit) *in that period* are playing an important role. So let me start with a bit of institutional context on the two audit offices.

The U.S. General Accounting Office (GAO, known since 2004 by the new name "Government Accountability Office") has around 3,200 staff and is much bigger than the Netherlands Algemene Rekenkamer (Netherlands Court of Audit; NCA), which has a staff around 250. According to the GAO website:

> The U.S. Budget and Accounting Act created the GAO in 1921 when Congress realized the need to control growing government expenditures and debt after World War I. Until the end of World War II, GAO primarily checked the legality and adequacy of government expenditures. After World War II, as government responsibilities and programs grew, so did GAO. The focus of our work shifted toward helping Congress monitor executive branch agencies' programs and spending. (GAO, n.d., para. 4–5).

After the early '70s, evaluation became an important part of the work agenda of GAO, hiring scientists from a great diversity, investing in methodology, doing experiments to find out about effectiveness, etc. Chelimsky wrote:

> We began building this capability [for evaluations] informally through the 1970s in our Program Analysis Division, and then took the formal step, in 1980, of creating the Institute for Program Evaluation. This unit became, in 1983, GAO's Program Evaluation and Methodology Division (PEMD). (1990, p. 43)

When I met Ray for the first time, he was the director of operations at PEMD, and before working at GAO he was a professor at Cornell University.

The Algemene Rekenkamer, the Netherlands Court of Audit (NCA), has a history of over 600 years, but the name it uses now (and the location of the organization) was established in 1814. Since 1927 the Court, based on the Governments Accounts Law, had the duty to verify as much as possible "whether the government assets that are susceptible are sufficiently productive." This was not only a very restricted definition of efficiency, but more problematic was that the Court could only "verify" what traditional audits had found. There was no independent task regarding efficiency, let alone effectiveness. Since the amended 1976 Governments Accounts Law, the office "could [also]

pay attention" [1] to doelmatigheids- en doeltreffendheidsonderzoek' (performance auditing; program and policy evaluation and efficiency studies). In the early 1980s a new unit was established and tasked with doing performance- and evaluation-type work, and in the early 1990s it became one of three directorates of the NCA. Their tasks were (1) doing government-wide audits and evaluations [2] (looking into the efficiency, effectiveness and side effects of all major tools of government (sticks, carrots, sermons, pillories, contracts, inspections, certifications, privatization, centralizations / decentralization, quangocratization and "evaluations" as a tool of government); (2) case studies; (3) data collection and analysis, including—though seldom—articulating and testing underlying assumptions (today called theories of change). The directorate was also in charge of (4) compliance audits and performance audits of three ministries.

One of the first things I discussed with the then-president of the Court (Frans Kordes) in early 1988 was his memorandum "Quo Vadis" (Kordes, 1984). It outlined his vision and perspective of where the NCA should go to in the next 5 to 10 years, and why and how. The memorandum included a plea for more empirical work from an evaluation, performance auditing, and methodological perspective. I learned that visiting GAO and his meetings with Chuck Bowsher (then its auditor general) contributed strongly to the foundation for this and to a desire to have the NCA go for the broad spectrum of activities and tasks. It is possible that Kordes also met Chelimsky, but I couldn't check that, given his passing away a number of years ago.

A few years later, in 1994, I asked Kordes to be EES's first president, and he gladly accepted. The first EES congress was held in The Hague on 1 and 2 December, 1994. During that congress he met Eleanor Chelimsky, then (former) assistant comptroller general for program evaluation and methodology of the GAO, who was coincidentally also president of the American Evaluation Association at the time. Ray also met Frans Kordes, and they went together very well. I still see them standing in hallways and congress rooms chatting and discussing GAO, NCA, and the work that had to be done! Kordes and I published a short impression of that first EES congress in the Sage journal *Evaluation* (Leeuw & Kordes, 1995, pp. 122–124).

Rist's Intellectual and Leadership Developments Between the Mid-'80s and -'90s and Their Impact

In 1987 for me the first and maybe most impactful piece of work produced by Ray was his Chapter 16 in *Social Science Research and Government*, edited by U.K. sociologist Martin Bulmer and published in 1987. Why was this chapter of such a great importance and impact?

One reason is that the context of his chapter is the "description and assessment of the social science contributions within the mosaic of GAO activities" (1987, p. 304) in a time when GAO was, earlier than the NCA, broadening its focus and methodologies. It was (then and now) a feast to read Ray's analysis of what was going on in this huge organization. He described a variety of contributions that social science research provides to help meet the information needs of the Congress.

Secondly, Rist distinguished between theory-driven and question-driven analysis. However, most fascinating was his typology of questions asked and answered by GAO research: descriptive questions, normative questions, and cause-and-effect questions. For all of these he presented examples ranging from case studies and quasi-experiments to program operations and delivery of services examination and systematic reviews. Following Pressman and Wildawsky (1979) he had a sharp eye for the fact that "policies imply theories." With respect to normative questions, Rist makes the point that "answering normative questions is not normally thought of as part of the general methodological repertoire" (1987 p. 312). But he found that "working in the GAO has created a number of opportunities to explore the social science contributions to answering normative questions" (p. 312)

In 1989 he published in the *Journal of Public Policy* an article on management accountability (in the public sector), describing and analyzing the signals auditing and evaluation sent. He saw auditing and evaluation as

> two strategies which governments may use to ascertain if existing policies and programs are

[1] This is the translation of what is in the budget and accounts legislation. There were two goals in using these vague words: One was that nobody could hold the NCA responsible for an *annual* and *full coverage of efficiency/effectiveness* of the central government's organizations and policies. The second was that it allowed the NCA to develop its own approach/methodology doing this work.

[2] See Kordes, Leeuw, and van Dam (1991) for the philosophy behind government-wide studies, as well the example on subsidies.

> being administered as they ought. The analysis compares the strengths and weaknesses of auditing and evaluation as means of monitoring accountability. Three types of accountability are discussed. Auditing makes its strongest contributing to managerial accountability through a focus on fiscal or regulatory issues. Evaluation makes its strongest contribution in the area of program effectiveness. Both make contributions to program implementation. The two approaches are viewed as mirror opposites of each other. (Rist, 1989, p. 356)

I see this article as an early and quite innovative comparison of these two "worlds." The ways in which he specified the different contributions is very much to the point (for the 1990s and early 2000s). It prompted me, a few years later, to write about differences and similarities between performance auditing and evaluation (Leeuw, 1992); the differences between performance audits, new public management, and performance improvement (Leeuw, 1996a); and how to build bridges to combine audits and evaluation (Leeuw, 1996b).

The year 1989 brought a special issue of the *Knowledge in Society Journal*, starting with the introduction (nicely called "Forethoughts") by Ray. Several INTEVAL members had contributions: Erik Albaek, who wrote on designs and utilization; Andrew Gray and (the late) Bill Jenkins on evaluation in the United Kingdom, and (the late) Marie-Louise Bemelmans-Videc, who presented and discussed a then-typical Dutch approach to evaluations, the so-called reconsideration procedure.[3]

In 1990 the first book in the Comparative Policy Evaluation series under Ray's editorship and guidance appeared with Transaction Press: *Program Evaluation and the Management of Government: Patterns and Prospects Across Eight Countries*. Ray's chapter introduced an important—and today maybe even more relevant than in 1990—distinction between *the management of evaluations* and *management by evaluations*. This issue is still relevant, despite the large increase in the number of handbooks, textbooks, and cookbooks on the management of evaluations. Even more prominent are changes regarding the second "management," i.e., management *by* evaluations: Compared to 1990, now arguments and papers are published showing that there is not only perhaps too much of that, but also that "management by evaluations" can have and sometimes has unintended negative side effects. Think of the critique by Frey (2006) on "evaluitis," the critical discussion of routinization and evaluation machines by Dahler Larsen (2018), and Raimondo and Leeuw (2020) and Leeuw and Pleger (2023) on evaluation capture. So, in case there will be an anniversary edition of this book (in 2025 or 2030), it should not only focus on *patterns and prospects* but also on *unexpected, unintended, and negative side effects*.[4]

The year 1991 brought a seven-page-long interview with Ray by Paul Johnson, published in *Evaluation Practice* (in 1997 continued as the *American Journal of Evaluation*). This conversation was partly about his responsibilities as chair of the working group on policy and program evaluation, which was formed in 1986 by the International Institute on Administrative Sciences (IIAS), located in Brussels. In the beginning of INTEVAL, IIAS sponsored the academic members of INTEVAL and took care of some of their costs for the meetings. Paul Johnson also raised questions about the first INTEVAL book, one of the questions dealing with the wave theory of evaluations (developed by Derlien [1989]). Ray "sold" or, if you will, "marketized" the (development of) INTEVAL as a group and as a producer of knowledge like a salesman.

Later in the 1990s Ray and I edited and published, together with Dick Sonnichsen (head of evaluation and performance auditing of the FBI) a book called *Can Governments Learn?* (1994). The book examined organizational learning in the public sector and sought to understand the role policy and program evaluation can play in helping governments to learn. Among the societies studied were Belgium, Canada, The Netherlands, Sweden, and the United States. Rist, Bemelmans-Videc, and Vedung edited in 1998 the "great cookbook"

[3] My contribution originally planned for this special issue was not ready in time; it appeared in 1991 in the same journal, but without the "Forethoughts" of Ray.

[4] Dennis J. Cusley published a review of this book in *Evaluation Practice*. He, amongst others, is critical writing things like this: "The book is of some value. But in other respects this reviewer finds it wanting" (1990, p. 243). He also points to the many definitions that exist for evaluation. As far as I can see, Cusley forgot or completely overlooked the contribution of the book to the evaluation literature. Think of the wave theory presented by the late Hans-Ulrich Derlien, the different types of management of and by evaluations, the—indeed—often bumpy roads along which evaluation was developing in several countries; these are only a few points that are not even mentioned in the review.

Carrots, Sticks and Sermons: Policy Instruments and Their Evaluation. Together with a colleague from the NCA, we had a chapter on the evaluation of subsidies ("carrots") based on a government-wide evaluation of the 700 subsidies used by the Netherlands central government. Analyzing the underlying subsidization theory; confronting it with legality, compliance, and administrative data; checking the evaluability of the subsidies and whether or not the government had actually carried out or commissioned effectiveness and efficiency evaluations—this was all part of the NCA report and of the chapter in the "cookbook." The database was later used by me and other colleagues in academic journal articles.

This article ends around 2000. My collaboration with Ray never ended. When he left GAO and moved to the World Bank (in 1997) and I left NCA (1996) to take up the position of professor and dean of the humanities faculty of a Dutch university[5] as well as the chair of evaluation studies at Utrecht University (Netherlands), we continued our joint activities. Although incomplete, they concerned Ray's invitation to become one of the first lecturers of IPDET (I did that from 2000 till 2012); the World Bank invited me to evaluate the anticorruption program in two African countries (Ray joined partly) (Leeuw, van Gils & Kreft,1999); I edited, with Jos Vaessen (now Inter-American Development Bank), the INTEVAL book *Mind the Gap* (2010) and authored a number of chapters in other books in the series; Ray invited me to write two reports (one on the organization of evaluations [*Think Before You Leap*] and one on the World Bank's "readiness assessment approach," asking whether the underlying logic of this approach was valid). Many other joint activities followed.

Conclusions and What Can Be Learned from This *N* = 1 Case

I hesitate to present some learning insights from this single case, as it is a single case between two friends, but nevertheless I thought about these:

- When entering new fields, new organizations/institutions, and trying to implement new or improve existing research approaches (or paradigms or strategies), it is not enough to have vision statements or future (quo vadis) memoranda or whatever type of document. They are important but not enough.
- More important are people who take the lead, are persistent in this, do not shy away from headwinds, try to embed their vision in the behavior of others but always with dialogue and discussion. At the same time, it is important that such people, sometimes called champions, have social capital, i.e., have soulmates who talk critically and constructively with and against them and who create a nice atmosphere too. In the 70s, 80s, 90s of the last century, this mainly meant social-physical contact. Nowadays a part of that can be done digitally.
- Dialogues and discussions need to be open, i.e., not limited by protocols or (woke) ideologies. Participants need not to be afraid or against discussion where people are contradicting each other and are in favor of two-(or more)-sides-ism. Nobody has a claim to truth.[6] Sir Karl Popper's falsification perspective, and his critique on "the poverty of historicism," are nowadays more relevant than maybe ever before. At the same time, it is crucial that champions create opportunities for people / staff / students / PhDs to participate, learn, fight, love, and conquer.

Acknowledgments

Thanks to Richard Boyle and Rob van den Berg for comments on a draft version of this paper.

References

Bemelmans-Videc, M.-L., Rist, Ray C., & Vedung, E. (Eds.). (1998). *Carrots, sticks and sermons: Policy instruments and their evaluation* (Comparative Policy Evaluation Vol. 7). Transaction Publishers.

Chelimsky, E. (1990). Expanding GAO's capabilities in program evaluation. *The GAO Journal*, *8*, 43–53.

Cusley. D. (1990). [Review of the book *Program Evaluation and the Management of Government* by R. C. Rist (Ed.)]. *Evaluation Practice*, *11*(3), 242–243.

[5] While working at the NCA I was appointed a part-time professor in evaluation and policy studies at Utrecht University, Department of Sociology (1993–2005).

[6] I know that a discussion is possible on how "far" Popper's fallibilism and the falsifiability thesis reach. One perspective is that fallibilism is "everywhere" and "always" needed. Another position is more restricted, in the sense that one can never be certain of the truth of a statement of what is the case but one can be certain of the truth of a statement of what is not the case.

Dahler-Larsen, P. (2018). The skeptical turn in evaluation. In J. E. Furubo & N. Stame (Eds.), *The evaluation enterprise* (pp. 58–80). Routledge.

Derlien, H. U. (1989). Genesis and structure of evaluation efforts in comparative perspective. In R. C. Rist (Ed.), *Program evaluation and the management of government* (Chapter 9). Routledge.

Frey, B. (2006). Evaluitis – A new illness (Evaluitis – Eine Neue Krankheit). CREMA; University of Basel.

GAO. (n.d.). 100 years of GAO. https://www.gao.gov/about/what-gao-does/hundred-years-of-gao#Our%20History%20at%20A%20Glance

Johnson, P. (1991). Ray Rist talks about IIAS working group program evaluation [Interview]. *Evaluation Practice*, *12*(1), 45–53.

Kordes, F. G. (1987). De gedaanteverandering van de Algemene Rekenkamer. Voordracht Nederlands Genootschap van Hoofdredacteuren.

Kordes, F. G., Leeuw, F., & Van Dam, J. (1991). The management of government subsidies. In A. Friedberg et al. (Eds.). *State audit and accountability* (pp. 280–299). Jerusalem.

Leeuw, F. L. (1991). Policy theories, knowledge utilization, and evaluation. *Knowledge and Policy*, *4*, 73–91. https://doi.org/10.1007/BF02693089

Leeuw, F. L. (1992). Performance auditing and policy evaluation: Discussing similarities and dissimilarities. *Canadian Journal of Program Evaluation, 7*, 53–68.

Leeuw, F. L. (1996a). Performance auditing, New Public Management, and performance improvement: Questions and challenges. *Accounting, Auditing & Accountability Journal*, *9*(2), 92–110.

Leeuw, F. L. (1996b). Auditing and evaluation: Bridging a gap, worlds to meet? *New Directions for Evaluation*, *71*, 51–60.

Leeuw, F. L., & Kordes, F. G. (1995). Some impressions of the first European Evaluation Society conference. *Evaluation*, *1*(1), 122–124.

Leeuw, F. L., Van Gils, G., & Kreft, C. (1999). Evaluating anti-corruption initiatives: Underlying logic and mid-term impact of a World Bank program. *Evaluation*, *5*, 194–219.

Leeuw, F. L., Rist, R. C., & Sonnichsen, R. C. (1994). *Can governments learn? Comparative perspectives on evaluation and organizational learning* (Comparative Policy Evaluation Vol. 3). Transaction Publishers.

Leeuw, F. L., & Pleger, L.(2023). Evaluation capture, evaluator resilience, and the need for competencies of evaluators. *Journal of MultiDisciplinary Evaluation*, *19*, 46–53. https://doi.org/10.56645/jmde.v19i46.863.

Pressman, J., & Wildavsky, A. (1973). *Implementation*. University of California Press.

Raimondo, E., & Leeuw, F. L. (2020). Evaluation systems and bureaucratic capture: Locked in the system and potential avenues for change. In B. Perrin & T. Tyrrell (Eds.), *Bureaucracy and evaluation* (Chapter 9). Taylor & Francis Group.

Rist, R. C. (1987). Social science analysis and congressional uses: The case of the USA General Accounting Office. In M. Bulmer (Ed.), *Social science and social policy* (Chapter 16). Cambridge University Press.

Rist, R. C. (Ed.). (1989–90). Cross national perspectives on the policy uses (and abuses) of evaluation [Special issue]. *Knowledge in Society*, *2*.

Rist, R. C. (Ed.). (1990a). *Program evaluation and the management of government: Patterns and prospects across eight nations* (Comparative Policy Evaluation Vol. 1). Transaction Publishers.

Rist, R. C. (1989). Management accountability: The signals sent by auditing and evaluation. *Journal of Public Policy*, *9*(3), 355–369.