Education Review
Reseñas Educativas
Resenhas Educativas

CREATING NEW IDEAS IN EVALUATION

ERNEST R. HOUSE

In Spring, 1967, I was completing a doctorate in education at the University of Illinois when I received a phone call. Would I be interested in evaluating the Illinois State Program for Educating the Gifted and report the findings to the state legislature? I would have four years and a budget of $5m (2025 dollars). Uncertain, I walked over to the office of my statistics professor. “Sure, you should take it on,” he said. “But I don’t know anything about evaluating programs,” I said. “Nobody does,” he said. “Only a few papers have been written on the topic, including a recent one by Bob Stake. He’s next door. Go talk to him.”

I did take on the project, a risky choice compared to university jobs. I was 30. Four years later, I completed the evaluation successfully, one of the first large program evaluations in the country; I had learned evaluation by doing it. Over the next several decades, I continued to conduct and write about evaluations, helping develop evaluation as a distinct practice, profession, and discipline. At his retirement seminar in 2011, Michael Scriven, philosopher of science and first to conceive of evaluation as a discipline, said I had generated some of the most creative ideas in the field. Of course, others had also contributed significantly in the early years (see Williams, 2016).

How did I manage this? In *The Act of Creation* (1964), Arthur Koestler, the Russian novelist, explains that being creative in science, arts, and humor entails taking an idea from one framework and applying it to another frame, creating an unexpected insight. He cites many examples, such as Archimedes taking a bath, realizing how his body displaced the water, and applying that idea to measuring the volume of solids. Of course, the new idea must be relevant to a problem in focus.

In conducting evaluations, I often encountered problems. If the problem was significant enough, I sometimes looked to frames of reference outside evaluation. Many ideas I weighed against my childhood and teaching experiences. Occasionally, this led to new insights. Merging an idea from outside with an idea inside evaluation created a third idea. Sometimes the new

House, E. R. (2025, December 11). Creating new ideas in evaluation. [Acquired Wisdom Series]. *Education Review, 32*. https://doi.org/10.14507/er.v32.4389

ideas worked and were widely accepted. Sometimes they were ignored.

Of course, you only work with the intellectual frameworks you know. My frameworks were many. First was my childhood, happy at first, then troubled, difficult, rough. By the time I was 9, I had come to believe that adults did not make good decisions and that I had to think for myself. In other words, I became a critical, independent thinker at an early age. Sometimes I could be too critical.

My second framework was a superb liberal arts education at Washington University, St Louis. This opened new intellectual worlds for me. Physics, chemistry, biology, genetics, anthropology, ancient Greek history, social history, philosophy, psychology, French, and English literature, my major. For four years I studied intensely. These diverse courses provided pathways to future ideas. I continued to read widely in several areas.

After college, I taught high school English for 4 years. I learned what teaching was like. I was asked by University of Illinois staff to introduce new materials they were developing at their lab school, including rhetoric, semantics, and linguistics. After 2 years they offered me a job at the university to help develop materials, train teachers, and travel the state to see how these teachers were faring. Two more years and I had practical experience in schools to recognize what was possible and worthwhile.

My fourth framework was graduate work at the University of Illinois. James J. Gallagher, the leading authority on education of the gifted, offered me a scholarship to study for a doctorate. I took courses in educational history, philosophy, psychology, testing, statistics, and organizational theory. In general, I learned how educational research was conducted.

Three frameworks are easy to grasp; my childhood warrants a closer look. Early experiences influenced ideas I had years later.

Childhood

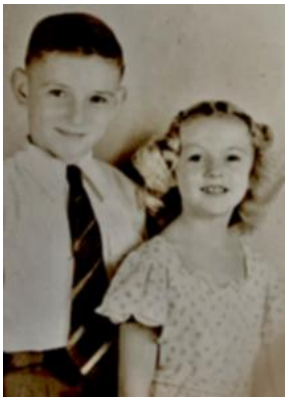
I was born in a boarding house, son of a taxi driver. The town was Alton, Illinois, an old riverboat town on the Mississippi, now industrialized. It was 1937, last stage of the Great Depression. We moved to a small stucco house on Cherry Street Alley that my grandmother rented for $5 a month. The family included my father, mother, grandmother and me, plus a few uncles and aunts in and out. The hustle, bustle, and warmth of that family remains among my happiest memories.

In winter of 1939, we lived on rabbits and squirrels the men hunted, and sweet potatoes they dug up on farms. After a few years my father, mother, and I moved to a house on Main Street, along with my baby sister Georgia. I accompanied my dad on his taxi runs, an exciting adventure. I had a toy steering wheel I turned as I sat beside him in the front seat. The two of us seemed inseparable. When I was 4, he was killed in a car crash during his midnight shift, hit by a

drunk driver. Twenty years later, I learned that my dad was an ex-con who had spent two and one-half years in Leavenworth Penitentiary.

After his death, we had no money. There was no support for widows or children. My mother, grandmother, sister, and I moved to a three-room rental. My grandmother took care of us while my mother worked three shifts in the war-time munitions factory. We lived on little. Nonetheless, those years were happy enough as I started school, though I never forgot the memory of my father nor his loss.

When I was 7, my mother married a man from the factory, and we moved to an isolated house on a lonely highway north of town. Unfortunately, the man was mentally ill. He and my mother had violent arguments, and he often held a loaded gun against my head. After 6 months, we ran away, and my mother left my sister and me at her sister's farm across the river in Missouri.



While my mother worked at the factory, my sister and I lived on the tenant farm with an aunt, uncle, and three cousins. No running water, no electricity, no car, no telephone. It was clear to us we were not wanted, a burden on the farm family. After 6 months, my mother arranged for us to live with another uncle and aunt among the oil refineries on the Illinois side of the river. After 6 months, we began feeling we belonged.

Abruptly, my mother and aunt had a falling out about our care. We were sent back to the farm in the dead of winter. This was a low point in our lives. I was 9 and my sister 7. We felt nobody wanted us. We were a burden to everyone. Those were dark days. I've told this story as experienced from a child's point of view in a memoir, *Cherry Street Alley (*House, 2015a*)*.

That memoir ends with my mother's third marriage, our return to the city, and a stable life. My new stepfather, although a rifleman in the fiercest battles of World War II, was a mild, gentle person. He helped provide the stability we needed to survive and thrive. Even so, the stress of those years changed us. By age9, I had decided adults made bad decisions that affected their children adversely. I had to think for myself to protect me and my sister. I became a critical independent thinker at an early age.

My First Evaluation

To begin my first evaluation, I read everything written about program evaluation in a month. There wasn't much. I appointed an advisory committee chaired by Bob Stake, which included Egon Guba and Dan Stufflebeam, leaders in the new field. No person was ever better served by such a committee. They were extremely generous with their time and ideas. Our meetings were exciting exchanges of ideas. It was a new world.

The conventional research was to administer standardized tests as outcome measures, either in surveys or designs with control groups. But the Illinois Gifted Program was complex, with 25 demonstration centers, 340 school districts receiving money for their local programs, plus teacher training workshops and research studies.

I had a sense that if I were to evaluate the program, I should be fair to what the program actually was, not reduce it to test scores that didn't capture its essence. Stake (1967) had written a paper calling for broad descriptions of programs. I based the

evaluation on his "countenance" model of what an evaluation might encompass.

At that time, education and social researchers were focused on quantitative methods as the path to truth, and I spent much of my budget training interviewers, recording interviews, and collecting other data that met those reliability demands. Joe Steele had developed an environmental press questionnaire based on the intellectual levels of Blooms Taxonomy. We used that innovative device to collect classroom data. Steve Lapan led the data collection team. Altogether we collected 42 different types of data from a stratified random sample of 34 school districts that we used to typify gifted classes. Tom Kerins helped with the analysis.

All this we reported to the state legislature in a series of reports throughout 4 years. To evaluate the schools demonstrating new teaching techniques, we not only questioned visitors to the schools but tracked them into their schools to see if they had changed. Most had not; not the finding anyone expected. This led to rethinking how demonstration schools should work. They must work with teachers within their particular settings.

Once we began dealing with schools, I quickly realized how intensely political evaluation was. We were making judgments about activities that affected people's careers. This was inherently political. We encountered many conflicts with local and state officials that had to be resolved through negotiation. I began seeing politics as inherent in the evaluation process. Conflict and negotiation were inevitable.

The Politics of Educational Innovation

I collected and published these empirical findings in a book (House, 1974a). If you introduce new curriculum materials to a school, those most likely to be interested are the younger teachers. The older ones are more set in their ways and perhaps skeptical that things can or should change. Yet the younger teachers have little power and are vulnerable to criticism. Hence, it's critical to have the school principal strongly supporting the new innovation. Otherwise, things do not go well. In short, you must understand the politics of the faculty.

The chances of success are greatly enhanced if the teachers undergo training in summer workshops and work in teams so they can support each other. Innovators must exert considerable effort to implement new ideas. There must be a strong inside advocate to make things succeed. Otherwise, the innovation fades away.

In later work, I expanded these ideas by positing three critical dimensions to achieving innovation: the technical, political, and cultural (House 1981; House & McQuillan, 1998). For example, if you are introducing the "new math," the teachers must understand the math itself. I have seen math presented by teachers who simply pass out worksheets and cannot explain the problems. This doesn't work.

There is also the culture of the school. When I was teaching high school English, I taught the lowest ability group as well as the top. Having read Chomsky's *Syntactic Structures* (1957), I told this class that we, as a class, would construct our own grammar. Chomsky claimed that those who speak a language understand the grammar implicitly, even if they can't parse it as teachers do. For once the class was interested; they were learning something others weren't. This lasted only a few days until the aging head of department came charging into my room. How dare I teach this esoteric stuff to these boys who didn't even know the real grammar? The boys

were not supposed to know something those higher up didn't. I had run afoul of the culture of the school about who was supposed to know what.

An important background story is that of the federal research, development, and diffusion plans for innovation. When Dave Clark was in the U.S. Office of Education, he envisioned that schools might innovate by establishing research centers that would create new ideas and educational laboratories that would help develop and infuse new programs into the schools. The plan, as drawn up by Clark and Egon Guba, was known as the Research, Development, Diffusion model of educational change.

As I evaluated the Illinois Gifted Program, I could see that the Illinois Demonstration Centers did not operate in the way envisioned. Teachers visited the demonstration centers, loved what they saw, but did not implement the ideas. They would say, "My principal would never let me do that." Adopting new ideas was far more involved. I used the Illinois data to demonstrate that the R, D, and D model was too rationalistic. In fact, adoption of an innovation was far more political and culturally laden.

I published this critique, and we argued in print (House et al., 1972). By then, I knew Clark and Guba well and regarded both of them highly. Over time, they agreed with us. When people are presented data contradicting what they believe, it takes a period of time for them to absorb the change. They must move their intellectual furniture around in their heads. The centers and regional labs could be useful once they recognized the complexities involved in educational change.

However, this new realization was already muted by national politics. In 1968, I had gone to Washington to show James Gallagher my evaluation plan and asked Gallagher, now number two in the Office of Education, where all the promised federal money for the labs and centers was. He said, "On boats going to Vietnam."

Lyndon Johnson's grand plan for education had been wrecked by the ambitions of the American foreign policy establishment, which argued that if Vietnam fell, all Asia would fall like dominoes to the communists. Vietnam is now a major trading partner, after millions of lives lost and billions of dollars wasted. The R, D, and D model was too idealistic, but it was better than the Domino theory. This episode was emblematic of education policies over the next many decades.

My book on the politics of educational innovation was well received. It countered simplistic ideas then prevalent. Innovators cannot simply introduce ideas and expect educators to adopt them. Education change is more complex. I should have revised the book but didn't. This reveals one of my regrettable traits. I was always moving on to new ideas, rather than revising old ones. Throughout my career I pursued what intrigued me intellectually at the time.

Disputes with Governments over Education Policies

If the government was not investing large sums of money in education reform, what was their approach? Nixon was now president. The new education policies arrived in the form of the Michigan Accountability System. It was based on student testing. There were tests at different grade levels; and if schools scored too low, they would be deprived of some funding the next year. There were few resources to help students or teachers. This plan had been developed with the approval of the feds and rumored to become the new federal policy.

When I was teaching high school, I had been active in the teachers' union. I had an idea how to challenge such policies—evaluate them. I conveyed the idea to union colleagues. Soon, the National Education Association contacted me. My idea was to set up a panel of experts to take a look at the untried Michigan system, like a royal commission in the UK. Our panel consisted of Dan Stufflebeam, Wendell Rivers, and myself as chair. The NEA supported our efforts, and we had a written agreement they were not to interfere with the panel's work.

Our panel reviewed the details of the Michigan system, conducted interviews, and collected testimony from dozens of people. After months of work, we wrote a report pointing out problems and deficiencies, derived from our analyses and the testimony of others (House et al., 1974). Our report received national publicity.

Since the panel was supported by major teacher unions, the Michigan authorities paid attention, albeit reluctantly. Some egregious features of the Michigan system, such as taking funding away from struggling schools, were modified. Our report circulated nationally and prompted states with similar accountability systems to modify their plans. Florida had such a system, and the NEA set up a panel there chaired by Ralph Tyler, following our review procedures.

Scriven called such evaluation of evaluations, "meta-evaluation." After all, evaluations should themselves be subject to review. Meta-evaluation became a core strategy. I used this to evaluate efforts, such as the Follow Through program, the evaluation of Jesse Jackson's Push/Excel program (House, 1988), the New York City Promotional Gates program, and to refute Charles Murray's findings in *Losing Ground* (1984), which claimed that Great Society programs were harmful (House & Madura, 1988). These meta-evaluations received attention in the mass media.

Unfortunately, that did not stop the spread of poorly conceived, top-down efforts and questionable studies. Those were initiated in the political and financial sectors of society, aimed at restraining education funding, often through mandating tests. The authorities would mandate standardized tests, as if the tests alone would cause good things to happen. This would be followed by reporting test scores rising or falling. Then, experts and politicians would speculate why they thought the scores behaved that way, without any evidence other than what they dreamed up. This circus of publicity has led to little improvement in education. The semblance of reform occurs, not real reform.

Earlier, in education history, after Sputnik, the National Science Foundation took on the task of improving science education by retraining science and math teachers in summer workshops. This was an effective investment in human capital, and it worked. But it was expensive. To reform American schools similarly would require large investments, more than governments were willing to spend. Currently, with the promotion of charter schools, voucher programs, and privatization, apparently governments have decided to circumvent traditional public schools altogether.

CIRCE and the Qualitative/Quantitative Dispute

Looking back a few years to 1969, the reduction in federal education funds meant that the regional lab in Chicago where my gifted program evaluation project was housed lost its federal funding. Our evaluation funding was from the state, but we had to move. Bob Stake and Tom Hastings generously offered us a home to

complete the evaluation at the Center for Instructional Research and Curriculum Evaluation (CIRCE) at the University of Illinois Urbana – Champaign

CIRCE was originally a standardized testing unit. Hastings, the director, studied with Ralph Tyler at the University of Chicago in an atmosphere influenced by John Dewey, and later both he and fellow student Lee Cronbach came to Urbana. Eventually, Hastings started the center and Cronbach left for Stanford. Stake was hired later as a psychometrics expert, but the focus of the center was turning to evaluation.

Tom Hastings
Ernie House
Mary Anne Bunda
Barak Rosenshine
Bob Stake
Terry Denny

CIRCE then became a hotbed of ideas with people like Hastings, Stake, Terry Denny, Gordon Hoke, Jim Wardrop, top graduate students, and several affiliated faculty, plus my group. The Friday sack lunch was the intellectual wellspring. Every Friday we invited someone to present an idea, and we discussed the idea critically, sometimes heatedly. Our seminar room was crowded. Many speakers were from other universities and countries. I tested my ideas in these exciting intellectual exchanges over the next dozen years, as did Stake and others. CIRCE became well known.

Bob Stake **David Berliner**

In 1972, Stake came back from a sabbatical convinced that the case study method developed by Barry MacDonald and others at the Center for Applied Research in Education (CARE) at the University of East Anglia, England, was worth developing as an evaluation approach. Several of us followed his lead. At that moment I was writing cases of the 34 Illinois school districts, trying to fit 42 types of data into an intelligible account; not an easy task.

The work on the Illinois Gifted evaluation had been done with regard to reliability criteria predicated on the notion that the findings might be replicated. Usually, reliability was based on correlations of one kind or another. We published some findings in journals that allowed only reports based quantitative data. It did occur to me that without the reliability effort, we could have found the same results in one year at a fraction of the cost. However, interview material would not have been considered reliable in the then dominant research framework.

The case study approach that the British developed was more journalistic, based on recorded interviews shown to those

interviewed for verification, and presented to the audience as a case study. We followed the looser format after Stake, who would eventually write the classic, *The Art of Case Study Research* (Stake, 1995). This move put us at odds with a large segment of the evaluation community, who used quantitative techniques exclusively and saw case studies as too subjective. Many believed only quantitative methods were valid.

Thus began what some called the "qual-quant" wars, disputes over whether qualitative methods were legitimate. At times the arguments were heated. Stake and I were active in these debates. The quants said the qual methods were too subjective and subject to biases. The quals argued, what's the use of being objective if you think an education program is reducible to a few numbers?

A few like me said that the quantitative studies were not as objective as claimed, but that they hid a number of value decisions made during the study, unrecognized by the evaluators themselves. We attacked the quantitative studies and the philosophy on which the methods were based. This dispute continued through the 1970s and 1980s until about 1994, when a compromise was reached within the newly formed American Evaluation Association (AEA).

Justice in Evaluation

While all this was happening, I wondered how students not in the Illinois gifted programs had fared. How had those not selected been affected? I hadn't given any thought to that in my evaluation. About that time I saw a review of John Rawls's *A Theory of Justice* (1971), considered by many scholars as the best work on moral philosophy of the 20th century. It was remarkably closely reasoned. I read it twice.

Rawls work was in the mainstream utilitarian philosophy tradition, which argued for the greatest good for the greatest number. However, Rawls wanted to modify the tradition because it often sacrificed the welfare of those less well off for the majority welfare. For example, when inflation was too strong, the government increased interest rates, thereby slowing the economy and forcing those in marginal jobs out of work, in effect, trading off the welfare of an unfortunate few for the welfare of many. Rawls thought this was morally wrong. I saw similarities in education.

Drawing on social contract theory, Rawls asked what type of government citizens would choose if they did not know in advance what position they would occupy in that society. He ended with a description of an idealized society that would offer opportunities, like the US, but would also provide strong economic support for those at the lower reaches of the economy, like Sweden. His vision was more egalitarian than the US in that inequalities were allowed only if they helped the least advantaged. For example, you could spend resources training MDs because they would benefit the least advantaged. Or, money spent training obstetricians is more morally responsible perhaps, than money spent training plastic surgeons.

In 1975, I wrote a paper asserting that evaluators should attend to social justice as a criterion in evaluations and have concern for those less advantaged, without saying exactly how this should be done, other than by following Rawls's two principles of justice (House, 1976). The paper, "Justice in Evaluation," caused a furor among many in evaluation and social science, who said that social justice had nothing to do with science. Science was about facts, not values. Lloyd Humphreys, a big name in the Illinois psychology department, demanded that interested faculty members should have a

private seminar on Rawls. We had about 15 faculty members involved led by me. It was a contested topic. Lloyd and I disagreed.

I argued that evaluation in a society necessarily involved values, as the name of the field implies, and that if evaluation were to be a social institution, it had to be concerned about justice. Rawls said, the first virtue of social institutions is social justice. Heated arguments continued for years, with many evaluators introducing justice considerations into their work, often in ways I never envisioned. As ethnic politics and feminist issues advanced, these advocates argued it was only "just" they should have their proper place in society. Eventually, social justice became a common conception in the field. Many others ignored the idea. As an early advocate, I became identified with social justice (House, 2004).

Evaluation and Validity as Argument

In summer of 1976, I was invited by Eva Baker to be a visiting scholar at UCLA. I had read a new work by two Dutch philosophers entitled *The New Rhetoric* (Perelman & Olbrects-Tyteca, 1969). They recast the old subject of rhetoric into the modern form of argumentation. I used their ideas to contend that the evaluation process itself was a form of argument where the evaluator presented evidence to bolster the case for the conclusions. Hence, conclusions were not certain, but subject to criticism, discussion, and dispute. Conclusions were plausible, not certain (House, 1977a). A bit esoteric, I thought, but T*he Logic of Evaluative Argument* monograph became popular.

For example, Lee Cronbach (1982), a major figure in measurement, used the idea to contend that determining the validity of standardized tests entailed an argument subject to evidence and challenge. Other experts such as economist Thomas Kane and psychologist Lorrie Shepard followed. Validity as argument became a concept across evaluation and measurement. A study or a test was valid subject to argument and evidence. Plausibility, not certainty, was what we sought and could achieve.

Before that, evaluators were locked into narrow processes for determining validity. These prevented the development of new ideas and use of new data collection techniques. For example, tests were often validated by correlating them with other tests similarly labeled, a bit of circular argument. Some argued that case studies provided a better picture of programs than test scores. The broader conception of validity opened a door to different approaches.

I noted something else. If you convince the leaders of a field, like Cronbach, others will follow, often begrudgingly. Intellectual disciplines are not voting democracies. Gene Glass, another leader in education research, introduced the idea of meta-analysis, a method for combining research findings quantitatively (1976). Meta-analysis not only transformed work in education but quickly spread to other social sciences and even medical research. It was revolutionary, though not without dissent. In 1978 at UCLA, Glass, Stake, and I had an adventitious conversation in which Stake and I argued for the legitimacy of qualitative studies. After long intense discussion, Glass became convinced there was a place for such studies. It was Stake's question, modestly advanced that changed minds: "If a school needed a new program to solve a problem, who would you trust to decide; an experienced principal or the results of your experiment?"

Why are leaders of disciples so important? First, they have done important

work. Second, they are highly skilled at arguing positions. They may be heavily invested in their favorite ideas, but, surprisingly, they are often open to new ways of seeing things, if the new ideas are well argued. That's true of those I've known personally, like Cronbach, Glass, and Don Campbell, the leader in experimental design in the social sciences realm.

The late 1970s were a time of great changes in my thinking. I now saw evaluation and validity in full societal context. Not only the politics, but the basic ideas themselves were socially derived from concrete particular historic origins and hence changeable, part of the ideas that we work with, the intellectual tools we are given. And if the intellectual tools are malleable, perhaps we can make them better. We can also use other disciplines to analyze the new field.

I completed *Evaluating with Validity* (1980), adding other ideas to evaluation such as argument and social justice. I wrote a paper with Norman Care, a Rawls expert, "Fair Evaluation Agreement," stipulating the conditions under which an agreement between parties could be considered fair. In another paper, I discussed the role of "coherence" as an important criterion in studies. How well do the pieces of evidence fit together? In another effort, I showed how underlying metaphors, such as industrial production, and images, such as the drunken driver, could affect evaluations. I distinguished between power and authority. The government might intervene and conduct evaluations without local participation because it had the power, but it would be better to secure cooperation through agreement and discussion. The book, although controversial at the time, eventually became an important addition to the evaluation literature. In retrospect, I was too rough on the quantitative people, which rankled them.

Professional Evaluation

Although evaluators were arguing with each other, we were also developing a new field of intellectual endeavor. I accepted the editing of a struggling evaluation journal, *New Directions in Program Evaluation*, and revived it. Evaluators started two professional organizations and merged them into the American Evaluation Association. Evaluation centers were established at several universities.

I was asked to write a regular column in *Evaluation News*, later renamed *Evaluation Practice*, and in *American Journal of Evaluation*. My dozen columns were on topics such as "Factional Disputes in Evaluation" (1984) "Drawing Evaluative Conclusions" (1986), and "Evaluating the FBI" (1988), in which I carried on a productive discussion with the head of the FBI internal evaluation unit about whether the unit was sufficiently protected from interference. My opinions were often controversial and generated discussions. I didn't particularly like writing the periodic column; I had a difficult time coming up with topics.

Although I participated in many professional events, I was not the organizer. Social organizing was never my forte. Others did much more. What I could do was think through what being an intellectual discipline meant, and for this I turned to philosophers of science, sociologists, and social historians. In *Professional Evaluation* (1993), I described how evaluation had developed from the social sciences and that the social sciences were shaped historically by political forces of the early 20th century. Social scientists back then could not criticize ideas of American "exceptionalism," which claimed there were no social classes in the US, nor could they criticize business and free enterprise. A few economists lost jobs

for supporting labor unions. Fear of socialism and communism were rampant.

That coaxed social scientists into claiming objectivity by employing quantitative techniques and ignoring social inequalities. Hence, evaluators inherited a set of techniques focusing on certain topics and methods, and neglecting others, particularly any value claims. These restricted methods were not always suited to evaluating complex social and educational programs.

I learned from philosophers, like Stephen Toulmin in *Human Understanding* (1972), that each discipline must use methods that suit the nature of what they study. Astronomers must use different methods than biologists. Program evaluators needed to discover appropriate methods for what they studied. Also, professional ethics were extremely important. Professions aspire to contribute to the welfare of society and have an ethical obligation to do so ethically. You would expect an MD to prescribe what was best for the patient, not what made the MD the most money. In addition, if it's a discipline, evaluation must maintain its honesty and integrity to be of value. Loss of integrity would be the worst fate for an intellectual discipline, and, in a hyper-capitalistic society where everything is reduced to money, an imminent danger.

Within the university, I was asked to serve on the new committee that evaluated departments. Pressures had come from the state government for accountability, and the University of Illinois decided to develop their own evaluation system before the state imposed one. They chose a professional review approach, along the lines of accreditation models that professors were familiar with. Accreditation of schools and colleges had been around since the early 20th century. Essentially, it was professionals evaluating other professionals by offering informed opinions. However, evaluating university departments internally was novel.

When I joined the committee, I was the junior person among senior professors from the powerful science departments. What they had wasn't bad, but it was new and they were amateurs. At my first meeting they were discussing doing away with the small journalism college because some of them disliked the dean. Reluctantly, in spite of my inferior status, I suggested there might be a better way to handle the situation, a less personal way. They didn't like me interfering, but they listened.

We worked out a better solution, and over time I helped professionalize the operation. The Illinois approach was one of the first university departmental evaluation systems and was widely adopted nationally and internationally (House, 1974b). Years later, I served on the first advisory panel when the Swedes decided their universities should conduct such evaluations. It was significant that the head of the Swedish higher education system, Sigbrit Franke, had a background in evaluation.

What might not be apparent to those outside academia is that professors often do not extend their disciplinary rigor to other situations. A major problem is control of biases. Academics might be rigorous inside their field, but when it comes to judging other departments and other personnel for promotion, they may fail to guard against personal biases. A discipline of evaluation would focus on their need to do so.

Causation and Methods

When I was in graduate school in 1966, James Gallagher put a model of experimental design on the board featuring a treatment group and control group. He called this the "medical model" of research.

He said it works well in medicine but not in education. No one knew why. This puzzled me, especially since such a design was often called "the gold standard."

After reading philosophers John Mackie's (1974) *Cement of the Universe* and Roy Bhaskar's (1978) *A Theory of Science* years later, I figured out why control group designs didn't work as well with education programs. The conception of causation that we inherited is called the regularity or Humean theory of causation, named after David Hume's influential analysis of cause (Hume, 1740; House, 1991). Regularity describes the conception.

Put simply, the reason that we know one event caused another event is that the first event took place before the other event regularly—regularity of succession. If an event occurred and another event occurred after it repeatedly, we would have reason to believe the events would occur together again. Succession of events is what we were after. In fact, Hume said that is all there is to causation, along with contiguity of the events. The research task is to determine the succession of events. Put succinctly, "If p, then q; p, therefore q." Following this idea, some evaluators believed the classic randomized control group design was perfect. No error could result.

However, social causation is more complex than the regularity conception indicates. Even with the same education program, there are different teachers in different places who produce different results. We might try to control for the teachers, but there are so many variables that influence the outcomes, researchers can't control all of them. Put another way, a program is not an integrated causal mechanism. Parts of the program might interact with elements in the environment to produce quite different effects.

Following Mackie, Cronbach (1982) devised a more complex formulation. "In S, all (ABC or DEF or JKL) are followed by P." In other words, in this particular setting, P, the outcome, may be determined by ABC or DEF or JKL. The problem for evaluators is that if A is the program, we only get P if conditions B and C are also present. So we could have A and not have the outcome P. More confounding, since P is caused by DEF and JKL combinations as well, we might not have the program A but still get P. Neither the presence nor the absence of A, the program, determines P. Regularity of events is not a definitive test of cause and effect. The classic control group design will not produce definitive conclusions if causation is this complex.

However, social causation is even more complex than Cronbach's formulation. Mackie's original formulation was this: "All F (A…B… or D…H… or …) are P…" The dots represent missing causal factors we don't know about. We have huge gaps in our knowledge of social events, gaps we don't know about, and gaps we don't even know we don't know about. We can never fill these gaps or be certain of all that is involved.

Sometimes qualitative studies, meta-analysis, and program theory seem to work better than experimental studies. Each approach takes account of a more complex social reality by framing the program and study more precisely, albeit in different ways. Qualitative studies show the interaction of people and events with other causal factors in context, which limits the causal possibilities and alternatives one must contend with (Maxwell, 1996). Meta-analysis uses individual studies, each of which occurred in separate circumstances of rich variation, which makes generalization more possible (Cook, 1993). Program theory delineates the domain investigated, which allows the questions evaluators pose to be

more focused, relevant, and testable (Lipsey, 1993).

In his work, Roy Bhaskar, a philosopher of science, contends that physical science investigators are able to predict and control the context and can conduct studies that yield stable findings by being able to identify and control intruding factors. But that's not possible in the social world. The social sciences do not have such replete theories and the social world is not as stable, uniform, or regular. Such designs can produce results, but they are far from certain. Findings require careful interpretation. That's why randomized controlled tests work well in evaluating drugs, but not in evaluating educational programs. The entities being examined in the physical world are different, as are the contexts of the studies and the scientific theories on which the studies are based.

As one example, in 1977 I was asked by Marge Martis at the Ford Foundation to take a look at the Follow Through evaluation. It was an attempt to determine the best early childhood program for the US. A dozen different programs were evaluated in a head-to-head contest. At first, developers of these programs were promised that outcomes suiting all programs would be used in the study. However, due to costs, these outcome measures were reduced to only a few measures, which favored particular programs. Some developers complained they were being treated unfairly.

With the support of the Foundation, I set up a panel that included Gene Glass, Les McClean, and Decker Walker (House et al., 1978). We concurred with those who analyzed the study data that the findings varied so much from site to site within each program that it was inappropriate to determine that one program was best. In other words, the same program had different results in different settings. That's what one would expect if there were many unknown causal factors at work. We disagreed with the analysts who reorganized the findings to draw conclusions in spite of this variability. Our critique received national publicity and drew counter arguments.

Beyond this, I endorsed a version of scientific realism outlined in "Realism in Research" as an alternative to the opposing qualitative/quantitative disputes (House, 1991). In the realism view, there is one complex reality; qualitative and quantitative methods offer different perspectives on this reality. This affirmed the multiple methods approach Jennifer Greene (2013) advocated. Realism as an ontology, as outlined by Bhaskar, seemed to make sense.

In 1994, Sharon Rallis and Chip Reichart organized the annual AEA conference around the theme of bringing the feuding factions together. They asked me to give the opening keynote. I suggested integrating the qualitative and quantitative perspectives around the realism ontology (House, 1994). After my talk, Peter Rossi, a demographer and major foe in the struggle, said he agreed with every word I said. He added that the argument now was about who got the money to do the studies, a bit of political realism perhaps.

Later books by Pawson and Tilley (1997) and Mark, Henry, and Julnes (2000) dealt with causation in evaluation by advancing realist conceptions similar to Bhaskar's. There remain disagreements about the utility of rigorous experimental designs. My view is they may provide useful information, if carefully interpreted.

Finance and Economics

I had never been interested in finance or economics. In 1993, I turned 56 and realized I had no plans for retirement. As I walked across campus to exercise with Phil

DeStefano, dean of education and later Chancellor of the University of Colorado, Phil mentioned his plans and the key role of investments. I decided to consider investing.

Investing was a fascinating topic, with interesting similarities and differences with evaluation. By investing my retirement funds, I coincidentally discovered I was good at the practical task of investing, which depends more on personality traits than intellectual ones, as Warren Buffet has noted. For example, if you can't stand to watch the prices of your stocks drop in the market, it doesn't matter how smart you are.

The investment literature is profuse. The statistics are familiar. For example, the Sharpe ratio comparing asset classes to one another is similar to Glass's meta-analysis comparing different psychotherapies (Smith & Glass, 1976). For most investors, the best advice remains Buffet's: Invest in the Vanguard S&P 500 fund and stick with it. The best research agrees with investing in low-cost index stock funds (Malkiel, 1973). After trying different strategies over three decades, I fully agree.

I did have an advantage. Where I grew up, I watched door to door salesmen ply shoddy goods in our working-class neighborhood. I warned my parents against their sales pitches, to no avail. The promises of the investment community sounded familiar. An essential thing to know about the investment industry is that it is riddled with conflicts of interest. Hence, low-cost index mutual funds with reliable firms are the way to invest. I have discussed investment issues from an evaluator's point of view elsewhere (House, 2015c).

As I read more economics, I saw ideas I could use. When I was in Stockholm in 1978, Herbert Simon had just won the Nobel Prize in economics and was giving his obligatory talk. Anyone could attend. His speech was an attack on neoclassic economics, particularly Milton Friedman. Simon argued that the quantitative models based on assumptions of perfect human economic rationality had little to do with how humans actually behaved. In grad school, I had read his work about humans operating with limited rationality (Simon, 1961). We had a brief chat; his presentation stuck in my mind. I saw similarities in evaluation.

How do people actually behave? In 1996, I applied transaction cost economics to education reforms suggesting how to appraise new innovations in advance. In Williamson's *Markets and Hierarchies* (1975), the key ideas are transaction costs, bounded rationality, and opportunism. The first is determined by what costs someone bears to change something, e.g., a new job means you have to move. The second is the idea that humans have limited rationality. Opportunism means that people may act in their own interest rather than that of the organization. Teachers do not always do what reformers want them to do. A potential reform might be thought through in advance by employing these criteria. I put these ideas in an article and book (House, 1996, 1998).

The writings generated some interest, but not a lot. Over the years I found that educational researchers are not interested in economic ideas. Reformers might promote the idea that markets are always correct and will magically transform things, as with school vouchers. That's an ideological slogan rather than a complex idea, used to justify what particular interest groups want. Transaction cost economics was one of my new ideas that had little influence.

Facts and Values

By the late 1990s an important task remained—to clarify the proper role of values in evaluation. The issue revolved around the fact-value dichotomy. Ken

Howe, a philosopher colleague, and I wrote *Values in Evaluation and Social Research* (House & Howe, 1999), which provided an extended analysis. The dichotomy is the belief that facts refer to one thing and values refer to a totally different entity. The dichotomy is a particularly difficult problem for evaluation since values lie at the heart of evaluation. I doubt anything in the field has caused more confusion.

The distinction between facts and values came into evaluation through the positivists influence on social science. The logical positivists thought that facts could be ascertained with certainty, and that only facts were the proper subject of science, along with analytic statements like "1 plus 1 equals two," true by definition. Facts were empirical and could be based on pristine observations, a position called "foundationalism."

On the other hand, values were feelings, emotions, or useless metaphysical entities. They were not subject to scientific analysis. People held certain values or did not. Values were chosen. Rational discussion had little to do with them. The role of the scientist was to determine facts. Others, politicians perhaps, could worry about values.

Donald Campbell accepted the fact-value dichotomy (Campbell, 1982). However, he did not accept foundationalism. He contended that there were no pristine observations on which factual claims could be based because all observations were influenced by theories and preconceptions that people held. What was a fact was influenced by prior beliefs, Knowledge was possible because, although you could not compare a fact to a pristine observation to see if the fact was true, you could compare a fact to the larger body of knowledge it related to. The fact should fit the whole body of beliefs. Occasionally, the body of knowledge had to be changed to accommodate facts. In any case, you were comparing a belief to a body of beliefs, not a belief to pure observation.

On the other hand, Campbell accepted the positivist conception of values. Values could not be determined rationally; they had to be chosen. He thought it was not the evaluator's job to choose values. Once values were determined by politicians, sponsors, or program developers, evaluators could examine the outcomes of programs and policies with criteria based on those values. This meant that evaluators could not evaluate the goals of programs, since goals were closely connected to values. Evaluators had to accept program and policy goals.

I believe Campbell had the correct idea about facts but not about values. We can deal with both facts and values rationally. Facts and values are not separate entities altogether, though they sometimes appear that way. Facts and values (factual claims and value claims) blend together in the conclusions of evaluation studies and blend together throughout evaluation studies. We might conceive facts and values as lying on a continuum like this:

Brute Facts——————— Bare Values

What we call facts and values are fact and value claims, which are sometimes expressed as fact and value statements. They are beliefs about the world. Sometimes these beliefs look as if they are factual without any value aspect, such as, "Diamonds are harder than steel." This statement fits at the left end of the continuum. There is little individual preference built into it.
A statement like "Cabernet is better than chardonnay" fits better at the right end of the continuum. It is suffused with personal taste. What about a statement like, "Follow Through is a good education program"? This statement contains aspects of both fact and

value. The evaluative claim is based on criteria from which the conclusion is drawn, and it must be based on factual claims as well. The statement fits towards the middle of the continuum, a blend of factual and value claims. Most evaluative conclusions fall towards the center of the continuum. Context makes a huge difference in how a statement functions, and a statement can be factual and evaluative simultaneously, such as "Napoleon was a great general."

Similarly, claims that might seem factual in one context might become evaluative in the context of an evaluation. Considerations like these led Scriven to see the possibility of rational discussion about evaluative statements, discarding positivism. Philosophers have similar views (Putnam, 2003).

Evaluative claims are subject to rational analysis in the way we ordinarily understand rational analysis. First, the claims can be true or false. Follow Through may or may not be a good education program. Second, we can collect evidence for and against the truth or falsity of the claim, as we do in studies. Third, the evidence can be biased or unbiased, good or bad. Finally, the procedures for the evidential assessment, whether the data are biased or unbiased, are determined by the discipline.

Of course, some claims are not easy to determine. Sometimes, it may not be possible to determine the truth of the claims. Also, we may need new ways to help us collect, determine, and process the fact-value claims. This analysis of facts and values is quite different from the fact-value dichotomy. In the old view, to the extent evaluative conclusions were value based, they were outside the purview of the evaluator. In the new view, values are subject to rational analysis by the evaluator and others.

In our book, Ken Howe and I categorized how different approaches to evaluation handled values. Howe was trained in American pragmatic philosophy, strongly influenced by John Dewey. Our approach offered three general principles to arrive at unbiased (or less biased) claims in evaluations (House &Howe, 1999, 2000). The principles are inclusion of all relevant stakeholder perspectives, values, and interests in the study; extensive dialogue between the evaluators and stakeholders; and extensive deliberation to reach valid conclusions in the study. We called this approach *deliberative democratic evaluation*, drawing on the deliberative democratic literature in political science and philosophy (Gutmann & Thompson, 1996). The book remains widely cited in discussions of values.

Turning Point

Abruptly in 2001, I reached a turning point in my life. An earlier one had arrived when I moved from the University of Illinois to the University of Colorado. My work group and work conditions at Illinois were superb. Also, it was a good place to raise a family. I was unlikely to find better elsewhere. I spent 20 years there as student and faculty. But the flatland and dreary winters wore on me. I had grown up in hills overlooking the Mississippi. In 1985, I took a job at the University of Colorado, Boulder, which had mountains and sunshine in winter.

The next turning point was in 2001. In 1999-2000, I was invited to spend a year at the Center for Advanced Study in the Behavioral Sciences, Stanford. For my obligatory talk to the other fellows, 45 social scientists, I described the developing evaluation field. I explained how the changing concepts of causation, values, and politics had played out. The talk was well received by a normally critical audience.

Some said it was one of the best of the year. The director and assistant director invited me to return another year. Alas, I never made it back.

In June, 2000, my sister Georgia died from cancer at age 60, and my Uncle Paul died a few weeks later at age 92. These deaths hit me hard. They were my links to the past. My sister had been by my side all through our traumatic childhood. She was living in our family house and had stayed in that industrial area, now racked by loss of jobs. Her attitude towards life might be characterized by her favorite song, Peggy Lee's "Is that all there is?"

My Uncle Paul had been raised in the orphanage with my father and his other brothers. Paul spent decades as a hobo before rejoining the family, reassembled by my father after his release from prison. Paul was my last link to my father's generation. I had spent hours talking to him about the old days. These losses painfully sundered personal ties to my past.

The next summer I felt a slight pain in my left arm. Soon, I found myself awaiting bypass surgery. This was a shock. For 30 years, I had been swimming 750 yards a day several times a week and looked in great condition. "Genetics," the MDs said. My left main heart artery – "the widow maker" – was 70% blocked. I could have collapsed at any moment, as Brooks Juhlin, my close friend from high school, had done the year before. I recovered quickly from quadruple bypass surgery.

I rethought my life. I was 63. Did I want to continue as usual? How many more years did I have? I decided to retire from my university position. Fortunately, my interest in investment had paid off handsomely. I had enough money to last. Sure, I could continue to accumulate more, but to what end? As a noted economist said when asked if he envied billionaires, "I have something they'll never have. I have enough." There were things I wanted to do and places I wanted to visit overseas. I was in a fortunate position in which I could continue to publish papers, give talks, and spend time wherever I wanted. Maybe I could tackle ideas others were unlikely to consider. At the end of 2001, I retired from the university.

I decided to concentrate on evaluation and abandon education. I had been doing both. Two things were wrong with education research. The first was that when a new government came into power it swept aside previous policies and pushed its own ideas. These new policies were increasingly based on ideology. My strategy of meta-evaluation was having less effect. Evidence didn't make much difference.

My second concern was the education research community. It was set in its ways. Evaluation had evolved into a separate discipline. Education research was forever focused on parsing standardized test scores. A sizable testing industry had evolved. At the center were the powerful Educational Testing Service and College Board. Although a few things had changed, such as allowing some qualitative work, the core remained the same. I saw little chance of changing it. I focused on evaluation.

Deliberative Democratic Evaluation

Heading into the new century, I developed deliberative democratic evaluation further. A prominent legal case involving segregation of minorities in Denver schools had been recast. The school district and plaintiffs representing the Hispanic community had agreed to a new bilingual program whereby those students not speaking English would move from Spanish speaking classes to mainstream classes over a few years. How the program would work

was spelled out in detail. Judge Matsch of the federal district court needed a court monitor to see that the agreed plan was implemented. I was asked if I would accept the job.

The relationship between Denver Public Schools and the Hispanic community had been long and bitter. Over decades interactions were acrimonious, with neither side trusting the other. I decided to use the deliberative democratic approach. It entailed three principles: inclusion, discussion, deliberation. I set up an advisory committee that included key people from the stakeholder groups—DPS officials, lawyers for the plaintiffs, and a lawyer from the U.S. Department of Justice, which had filed as co-plaintiff. I met with them several times a year.

I collected data about how well the program was being implemented and fed this information to the advisory group. Often, they would ask for more data, which I collected and brought back to the next meeting. This iterative process enabled us to be on the same page about the state of the program. Also, the discussions became reasonable, if occasionally heated. Some trust was established. Not all was perfect. One Hispanic activist insisted the school district was hiding large numbers of students who should be in the program but weren't. I could not find this phantom group. Not everyone was satisfied by the conciliatory process.

This monitoring went on for 5 years. I wrote periodic reports on the implementation of the program to the federal court, and these were made public. The program was big news in Denver, and the reports were eagerly awaited by the media. I asked the advisory group if they wanted me to comment publicly. They declined. When the reports came out, I refused comment beyond what was written; the stakeholders put their own slants on the reports. No one tried to tell me what to include, and I never asked. In the future, I might rethink my policy of remaining silent.

The politics and cultural dynamics were complex and fascinating. For example, the Hispanic community consisted of different subgroups. The students not speaking English were recent immigrants from Latino countries. They had little prior education. Many teachers and administrators were from the old Santa Fe culture, which had settled southern Colorado and northern New Mexico centuries ago. Others were educated Hispanics whose ancestors had immigrated some time ago. The activist lawyers were different yet again. These groups from different subcultures and social classes did not share the same interests and values. I've published details elsewhere (House, 2012). All in all, I was satisfied with the results.

Writing a Novel

I noted from my talk at the Center for Advanced Study in the Behavioral Sciences how well my story about my work in New York City had fared. Rather than explain the politics of evaluation, I told the story of how NYC politics played out in our work there. Political dynamics are difficult to grasp without specifics. I learned a lot about American politics by reading Robert Penn Warren's *All the Kings Men* (1946). Why not write a novel about evaluation politics? Grad students could learn what to expect.

In 1980, while I was still at Illinois, I had received a call from a staffer in Deputy Mayor Wagner's office in New York City. The Mayor and Chancellor of schools had initiated the Promotional Gates program, which called for flunking kids in certain grades if they did not attain specified test scores. The policy staff in the Deputy Mayor's office were skeptical it would work. They needed someone to examine the

evaluation that was being conducted by the Chancellor's research office. Their findings showed spectacular test score gains.

I said to the staffer, why not get someone close by? Plenty of experts in the NYC area. She said they are all afraid they will be blackballed from working with the schools if they are critical. The politics of NYC are rugged, she said. We called around the country, and people said you're the toughest evaluator. We need someone able to withstand the political pressures. This blatant appeal to my vanity worked. I took on the challenge.

I set up a panel with two colleagues, Bob Linn, a leading test expert, and Jim Raths, elementary education expert, with me as chair. The first thing Linn found was that in the evaluation of the summer workshops, the school researchers had failed to account for regression to the mean. Instead of huge gains, they had none. I travelled to New York to face the furious Chancellor, at his request.

The Vice Chancellor, the hatchet man, kept me waiting half an hour outside his office. It was just him and me.

"We hired you to help us, not hurt us," he said.

"We are helping you, "I said. "You don't have any test gains."

"I can hire any number of consultants to say what I want them to say. They're a dime a dozen."

"No doubt," I said, "But somewhere out there, if not in New York City, there are experts who will know you've used the wrong statistic. Let me put it this way. You're using the wrong caliber ammunition in your gun. If that's what you want to do, fire away."

Roaming the streets of the city during several days they had kept me waiting, I had come up with an appropriate metaphor. He was silent for a full minute, though it seemed much longer.

"Ok," he said, to my relief. He wanted us to help them straighten out their data analysis and not criticize them publicly. I agreed. We helped the research staff and made some suggestions. During this period I was sending confidential reports, one copy to the Mayor and one to the Chancellor. This worked well until the Village Voice got hold of our reports and printed excerpts in a front page exposé. Someone in the Comptrollers office, the third power center of NYC controlling budgets, had leaked them.

One outcome was a public hearing with a committee governing the city. During our testimony, abruptly a councilman realized we were from Illinois. He asked me why the schools were using evaluators from so far away.

"Well, Councilman," I said, "It's because no one in New York City seems to trust anybody else." The entire room burst into sustained laughter. I continued my testimony with elevated credibility. This was a drama I thought might make a good novel.

I took the New York experience and fictionalized it, changing plot and characters. To retain readers attention on a dull topic, I made each chapter short and focused on only one scene to keep the narrative moving. The literary model I had in mind was the noir detective story, where the detective stumbles into a situation he doesn't fully understand and must work through hidden complexities to discover what's really going on. I was a fan of Dashiell Hammett and Raymond Chandler. I kept *The Maltese Falcon* (1930) in mind as I put the fictional evaluator in similar situations as Humphrey Bogart in the movie (1941).

Overall, the novel was a readable attempt to portray the politics of evaluation, and it received good reviews from leaders in the field. The title *Regression to the Mean* (House, 2007) was a double entendre, referring both to an arcane statistical artifact at the center of the dispute (explained in the novel) and to basic human behavior. Basically, humans are still human and behave that way.

A few evaluators taught the novel. However, teaching a novel was strange to those in evaluation, who had mostly technical educations. Although grad students enjoyed reading it, professors found it awkward to teach or thought a novel was inappropriate in evaluation classes. From my perspective, it was another innovative idea that failed to be adopted.

It remains the sole novel about evaluation (House, 2007). It was fun to write, quite different from writing articles or memoirs. One fan was eminent evaluator Carol Weiss of Harvard, one of the best writers in the field. She really liked the novel and decided to write one herself. She gave up after 30 pages. It's more difficult than it looks, she told me. Yes, it is. It requires a mental process in which you're creating scenes in your head, far removed from social science analysis.

Biases in Evaluation

In 2006, psychologist Mel Mark, president of the AEA, asked me to give a keynote at the annual conference about something new. My first thought was to look at weaknesses in experimental designs. When I looked closely at drug studies, what surprised me was that who conducted the study made a huge difference. If the sponsors of a new drug did the study, the findings were four times more likely to be positive than if a university did it (House, 2008).

In my talk I explored precisely how deliberate biases had occurred in randomized, double blinded studies, so-called "gold standard" studies, and why they occurred. Over the years government funding for the FDA had diminished to the point where it could not fund the drug studies. The FDA asked for help from the drug companies themselves, who gladly complied. Gradually, the drug companies took over the drug studies completely. They then deliberately organized studies to show positive results for their drugs. I showed how they did this.

In his study of capitalism over several centuries, the French Annal historian Fernand Braudel (1981) explained how an extreme turn to markets eventually reduces the capacity of a government to do its work effectively. People lose faith in the government and turn to private corporations to perform certain tasks. But these organizations stress profits, not the public welfare.

This seemed to be the case with the drug studies. Defective evaluations of drugs had already resulted in hundreds of thousands of people dying. I stressed that this was a foreboding sign for evaluation. As government becomes less effective because of cutbacks, it was likely to try to control information about events, since it could not control the events themselves. This meant controlling evaluations. By 2025, this paper seemed prophetic.

Over the next few years, I pursued the idea of conflict of interest bias in evaluations further. Scriven (1976) had some excellent ideas about how to curtail biases in general. At his retirement seminar in 2011, he expounded on the notion of evaluation as a trans-discipline that could be useful in many settings. I took the opportunity to analyze the recent subprime mortgage fiasco in detail from an evaluation

perspective. I showed that step by step in the mortgage assessment process, the packaging of mortgages, and the sales to unsuspecting parties, conflict of interest biases had distorted and corrupted everything. If honest assessments of the mortgages had been done at several stages, there would not have been a massive global financial crisis (House, 2013).

My third paper recommended that conflict of interest biases be added to Campbell and Stanley's (1963) classic list of technical bias threats. I listed a dozen ways that even randomized controlled studies could be biased (House, 2011). Moving beyond technical threats to human ones was radical. Doing so might open consideration of threats to validity to other human factors. Indeed, I thought there might be such factors and that a category of human biases could be added to the technical ones.

I could also see that attending to potential biases was an excellent way of improving evaluations. I was impressed by Kahneman and Tversky's work, summarized in *Thinking, Fast and Slow* (2011). Rather than trying to model "correct" thinking, they looked at the errors people made repeatedly and suggested how to correct those. For example, they demonstrated that investors weighed losses twice as heavily as gains. Economic theory had always weighted a hundred dollar loss the same as a hundred dollar gain. In fact, investors acted differently, treating losses twice as heavily as gains. They were "loss averse" and behaved differently in certain situations.

Kahneman attributed their understanding of people's actual behavior to Herb Simon's work and the developing field of behavioral economics. Control of biases seemed to be a productive pathway for evaluation to explore in the future.

In 2017 I analyzed racial bias as a threat. The death of Michael Brown in a St Louis suburb 15 miles from my hometown startled me. I had thought that racism would fade away as older people died. Clearly, I was wrong. Drawing on recent seminal works, I analyzed racism from its inception during the European takeover of much of the world, brilliantly explicated in *Empire of Cotton* (Beckert, 2014), to its manifestation in the US. I looked at how racism is embedded unconsciously in the mind, using Kahneman's fast and slow thinking, ending with ways for evaluators to protect against the bias. In the article, I used personal examples from my family and myself, showing how racial bias is passed through families and how racism is reproduced. My personal examples added considerable punch to the piece. Sharon Rallis, editor of *AJE*, was a huge help in publishing this very long analysis (House, 2017).

All in all, I foresaw a tough future for conflicts of interest in evaluation as the society entered an intense period of hyper-capitalism in which the information sources themselves would be at risk and severely contested. Reviewing this paper, I just learned that private equity groups had purchased the companies advising and regulating drug studies.

Practical Wisdom; The Moral Fiduciary

Anyone who has conducted an evaluation knows there's the way you planned to do the study and the way you actually did it. Usually, something happens that throws you off track and forces you to wonder about what to do next. How well you handle such situations often involves practical wisdom. Good evaluation depends on practical knowledge learned through on-the-job experience.

The concept of practical wisdom can be traced back to Aristotle, who conceived the

idea by watching blacksmiths, carpenters, and other craftsmen around Athens. The workmen did not follow rigid rules. The materials they worked with and tasks they performed were too varied and irregular. New applications posed new problems. When stonemasons had to carve rounded columns, they took a stiff, flat ruler and changed it to a flexible one that could measure curved surfaces. Aristotle thought that practical wisdom consisted of such appropriate actions taken in particular work contexts.

In *Practical Wisdom,* Schwartz and Sharpe (2010) updated Aristotle's concept. They agreed that practical wisdom consists of doing the right thing in the special circumstances of performing the job. It requires the ability to perceive the situation, think about what's appropriate, and act. It involves making choices among desirable outcomes that conflict with one another and choosing between better and best possibilities.

In the evaluation community, Tom Schwandt (2005) has written insightfully about professional practice and practical knowledge. Professional practice is a complicated set of transactions in which practitioner judgments are tested. "Our everyday practice as teachers, managers, social service workers, and health care providers tells us that no escape from these dilemmas can be found. We are…always on the 'rough ground' where values, personalities, evidence, information, feelings, sensitivities, emotions, affect, ambiguities, contradictions, inconsistencies, and so forth are simultaneously in play as we try to do the right thing and do it well" (Schwandt, 2005, p. 99).

I could see that practical wisdom applied to evaluation practice as well (House, 2015b, 2023a). After explaining the concept, I used examples from my Denver experience. For example, back then people had advised me to act authoritatively and autonomously as federal court monitor. However, my experiences suggested that better results could be achieved by listening to stakeholders and involving them.

Another practical decision was about data collection. An outside firm had been hired by the school district to set up a quantitative monitoring system. Perhaps I could rely on that. However, my previous experiences indicated that such systems promised more than they delivered and took a long time to set up. I couldn't rely on that. Another possibility was to collect questionnaire data. Again, my experiences suggested self-reports produced much false information; not good for a big court case. I decided to use trained interviewers in on-site visits following a checklist that I could double check with other information.

In my book, *Evaluating: Values, Biases, and Practical Wisdom***,** I provided several Denver examples and also took scenes from my evaluation novel and analyzed scene by scene ways in which the fictional evaluator employed practical wisdom. Finally, I specified how evaluators might validate practical wisdom insights. All in all, practical wisdom in evaluation was a concept worth pursuing and others have done so (see Hurteau et al., 2023).

Another key issue in evaluation is how evaluators handle the different values, interests, and demands of various stakeholder groups in the evaluation. Different approaches manage this issue in different ways. The oldest approach is to simply deny that divergent or conflicting demands exist. The assumption is that there is one common goal and that goal is adequately reflected in the outcome measure of the study, such as test scores in education or GDP with the economy.

I have argued that as American society becomes increasingly fragmented, the idea of one measure reflecting the welfare of the entire society is less tenable (House, 2020, 2023b). Another role the evaluator might play is that of a "moral fiduciary." A fiduciary in finance is a professional consultant who is sworn to protect the interests of the client being served. If investment professionals are not fiduciaries, they are free to recommend products that make the most money for themselves, not for clients. Frankly, that's what they often do. However, if a fiduciary does that, the client can sue them. Being a fiduciary is a legal status.

A moral fiduciary in evaluation would be pledged to protect and balance the interests of all those involved in the evaluation. How the evaluator weighs and balances all interests requires many considerations and long discussion. But the point is that no one would be ignored or "sold out" in the evaluation. These decisions about weighing and balancing various interests often arise in drawing conclusions, particularly in the emphasis given to various findings. Again, this role casts the professional evaluator as a person of considerable integrity, who is able to portray and defend the varied interests in a democracy.

I now saw evaluation in an evolutionary framework. As individuals we improve our thinking by becoming aware of our mistakes and biases. At the center of our thought process lies evaluative thinking. Humans must be able to distinguish good and bad at a fundamental level. This process is modeled by fast and slow thinking. Our early evaluative thinking is heavily influenced by family, for better and worse. We become more adept in later life, as we acquire wisdom.

At the societal level, the institution of evaluation evolves, drawing from preceding ideas in the social sciences. Those ideas are modified as evaluation becomes aware of the institutional demands. It becomes more adept and complex. This occurs in liberal democratic, capitalist societies as people deal with choices they must make individually and collectively. It's not accidental that evaluation as an institution developed first in the US, the most capitalist society in the world. Evaluation as an institution depends partly on the society in which it takes place. Its future in a hyper-capitalist society in which everything is turned to profit is unclear, as is that of science.

Critical Colleagues

During a long career, how did I test my own ideas? Certainly, by publishing them. But before that I had a group of critical colleagues I sent my papers to before public exposure. Before the Internet, academics talked about their Invisible Colleges.

Although I've deliberately emphasized new ideas from outside evaluation, I was also reading the evaluation literature itself. That includes work by important pioneers such as Stake, Scriven, Stufflebeam, Michael Patton (2013), Marv Alkin (1990), and Carol Weiss (2013). I was in frequent contact with them and sometimes asked for help. These scholars could write their own versions of the early days of evaluation. Later evaluators I followed included Jennifer Greene, Tom Schwandt, Mel Mark, and Stafford Hood (Hood et al., 2005). Evaluation practice as manifested in the literature gave me a sense of the field in addition to my experiences (House, 2015d)

My initial feedback was from the CIRCE group at Illinois, led by Stake and Hastings. We traded ideas daily and papers in early drafts. Several top grad students helped with

ideas and co-authored papers over those years. I circulated papers and received comments, many of which I incorporated. All in all, the process improved my work considerably.

Beyond the CIRCE group, a few colleagues were especially helpful and influential. Steve Lapan headed the data collection effort in my first evaluation. We became close friends, and I continued to send him early drafts. Steve was an excellent teacher and had an unparalleled knowledge of public schools. He was my primary resource for what would fit the schools. He could also look at passages where I was struggling and suggest better ways of expressing the ideas.

We wrote a book together, *Survival in the Classroom* (1978), with teachers as the audience. It was by far my best seller. Steve was remarkable for his insights about teaching and about people generally. He had a strong influence on the College of Education at Northern Arizona University. My wife Donna and I visited him and his wife Pat regularly, first in Chicago, then in Sedona and San Diego. Our friendship stretched back to my first evaluation. His untimely death in 2011 was a severe personal loss.

Barry MacDonald at the CARE group in England was another strong personal influence. In 1975, I spent 4 months in England, my first long stay abroad. I learned that societies and cultures are different, a surprise for an insular Midwesterner. I learned that what Brits meant by democracy was not quite what we meant. For example, for Barry democratic evaluation meant protecting sources of information, like teachers, from possible abuse by those above. The social class structure loomed heavily in his thinking. He was an angry working-class Scot, intelligent enough that he was sent to grammar school, where he was ridiculed by faculty and students for his accent. As he became a leading evaluator, he never forgot.

When I tried to apply his ideas in the US, they sometimes backfired. For example, when I sent an evaluation report without recommendations, I nearly lost the contract. We changed the case study approach considerably, and when Howe and I developed deliberative democratic evaluation, we relied on ideas in the US. Barry and I both saw evaluation as political but our politics were different.

His ideas were valuable because he had an asocial streak that made some insights unusual. If we were talking about Michael Apple, whose work we admired, he might say, “No need to critique the ruling elite from a radical perspective. The ruling class can’t even meet the goals they set for themselves.” We spent time together in England and US, trading ideas and insults. We both had a sharp sense of humor. I learned to like English bitters and more than I needed to know about malt whiskey. As his health failed, I met him in Australia, where a daughter lived, sitting on the veranda of her Federation house in Melbourne, talking about old times. He died in 2013 (see Norris, 2015).

I learned an important lesson being overseas. Each society, country, and culture is different. I spent considerable time in England, Australia, Sweden, and Spain, and visited dozens of other counties. Norway is different from Sweden, and Australia different from New Zealand. Whenever I gave a talk in another country, I prefaced it with, “I’m going to talk about this project in Denver. Every society is different. If you find something useful, good. If not, ignore what we did.” When I hear Americans advocating something overseas, I shudder. What they are selling probably didn’t work in the US either.

Beyond work contacts, two long-time personal friends reviewed my papers. Gary Waldman, a physicist, and I met in our honors rhetoric class our first year in college. We became close friends through college until now. Gary has the far-ranging intelligence and liberal arts background to enable him to understand topics outside physics. He's an ardent admirer of Proust, reading *Remembrance of Things Past* three times. I managed it once. Our friendship has continued for 70 years.

Dave Harvey, a sociologist, is from my hometown. We reconnected in graduate school. Later, as he took a job at the University of Nevada, Reno, I visited him often. His left-wing, critical theory orientation kept me aware of how society was perceived outside the prevailing social consensus. It has been productive to balance diverse views of society against each other. Dave has a great sense of humor, like most of my close friends.

Biography, Family, and Deep Values

As I aged and moved away from active practice, I continued my interest in influences on evaluations by considering personal biography. I knew many evaluators and had an idea how their backgrounds affected what they did. However, I didn't want to speculate. If I were to explore the issue, I should consider myself, even though self-examination challenges credibility.

I had an advantage. I had already written a childhood memoir. Years ago one of my children said something demeaning about children from less wealthy homes. What could my kids know, living in a middle-class professor's home? To inform them, I began writing about my childhood. As I wrote about events, like my father's death, I dredged up emotions associated with those events. I relived the emotions and, in a sense, reconfigured them. The writing became therapeutic. Even today, rereading about my dad's death might bring tears to my eyes, 84 years after the event. Such is the power of memories.

I wrote that memoir in the 1970s and revised it from time to time. I didn't publish it while the people were alive; I didn't want to hurt anyone. My mother would have been deeply disturbed by some revelations. By 2015, everyone mentioned had died, except for me. I published the memoir, *Cherry Street Alley* (House, 2015a). I also published a paper about how my childhood had affected my work (House, 2015b). David Williams (2016) collected similar reflections of seven influential "pioneers" in the field, including me.

Into my late 80s, I have continued to explore family influences on evaluations. In the first memoir, I described my childhood until age 9 when I had developed a critical, independent way of thinking. In retrospect, I see how I became a professional evaluator, a vocational field that did not exist until three decades later, and that I helped create. I helped develop a field in which I belonged.

Over the years I extended this venture into memories by creating portraits of family members who influenced me, including my mother, grandmother, sister, stepfather, and two uncles. I continued the story of schooling through high school. In these portraits I've focused on my interactions with relatives and strived to understand their lives apart from mine. My original motive was to remember people who deserved to be remembered. My current interest is to discern how early interactions influenced what I did later. Even if I don't see connections, perhaps future scholars looking at these detailed personal interactions might see connections to my work.

My sense is that the things evaluators do are influenced by what I call "deep values." For example, why did I puzzle about what happened to students not chosen for the gifted classes after my first evaluation? I trace this concern back to my mother, who was extremely moral. She worried about the welfare of others, even people she didn't know. I was never as concerned about others as she was, but the sensitivity was there.

I recognize the similarity of some childhood thoughts to later ideas. One is that the role of the evaluator should be that of a "moral fiduciary," one who protects the interests of those involved in the evaluation, an echo of my belief that I needed to be an independent thinker to protect my sister and myself. The moral fiduciary idea emerged seven decades after childhood, no doubt mixed with other experiences.

During many intense debates, I was subject to strong criticisms. The toughness of my father and uncles, who had undergone hard times as orphans, had also shaped me. I could withstand criticisms and respond appropriately. Of course, other scholars advocated ideas from their own backgrounds. Ultimately, early events influence behavior but do not determine it entirely.

Although my professional ideas were often based on seminal works in philosophy and social science, the motivation, direction, and values that led me to develop those ideas intellectually came from personal experience and family history. This enduring influence I call "deep values" because the values are long term, pervasive, embedded, and partly subconscious. More generally, I think deep values are biologically emergent. They come from individual experiences and those of family, with family experiences conveyed through family stories and behaviors. They are an important part of "acquired wisdom" in the deep sense.

As I look back nearly 60 years to the beginning of the evaluation endeavor, I have a feeling of solid accomplishment. We, my colleagues and I, built something that helps people and makes society better, something worthwhile. In my contribution, I sense the morality of my mother and the verve and toughness of my dad and uncles. Since none had the opportunity to attend high school, they might not have understood what evaluation is about, but they would be proud of my contribution. I hope going forward that evaluation can maintain its honesty and integrity, so essential to its social and intellectual significance.

References

Alkin, M. (1990). *Debates in evaluation*. SAGE.

Beckert, S. (2014). *Empire of cotton: A global history*. Vintage.

Bhaskar, R. (1978). *A theory of science*. The Harvester Press.

Braudel, F. (1981, 1982, 1984). *Civilization and capitalism, 15th to 18th centuries*, 3 Vol. Harper Row.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Rand McNally.

Campbell, D. (1982). Experiments as arguments. *Evaluation Studies Review Annual*, *7*, 117-128.

Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds). *Understanding causes and generalizing about them.* [New Directions in Evaluation, No. 57] (pp. 39-82). Jossey-Bass.

Chomsky, N. (1957). *Syntactic structures*. Mouton de Gruyter.

Cronbach, L. (1982). *Designing evaluations of educational and social programs*. Jossey-Bass.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(1), 3-8.

Greene, J. C. (2013). Making the world a better place through evaluation. In M. C. Alkin (Ed.). *Evaluation roots* (2nd ed., pp. 208-217). SAGE.

Gutmann, A., & Thompson, D. (1996). *Democracy and disagreement*. Belknap.

Hammett, D. (1929). *The Maltese falcon*. Knopf.

Hood, S., Hopson, R., & Frierson, H. (Eds.). (2005). *The role of culture and cultural context*. Information Age Publishing.

House, E. R. (1974a). *The politics of educational innovation*. McCutchan Publishing.

House. E. R. (1974b). The politics of evaluation in higher education. *Journal of Higher Education*, *45*(8), 618-627. Reprinted in F. G. Caro (Ed.), (1977). *Readings in evaluation research* (2nd ed.). Russell Sage Foundation.

House, E. R. (1975). Justice in evaluation. In G. V Glass (Ed.), *Evaluation studies review annual Vol. I*. (pp. 75-100) SAGE Publications. Also in R. A. Schmuck & P. J. Runkel (Eds.) (1977). *Second handbook of organization development in schools*. (pp. 505-517). Mayfield Press.

House, E. R. (1977a). *The logic of evaluative argument.* Center for the Study of Evaluation, University of California at Los Angeles.

House, E. R. (1980, 2010). *Evaluating with validity*. SAGE. Reissued, Information Age. (Also in Spanish).

House, E. R. (1981). Three perspectives on innovation: Technological, political and cultural. In R. Lehming & M. Kane (Eds.). *Improving schools: Using what we know*. SAGE.

House, E. R. (1984). Factional disputes in evaluation. *Evaluation News, 5*(3), 19-21.

House, E. R. (1986). Drawing evaluative conclusions. *Evaluation Practice, 7*(3), 35-39.

House, E. R. (1988a). *Jesse Jackson and the politics of charisma: The rise and fall of the PUSH/Excel program*. Westview Press.

House, E. R. (1988b) Evaluating the F.B.I. *Evaluation Practice, 9*(3), 43-46.

House, E. R. (1991). Realism in Research. *Educational Researcher, 20*(6), 2-9.

House, E. R. (1993). *Professional evaluation: Social impact and political consequences*. SAGE.

House, E. R. (1994). Integrating the quantitative and qualitative. In C. R. Reichart & S. F. Rallis (Eds), *The qualitative quantitative debate: New perspectives.* [New Directions in Evaluation, No. 61] (pp. 13-22). Jossey-Bass.

House, E. R. (1996). A framework for appraising educational reforms. *Educational Researcher. 25*(7), 6-14.

House, E. R. (1998). *Schools for sale: Why free market policies won't improve America's schools and what will*. Teachers College Press. (Also in Spanish.)

House, E. R. (2004) Social justice. In S. Mathison (Ed.). *Encyclopedia of evaluation* (pp. 393-396). SAGE.

House, E. R. (2007). *Regression to the mean: A novel of evaluation politics*. Information Age Publishing.

House, E. R. (2008). Blowback: Consequences of evaluation for evaluation. *American Journal of Evaluation, 29*(4) 416-426.

House, E. R. (2011). Conflict of interest and Campbellian validity. In H. T. Chen, S. I. Donaldson & M. M. Marks (Eds), *Advancing validity in outcome evaluation: Theory and practice* [New Directions in Evaluation, No. 130] (pp. 69-80). Jossey-Bass.

House, E. R. (2012). Democratizing qualitative research. In S. D. Lapan, D. M. Q. Quartaroli & F. J. Reimer (Eds.). *Qualitative research* (pp. 451-472). Jossey-Bass.

House, E. R. (2013). Evaluation's conflicted future. Pp. 63-72 in Donaldson, S. I. (Ed). *The future of evaluation in society: A tribute to Michael Scriven.* Information Age.

House, E. R. (2015a). *Cherry street alley*. Create Space, Amazon.

House, E. R. (2015b). *Evaluating: Values, biases, and practical wisdom*. Information Age Publishing.

House E. R. (2015c). Decision Making via evaluation: What's Marv's opinion worth. In C. Christie & and A. Vo (Eds). *Issues in evaluation use and decision making in society. A tribute to Marvin C. Alkin*. Information Age.

House, E.R. Interview. (2015d). The oral history of evaluation: The professional development of Ernest House. Interview by AEA Oral History Project Team: R. L. Miller, J. King, M. Mark & V. Caracelli. *American Journal of Evaluation, 36*(2) 270-232.

House, E. R. (2016). Childhood influences on my work. Pp. 83-88 in Williams, D. D. (Ed). Childhood influences on the work of seven North American evaluation pioneers. *New Directions in Evaluation, 150*.

House, E. R. (2017). Evaluation and the framing of race. *American Journal of Evaluation, 38*(2), 167-189.

House. E. R. (2020). Evaluating in a fragmented society. *Journal of Multi-Disciplinary Evaluation. 16*(36), 26-36.

House, E. R. (2023a). The practical wisdom of evaluators. In M. Hurteau & T. Archibald (Eds.). *Practical wisdom for an ethical evaluation practic*e. Information Age Publishing.

House, E. R. (2023b). Deliberative democratic evaluation: The evaluator as a moral fiduciary. In M. C. Alkin & C. A. Christie (Eds). *Evaluation roots* (3rd ed., pp. 110-117). SAGE.

House, E. R., Glass, G. V., McLean, L. D., & Walker, D. (1978). No simple answer: Critique of the Follow Through evaluation. *Harvard Educational Review, 48*(2), 128-160. Reprinted in T. D. Cook et al., *Evaluation studies review annual*, Vol. 3, SAGE.

House, E. R., & Howe. K. R. (1999). *Values in evaluation and social research.* SAGE Publications. (Also in Spanish and Chinese)

House, E. R., & Howe, K. R. (2000). Deliberative democratic evaluation. Pp. 3-12 in Ryan, K. E. & DeStefano, L. (Eds). *New Directions in Evaluation*, *85*(1), 3-12.

House, E. R., Kerins, T., & Steele, J. M. (1972). A test of the research and development model of change. *Educational Administration Quarterly*, *8*(1), 1-14.

House, E. R., & Lapan, S. G. (1978). *Survival in the classroom*. Allyn and Bacon.

House, E. R., & Madura, W. (1988). Race, gender, and jobs: Losing ground on employment. *Policy Sciences, 21*(2), 351-382.

House, E. R., & McQuillan, P. (1998). Perspectives on innovation. Pp. 198-213 in Lieberman, A., Fullan, M. Hargreaves, A. & Hopkins, D. (Eds.). *International handbook of educational change*. Kluwer.

House, E. R., Rivers, W., & Stufflebeam, D. L. (1974). An assessment of the Michigan accountability system. *Phi Delta Kappan*, *55*(10), 663-669.

Hurteau, M., & Archibald, T. (Eds.). (2023). *Practical wisdom for an ethical evaluation practice*. Information Age Publishing.

Hume, D. (1978) (1740). *An abstract of a treatise of human nature* (2nd ed.). Clarendon Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Strauss, and Giroux.

Lipsey, M. W. (1993). Theory as method: Small theories of treatments. In L. B. Sechrest & A. G. Scott (Eds.). *Understanding causes and generalizing about them.* [New Directions in Evaluation, 57] (pp. 5-38). Jossey-Bass.

Koestler, A. (1964). *The act of creation*. McMillan.

Mackie, J. L. (1974). *The cement of the universe*. Clarendon Press.

Malkiel, B. G. (1973). *A random walk down Wall Street*. W. W. Norton.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation: An integrated framework*. Jossey-Bass.

Maxwell, J. A. (1996). *Using qualitative research to develop causal explanations*. [Working Papers]. Harvard Project on Schooling and Children.

Norris, N. (2015). Democratic evaluation: The work and ideas of Barry MacDonald. *Evaluation, 21*(2), 135-142.

Patton, M. Q. (2013). The roots of utilization focused evaluation. In M. C. Alkin (Ed.), *Evaluation roots* (2nd ed., pp. 292-303). SAGE.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. SAGE Publications.

Proust, M. (1981). *Remembrance of things past*. Random House.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Schwandt, T. A. (2005). The centrality of practice to evaluation. *American Journal of Evaluation, 26*(1), 95-105.

Schwartz, B., & Sharpe, K. (2010). *Practical wisdom.* Riverside Books.

Scriven, M. (1966). Causes, connections, and conditions in history. In W. H. Dray (Ed.). *Philosophical analysis and history* (pp. 240-254). Harper and Row.

Scriven, M. (1976). Evaluation bias and its control. In G. V. Glass, *Evaluation studies review annual* (Vol 1, pp. 119-139). SAGE Publications.

Simon, H. (1961). *Administrative behavior*. McMillan.

Smith, M. L., & Glass, G. V (1976). Meta-analysis of psychotherapy outcomes studies. *American Psychologist, 32*(9), 752-760.

Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record, 68*(7), 523-540.

Stake, R. E. (1995). *The art of case study research*. SAGE Publications.

Toulmin, S. (1972). *Human understanding. Vol. 1*. Princeton University Press.

Warren, R. P. (1946). *All the king's men*. Harcourt Brace.

Weiss, C. H. (2013). Rooting for evaluation: Digging into beliefs. In M. C. Alkin (Ed.). *Evaluation roots* (2nd ed., pp. 130-143). SAGE Publications.

Williamson, O. E. (1975). *Markets and hierarchies*. Free Press.

Acquired Wisdom Series