

Evaluation Use Theory, Practice, and Future Research: Reflections on the Alkin and King *AJE* Series

American Journal of Evaluation
2020, Vol. 41(4) 581-602
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1098214020919498
journals.sagepub.com/home/aje



Michael Quinn Patton¹ 

Abstract

Marvin Alkin and Jean King published three *AJE* articles on evaluation use over four years, a coherent and comprehensive series covering the historical development of evaluation use, definitions and factors associated with use and misuse, and theories of evaluation use and influence, concluding with assessment of the first 50 years of use research. They conclude with recommendations for future theory development and research on evaluation. I draw a different set of conclusions and pathway forward. Where they seek a common universal operational definition of evaluation use, I propose treating use as a thick sensitizing concept that invites diversity of context-specific meanings. Where they find evaluation use theory inadequate, I argue that it is sufficient for its purpose. Where they seek more development of evaluation-specific utilization theory, I propose drawing on more established and validated theories from social sciences to explain and illuminate evaluation use as occurring in complex dynamic systems.

Keywords

utilization-focused evaluation, use studies, theory, defining evaluation, sensitizing concepts

It's all about criteria. Criteria are the basis for evaluative judgment. Determining that something is good or bad, successful or unsuccessful, works or doesn't work, and so forth and so on requires criteria, which brings us to the remarkable *AJE* series on evaluation use by Marvin Alkin and Jean King that addresses, among many core issues, by what criteria theories of evaluation should be judged.

I am afraid that many *AJE* readers may have missed the cumulative significance of their three articles because they appeared over a span of four years. Together they form a coherent and comprehensive set; anyone teaching evaluation would do well to package them as a set for student engagement. Taken together, they are a *tour de force*. They cover the historical development of

¹ Utilization-Focused Evaluation, St. Paul, MN, USA

Corresponding Author:

Michael Quinn Patton, Utilization-Focused Evaluation, 740 Mississippi River Blvd. South, Suite 21E, St. Paul, MN 55116, USA.

Email: mquinnp@gmail.com

evaluation use (Alkin & King, 2016), definitions and factors associated with evaluation use and misuse (Alkin & King, 2017), and theories of evaluation use and influence, concluding with thoughts on the first 50 years of use research (King & Alkin, 2019). Their overview and synthesis are comprehensive, insightful, and generative. They conclude with important recommendations for future theory development and research on evaluation use based on criteria for definition sufficiency and theory validity. Using different criteria, I draw a different set of conclusions, and in this addendum to their series, I offer a different pathway forward. The Alkin and King series crystalized for me this alternative path.

Defining Use

Alkin and King open their second article on definitions with my quote of Samuel Butler: “Definitions are a kind of scratching and generally leave a sore place more sore than it was before” (Patton, 1988, p. 304). They aspire to avoid additional soreness, but in my case, the effect has been the opposite. To alleviate the increased definitional soreness they’ve aggravated, I was stimulated to write this rejoinder. They review the great diversity of definitions of both evaluation and use and examine the interplay between definitions and theories. They agree with critics that these constructs are “ill-defined and diffuse,” though they think that distinguishing use from influence helps. But, they go on to lament “a lack of specificity of these concepts’ outcomes.” They ask: “What exactly does evaluation use or evaluation influence, a type of use, look like in different settings?” They then conclude:

If scholars can agree on a definition—ours or someone else’s—and explicit outcomes of evaluation use and create validated instruments for measuring it across various locales, studies would have common metrics for comparison across contexts.

To summarize, we believe that future research should pay close attention to the evolving contexts of evaluation use and of *the need for a common definition and outcomes* . . . (King & Alkin, 2019, p. 451, emphasis added).

Later, on the same page, they call for “evaluation use scholars” . . . to use “*common outcome measures* then the important features of use contexts and of the mechanisms that result in use in them may finally make the content of a unifying theory evident” (p. 451, emphasis added).

In an analysis and synthesis that describes diverse approaches and emphasizes the importance of contextual variation and sensitivity throughout, in the end, when it comes to recommendations, they fall back on the old positivist paradigm admonition that the solution to all conceptual challenges is standardized operational measurement. The search for precision is dominant, alluring, and misguided, at least from a pragmatic constructivist perspective. Look at the questions they pose and the adverb at the core of their questions.

- What **exactly** does it mean for evaluation “to speak truth to power”? (p. 450)
- What **exactly** does evaluation use or evaluation influence, a type of use, look like in different settings? (p. 451)
- What **exactly** does involvement or engagement mean in relation to use, and how might you measure it? Who **exactly** needs to be involved? (p. 452, boldface emphasis added)

Asking for exactness evokes a desire for *precision* and risks doing so at the expense of *accuracy*. Precision is concerned with *exact* measurement reliability, producing consistent results across people and contexts. Accuracy is concerned with correspondence to reality, a focus on validity. Intelligence tests are precise but inaccurate, as are personality tests of all kinds. Where human beings are concerned, standardization leads to the illusion of accuracy by achieving precision.

Exhibit 1. Alternative Inquiry Questions for Research on Evaluation.

King & Alkin (2019) Search for Precision	Alternative Search for Understanding Diversity
1. What exactly does it mean for evaluation “to speak truth to power”? (p. 450)	1. What are the variety of ways evaluators speak truth to power? What are the implications of that diversity?
2. What exactly does evaluation use or evaluation influence, a type of use, look like in different settings? (p. 451)	2. What are the diverse manifestations of evaluation use and influence in different settings? What are the implications of that diversity?
3. What exactly does involvement or engagement mean in relation to use, and how might you measure it?	3. What various forms do involvement or engagement take in relation to use, and how might these be documented and understood?
4. Who exactly needs to be involved? (p. 452)	4. What are the options for involvement, the basis for choosing among options, and the implications of those choices?

Nobel-prize-winning psychologist and decision scientist, Daniel Kahneman (2011), in his best-selling book, *Thinking, Fast and Slow*, describes developing instruments to select military personnel for officer training in Israel that had the advantage of being very precise and manifestly inaccurate, but the precision gave the illusion of meaningfulness and utility, and so the military embraced the inaccurate instrumentation. I doubt not that someone can create a standardized instrument with metrics for measuring use across various locales to allow comparison across contexts, but to do so would sacrifice accuracy (reality) and meaningfulness for precision because the reality is that use cannot be characterized by a universal operational outcome or uniform metrics. The reality is that use is variable, contextual, diverse, messy, and complex. Exhibit 1 contrasts the King and Alkin “exactly” questions with “diversity” questions. The contrasting questions reveal different approaches to understanding evaluation and use.

The Alkin and King Definition of Evaluation Use

Alkin and King (2017) note that “since the earliest discussions of evaluation use, most evaluation writers have not provided a *precise* definition of the term, but instead commented on the components of what counts as use” (p. 436, emphasis added). After reviewing various attempts to define evaluation use, they provide their own.

In 1979, Alkin, Daillak, and White attempted to reconcile the then prevalent and seemingly disparate definitions. Their definition took the form of an adapted Guttman-mapping sentence, a format helpful as a theoretical depiction. Almost 40 years later and based on what we now know, we provide a definition of evaluation use in this format (see Exhibit 2) and examine the five “matrices” of the Guttman-mapping sentence that define instances of use. (Alkin & King, 2017, p. 437)

The framework is labeled “Use defined with an adapted Guttman-scaling mapping sentence.” In presenting the framework in Exhibit 2, Alkin and King (2017) conclude that “taken together, the five matrices in this inclusive Guttman-scale definition allow an explicit definition of instances of evaluation use” (p. 439).

But the “mapping sentence” is not a *definition*. It is a framework for describing evaluation use in a particular instance. It is quite valuable as a framework for documenting the nature of evaluation use, but what the mapping sentence generates at the end (the final frame following “IS” at the end of the mapping sequence), when the content in each of the five parts has been specified, is evaluation use in

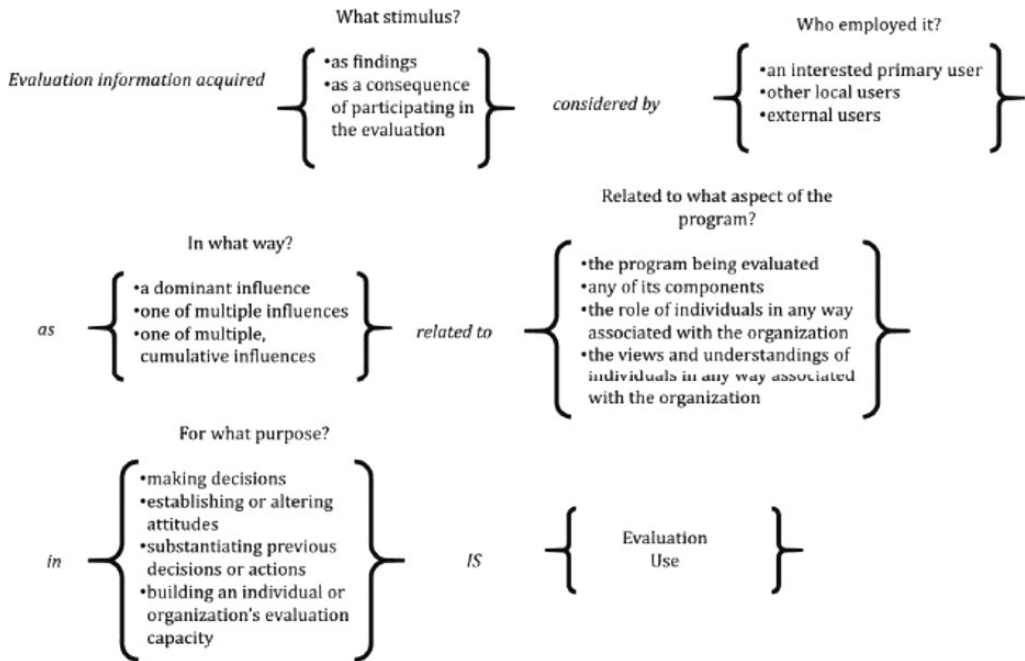


Exhibit 2. Use defined with an adapted Guttman-scale mapping sequence.

Source: Alkin and King (2017, p. 439).

a specific instance and context. It is a conceptual framework for describing evaluation use based on dimensions hypothesized by evaluation use theories to be explanatory.

Now remember, what they are striving for

If scholars agree on a definition—ours or someone else’s—and explicit outcomes of evaluation use and create validated instruments for measuring it across various locales, studies would have common metrics for comparison across contexts.

To summarize, we believe that future research should pay close attention to the evolving contexts of evaluation use and of *the need for a common definition and outcomes . . .* (King & Alkin, 2019, p. 451, emphasis added).

Let’s look at the adequacy of their proposed dimensions as the basis for a common definition. To save space, I’ll examine just two—the first and last.

Stimulus. The option for “stimulus” for use is evaluation information acquired “as findings” or “as a consequence of participating in an evaluation.” That is a modest start, but consider operationalizing (identifying and measuring) all the forms the “stimulus” might actually take. Here is a beginning list to illustrate the complexity involved: conclusions, interpretations, and/or recommendations with what degree of clarity; certain results, tendencies, possibilities, mixed findings, muddled findings, and/or inconclusive findings; positive, negative, or mixed; statistical analyses, qualitative patterns and themes, and/or merged methods; firsthand receipt, secondhand receipt, or thirdhand receipt of stimuli. The possibilities go on and on.

For what purpose. Alkin and King list four illustrative purposes. In fact, there are hundreds of possibilities of which here are a few: improvement, accountability, knowledge generation,

development, empowerment, cutting funding, scaling, and/or encouraging dialogue; opening debate, fomenting debate, deepening debate, enlarging debate, and/or ending debate; and compliance, solving problems, and/or transformation. There is no definitive operational list of purposes because there cannot be; the possibilities are too vast. We do have some broad, sensitizing categories of purposes (e.g., formative, summative, developmental, knowledge generating, accountability, and monitoring—see Patton, 2008, chapter 4). But, operationalizing use with the Alkin and King conceptual map, or any other precise definition, will be constraining and, ultimately, inadequate and inaccurate. We need to think beyond positivist operationalism. Pragmatic constructivism offers an alternative.

Beyond Operationalism: Use as a Thick Sensitizing Concept

Use should be treated as a thick sensitizing concept, not an operational concept. I raised this distinction with regard to process use (Cousins, 2007; Patton, 1998, 2007) when Amo and Cousins (2007) sought to operationally define process use for the same reasons that Alkin and King want to operationally define use and influence with standardized metrics. Let me repeat and elaborate the argument.

In a dominant paradigm of social science research, concepts only become real and meaningful when they have been operationalized, which means that the concept can be standardized and measured quantitatively. *The SAGE Encyclopedia of Social Science Research Methods*, in an entry on operationalization, affirms the scientific goal of standardizing definitions of key concepts. It notes that concepts vary in their degree of abstractness, using, as an illustration, the concepts human capital versus education versus number of years of schooling as moving from high abstraction to operationalization. The entry then observes,

Social science theories that are more abstract are usually viewed as being the most useful for advancing knowledge. However, as concepts become more abstract, reaching agreement on appropriate measurement strategies becomes more difficult. (Mueller, 2004, p. 162)

This is interesting. Abstraction is useful for advancing knowledge and building theory. Use and influence are somewhat abstract, to be sure, and this very quality of abstraction makes it difficult to reach agreement on how to measure (operationalize) them. The entry continued,

Social science researchers do not use [operationalization] as much as in the past, primarily because of the negative connotation associated with its use in certain contexts. (p. 162)

The entry discusses the controversy surrounding the relationship between the concept of intelligence and the operationalization of intelligence through intelligence tests, including the classic critique that the splendidly abstract concept *intelligence* has been reduced by psychometricians to what intelligence tests measure. Here, we have a dramatic manifestation of banality, taking a critically important idea—intelligence—and reducing it to what psychometricians can measure on a universal standardized test. Increasingly, as the entry notes, researchers have recognized this slippery slope.

Operationalization as a value has been criticized because it reduces the concept to the operations used to measure it, what is sometimes called “raw empiricism.” As a consequence, few researchers define their concepts by how they are operationalized. Instead, nominal definitions are used . . . and measurement of the concepts is viewed as a distinct and different activity. Researchers realized that measures do not perfectly capture concepts, although . . . the goal is to obtain measures that validly and reliably capture the concepts. (p. 162)

It appears that there is something of a conundrum here, some tension between social science theorizing and empirical research. Yet, a second entry in the *Encyclopedia of Social Science Research Methods* sheds more light on this issue.

Operationalism began life in the natural sciences . . . and is a variant of positivism. It specifies that scientific concepts must be linked to instrumental procedures in order to determine their values In the social sciences, operationalism enjoyed a brief spell of acclaim.

. . . Operationalism remained fairly uncontroversial while the natural and social sciences were dominated by POSITIVISM but was an apparent casualty of the latter's fall from grace. (emphasis in the original; M. Williams, 2004, pp. 768–769).

The entry elaborates three problems with operationalization, each of which applies to the challenge of defining use. First, *underdetermination* is the problem of determining “if testable propositions fully operationalize a theory” (p. 769). Examples include concepts such as homelessness, poverty, and alienation that have variable meanings in different social contexts. What *homeless* means varies historically and sociologically.

The second problem is that objective scholarly definitions may not capture the subjective definition of those who experience something. Poverty offers an example: What one person considers poverty, another may view as a pretty decent life. The Northwest Area Foundation, which has as its mission poverty alleviation, has struggled to try to operationalize poverty for outcomes evaluation; they found that many quite poor people in states like Iowa and Montana, who fit every official definition of being in poverty, did not even see themselves as *poor* much less *in poverty*. For them, poverty was an inner-city phenomenon not relevant to their rural communities.

The third problem is disagreement among social scientists about how to define and operationalize key concepts. The second and third problems are related, in that one researcher may use a local and context-specific definition to solve the second problem but that context-specific definition is likely to be different from and in conflict with the definition used by other researchers inquiring in other contexts. One way to solve the problem of definition is to abandon the search for a standardized and universal operational definition and treat use as a sensitizing concept.

Sensitizing and Illuminating Concepts

Sociologist, Herbert Blumer (1954), is credited with originating the idea of the “sensitizing concept” to orient fieldwork. Sensitizing concepts in the social sciences include loosely operationalized notions such as victim, stress, stigma, and learning organization that can provide some initial direction to a study as one enquires into how the concept is given meaning in a particular place or set of circumstances (Schwandt, 2001). The observer moves between the sensitizing concept and the real world of social experience giving shape and substance to the concept and elaborating the conceptual framework with varied manifestations of the concept. Such an approach recognizes that while the specific manifestations of social phenomena vary by time, space, and circumstance, the sensitizing concept is a container for capturing, holding, and examining these manifestations to better understand the patterns and implications.

Philosopher Elizabeth Minnich worked with Hannah Arendt on “illuminating insights, ideas that ask to be brought into conversation.” Trying to narrowly and operationally define such illuminating ideas leads to “conceptual errors,” what Hannah Arendt called, more dramatically, “metaphysical fallacies” (E. K. Minnich, personal communication, November 11, 2019; Minnich & Patton, 2020). Trying to operationalize sensitizing concepts is such a “conceptual error.” I like the *gravitas* of “metaphysical fallacy.” Sensitizing concepts constitute illuminating insights about something that deserves attention and, to be sure, conversation. But such conversations and dialogues are not

expected to yield operational definitions. Instead, they yield conceptual insights that are illuminative about manifestations of situational diversity and complexity.

Thick Concepts

More recently, ethical and linguistic philosophers have been delving into and unpacking “thick concepts.” This, I suggest, is another way of thinking about and understanding use, a pragmatic constructivist alternative to operationalization. Thick concepts are especially relevant to evaluation because they are terms that are both descriptive and evaluative.

B. Williams first introduced the phrase “thick concept” in his 1985 book, *Ethics and the Limits of Philosophy* . . . [H]is use of the phrase was assimilated from Geertz’s (1973) notion of a thick description—an anthropologist’s tool for describing “a multiplicity of complex conceptual structures, many of them superimposed upon or knotted into one another.” Incidentally, Geertz borrowed the phrase “thick description” from Ryle (1971) who took thick description to be a way of categorizing actions and personality traits by reference to intentions, desires, and beliefs. Although Geertz’s and Ryle’s notions of thick description influenced Williams’ terminology, their notions did not necessarily involve evaluation. By contrast, Williams’ notion of a thick concept is bound up with both evaluation and description. Or, in Williams’ terms, thick concepts are both “action-guiding” and “guided by the world.” They are action-guiding in that they typically indicate the presence of reasons for action, and they are world-guided in that their correct application depends on how the world is (B. Williams, 1985, pp. 128, 140–141; Kyle, n.d., p. 1).

I find the notion of treating use as a thick concept quite compelling and illuminative: Use is action-guiding in that the very term indicates the presence of reasons for action, and use is world-guided in that its correct application depends on how the world is (i.e., the situation and context within which use occurs or is intended to occur).

Thick Evaluation

Linguistic philosopher, Simon Kirchin (2019), has written a book entitled *Thick Evaluation* that examines in great depth “the crucial distinction between ‘thin’ and ‘thick’ concepts.”

We use evaluative terms and concepts every day. We call actions right and wrong, teachers wise and ignorant, and pictures elegant and grotesque. Philosophers place evaluative concepts into two camps. Thin concepts, such as goodness and badness, and rightness and wrongness have evaluative content, but they supposedly have no or hardly any nonevaluative, descriptive content: they supposedly give little or no specific idea about the character of the person or thing described. In contrast, thick concepts such as kindness, elegance and wisdom supposedly give a more specific idea of people or things. Yet, given typical linguistic conventions, thick concepts also convey evaluation. Kind people are often viewed positively whilst ignorance has negative connotations. (online)

Kirchin focuses on the debate between “separationists” and “nonseparationists.” Separationists think that thick concepts can be separated into component parts of distinct evaluative meaning, exposing their often very “thin” and nonevaluative content, while “nonseparationists” deny this and treat thick concepts as wholistic in meaning. Thick evaluation argues for a version of nonseparationism that would, without irony, include the concept of evaluation itself and evaluation use.

Without explicitly designating the terms as sensitizing, illuminating, and/or thick, evaluators commonly use concepts that exhibit those characteristics to inform their understanding of what to pay attention to, and the resulting mindfulness is both descriptive (what is happening) and prescriptive (i.e., evaluative, guiding what to do). Consider the notion of *context*, which King and Alkin emphasize as critical to understanding use and influence. Any particular evaluation is designed within some *context*, and we are admonished to take *context* into account, be sensitive to *context*,

and watch out for changes in *context*. But what is *context*? In 2009, the theme of the annual conference of the American Evaluation Association (AEA) was *context and evaluation*. Animated discussions ensued among those attempting to operationally and specifically define context and those comfortable with contextual variations in meaning. Those seeking an operational definition of context ranted in some frustration about the ambiguity, vagueness, and diverse meanings of what they, ultimately, decided was a useless and vacuous concept. Why? Because it had not been (and could not be) operationally defined—and they displayed a low tolerance for the ambiguity that is inherent in such sensitizing thick concepts.

Defining Use as a Thick Sensitizing Concept

Exhibit 3 provides straightforward, broad definitions of use, utilization, utility, and other concepts related to evaluation use. In this approach to conceptual definitions, *evaluation use is a thick sensitizing concept*. A thick sensitizing concept raises consciousness about something and alerts us to watch out for it within a specific context and to take action according to what we observe and understand (interpret and judge) flowing from the observance. That is what the concept of *evaluative use* does. It says, evaluations are used in a variety of ways by diverse people in different contexts. Watch out for how people are using evaluations. Pay attention. Attempt to understand what is happening. The evaluation may be producing insights and learnings quite apart from recommendations for action. Work to understand what is going on as people engage with evaluation. Help the people in the situation pay attention to their use options, if that seems appropriate and useful. Perhaps even make systematic evaluation use a matter of *intentionality*.

But do not judge the maturity and utility of the concept of use by whether it has “achieved” a standardized and universally accepted operational definition. Judge it instead by its utility in sensitizing us to pay attention to the variety of ways in which evaluation leads to action based on interpreting findings and, beyond findings, any thinking and/or action that is stimulated by the very processes of determining evaluation questions, establishing evaluation criteria for judging success, making methods choices, interpreting data, and rendering judgments of merit, worth, significance, and goodness.

Exhibit 3. Nominal Definitions of Use as Thick Sensitizing Concepts.

Evaluation Use: Whatever understandings, learnings, actions, changes, attitudes, and/or knowledge follow from evaluation findings and/or processes.

Evaluation Users: Individuals, program and project personnel, decision makers, and policy makers, and anyone who uses an evaluation in any way.

Primary Intended Users: Specific people expected to use a specific evaluation.

Primary Intended Uses: What intended users expect to do with an evaluation.

Findings Use: Application of evaluation findings.

Process use: Effects of an evaluation on those involved in some way (the effects of the processes of evaluation from beginning to end).

Utilization: The processes by which, and degree to which, an evaluation has effects, both intended and unintended.

Utility: Potential applications of evaluation findings, usability.

Defining Pluralism and Use

I once was asked to consult with an international initiative aimed at promoting pluralism in the world. Evaluators assigned to this task of evaluating whether pluralism was increasing or decreasing

in the world came to see me for assistance in coming up with a standardized instrument for measuring pluralism. After they presented this challenge, I said, “Let me repeat back to you what I hear you asking. I want to be sure I understand. You want to create a standardized measure of pluralism?”

Yes, they affirmed, that was their assignment. I said, “Let me say that again and invite you to listen very carefully. You want to create a standardized measure of PLURALISM.” They nodded affirmatively. “Let me try one more time. You . . . want . . . to . . . create . . . a . . . standardized . . . instrument . . . to . . . measure . . . *pluralism*.” That third time they understood and said, “So, perhaps it is not possible?” “Oh, it’s possible,” I replied. “It’s just not desirable or sensible, but it’s possible.”

Creating a standardized, universal measure of use based on a common operational definition is as problematic, in my judgment, as is a standardized, universal definition of pluralism, for USE is PLURALISTIC.

Defining Evaluation

While we are considering definitions, we might as well take on the even larger challenge of defining evaluation. In 2014, a group of Canadian evaluators addressed the issue of defining evaluation under the auspices of the Canadian Evaluation Society (CES). They reviewed the evaluation literature and found many competing and conflicting definitions. They held interactive sessions with CES members in a consultative process that included social media and in-person discussions. They concluded that “no single definition currently exists. Further, there are indications that a shared definition would be difficult to achieve . . .” (Poth et al., 2014, p. 87). They then concluded that not having a consensus definition is fine.

[A] definition is currently neither necessary nor desirable. Rather than developing a definition, perhaps a greater contribution to evaluation practice would be to embrace the multifaceted nature of evaluation and expertise of evaluators by describing the continuum of purposes, approaches, activities, and contexts in use within the Canadian evaluation context. This would provide the tools to talk about the strength of the breadth of the evaluation function and its ability to respond to the needs of various organizations and truly make a difference. (Poth et al., 2014, p. 99)

But, the notion that the field needs a definitive, universal, standardized, and consensus definition won’t go away. Most recently, Gullickson (2020) undertook an extensive review of attempts to define evaluation, likening the current state of diverse definitions to “the tale of the blind men of Indostan, who disputed furiously about whether the same elephant was a wall, snake, rope, tree, spear or fan because they were each touching a different part” (p. 1). She argues that “if evaluation is to emerge as a distinct profession, a more robust and agreed-upon definition is needed” (p. 1).

All parties need a shared understanding of what constitutes evaluation and evaluation quality to know whether they are delivering it/getting it. Understanding quality in evaluation begins at the same place as it does for any other evaluand—with the definition. (p. 1)

There’s the rub. For any other evaluand, the definition is contextual. Gullickson undertakes a comprehensive review of evaluation definitions, the logic of evaluation, and offers a universal definition based on two core criteria: “fully describe, fully judge.” Earlier, I took issue with the adverb “exactly” as seeking a level of precision that is neither possible nor necessary. What exactly is use? Let me illustrate by applying that adverb to these criteria: What exactly is “fully”? To “fully” describe and “fully” judge evokes for me the old ideal of “economic man” who, in all his sexist glory, had complete information and made completely rational decisions. Such a being never existed. What constitutes “fully” with regard to either description or judgment will be subject to

interpretation and context. Where words and human beings come together, variations in meaning come with the territory.

In the end, the definition Gullickson offers, though based on “fully describe, fully judge,” is less aspirational.

Evaluation is the generation of a credible and systematic determination of merit, worth, and/or significance of an object through the application of defensible criteria and standards to demonstrably relevant empirical facts. (p. 4)

She then proceeds at length to argue that this definition will solve problems of professional identity, certification, competencies, debates about approaches, communications among evaluators, communications with the outside world, training evaluators, and focusing research on evaluation. She concludes that the main reason that her definitive definition will be resisted is the self-serving interests of evaluation theorists and practitioners who advocate their own narrow perspectives which would be diminished in importance under this universal definition. Indeed, many who call themselves evaluators would be exposed as not being engaged in evaluation at all—Ouch.

Gullickson, like many others she cites, believes that evaluation needs a universal, agreed-on, and standardized definition to gain credibility and flourish as a profession. Has the field been languishing? We have Voluntary Organizations for Professional Evaluation in more than 125 countries around the world. All major national and international organizations have evaluation units. Annual conferences of the major professional organizations have been growing. Our evaluation journals have high impact ratings.

Gullickson believes that “because we lack a cohesive definition,” we “reduce our ability to create influence on society” (p. 7). My experience in this regard is limited, to be sure, but I’ve never heard of any policy maker, or person of influence or, indeed, anyone at all, say, “until you evaluators have an agreed-on definition of evaluation, what you say lacks credibility.”

The search for a universal, standardized, agreed-on, and enforced definition remains alluring to many. It is based on an exclusive view of the profession, a “clearly bounded profession” (p. 7), one in which, as Gullickson argues, the boundaries are clear about what must be done (“fully describe, fully judge”), what constitutes evaluation and what does not, and who is in and who is out. An inclusive view, in contrast, offers a big tent approach with diverse perspectives and contextually specific definitions. Let me add that I quite appreciate Gullickson’s definition and will cite and use it. I just don’t think it should be endorsed as THE definitive definition of evaluation nor should any other.

Everything I have said about the thick sensitizing nature of use applies doubly so to the term evaluation. We don’t have a common definition of evaluation. We do have a dominant definition: Scriven’s definition of evaluation as judging the merit, worth, and significance of something. Scriven’s criteria are the core of the definition Gullickson offers. The determining merit and worth definition have been with us for over 50 years and appear, from my reading, to be the most commonly cited in evaluation and research textbooks. Yet, few understand what it actually means, which may help explain why it is not yet universally agreed-to. I do evaluation workshops for hundreds of participants every year. I routinely asked participants to define evaluation and get a variety of definitions. More experienced evaluators usually cite some version of Scriven’s definition using the terms “merit” and “worth,” which were the focus of his original definition. I then ask them what the difference is between merit and worth. Virtually no one can make the distinction, which is actually quite an important one. They will hesitantly guess at the difference, but even when I tell them what the two meanings are, they remain unsure which meaning applies to worth and which applies to merit. Later, Scriven added significance. So, evaluation is judgment of merit, worth, and significance—and just what does that clarify? Aren’t merit and worth significant? Significance is

certainly a judgment, or at least it should be, but not one that lends itself to operationalization. It is an interpretation, indeed, a contextual interpretation that flows from thoughtful dialogue.

I cite as evidence in this regard the American Statistical Association (ASA) manifesto on statistical significance. In a special issue of ASA's official journal, *American Statistician* (March 2019), the editors banned the term "statistical significance." Banned it and any variations of it. Their rationale is that $p < .05$ (or $.01$) has always been arbitrary and through overuse, has become a rigid, mechanistic target without merit or worth, or for that matter, significance. Substantive significance (what the data mean) trumps statistical significance. In place of statistical significance, the *American Statistician* offers four principles to guide statistical interpretation:

1. accept ambiguity,
2. be thoughtful,
3. be open, and
4. be modest.

These are the principles that I suggest should inform discussions about how to define both evaluation and evaluation use. Accept ambiguity. There is not, cannot be, and should not be a standardized, universal definition of evaluation. When I chaired an AEA Task Force to define evaluation, we generated instead a discussion of variations on the theme and recommended that no official definition of evaluation be adopted by AEA, but that, instead, the description of the varied meanings and applications of evaluation be left open for ongoing thoughtful discussion. (For the statement on what evaluation is, see AEA, 2014; for how that statement was generated, who was involved, and why we recommended living with ambiguity, see Patton, 2018, pp. 6–17).

So, accept ambiguity, stay open, and be thoughtful about what use means. Place and keep any particular definitions in context. New manifestations of use will continue to emerge with new approaches to evaluation, among which I would mention developmental evaluation (Patton, 2011), principles-focused evaluation (Patton, 2018), and Blue Marble Evaluation (Patton, 2020). These emergent types of use were not anticipated in the early days of studying use.

For a final thought on accepting ambiguity, let me offer the insight of another evaluation theorist. Distinguished evaluation scholar, philosopher of science, and recipient of the AEA Paul F. Lazarsfeld Theory Award (2002, "presented to an individual whose written work on evaluation theory has led to fruitful debates on the assumptions, goals, and practices of evaluation"), Tom Schwandt suggests making peace with the inherent and inevitable ambiguity of wide-ranging notions like evaluative thinking.

Clarity and some degree of agreement on, if not precision in, evaluation terminology is of course useful, for if we are to nurture evaluative thinking among evaluators and stakeholders we need to be fairly clear on just what we are aiming to foster. Nonetheless, if complete conceptual clarity were a necessary prerequisite to effective learning and practice, we would likely never escape the weariness and, often, the tedium of definitional debate, particularly in the field of social inquiry that does not necessarily demand precision in use of certain terms. (quoted by Vo & Archibald, 2018, p. 140)

Let me emphasize that final sentence: "Particularly in the field of social inquiry that does not necessarily demand precision in use of certain terms." Turns out that definitional ambiguity isn't just the case for social inquiry and evaluation. The 2019 Nobel Prize in Physics went to James Peebles for his contributions to our understanding of the evolution of the universe and Earth's place in the cosmos. Peebles' research has revealed a universe in which just 5% of its content is known, the matter that constitutes stars, planets, trees—and us. The rest, 95%, is unknown dark matter and dark energy, a mystery, and challenge to modern physics, which cannot define what either dark matter or dark energy actually is.

Or consider this from biology. The latest research on human origins argues that an oasis known as the Makgadikgadi–Okavango wetland along Botswana’s Zambezi River was the ancestral “homeland” for all modern humans today. The researchers studied mitochondrial DNA—“genetic material stored in the powerhouse of our cells that is passed from mother to child”—of current residents across southern Africa. Then, “they layered the genetic data with an analysis of past climate and modern linguistics, as well as cultural and geographic distributions of local populations” (Wei-Hass, 2019). In commenting on the increasing complexity and ambiguity of human origins, Carina Schlebusch, an evolutionary geneticist at Uppsala University in Sweden, observed that

all of this work also circles the increasingly confusing definition of a species. While humans like to put everything in boxes, nature doesn’t fit into tidy categories, Schlebusch says. There are no distinct lines between one species and the next; everything works in shades of gray. (Wei-Hass, 2019)

Wittgenstein’s Use Theory of Meaning

In contemplating definitions, classifications, and categories, let us move from biology to philosophy. Philosopher Ludwig Wittgenstein (1889–1951) studied, illuminated, and philosophized about language. In contemplating the nature and challenges of defining words, he conceptualized a *Use Theory of Meaning* which posits that “meaning is use.”

For a large class of cases—though not for all—in which we employ the word meaning it can be defined thus: the meaning of a word is its use in the language. (Wittgenstein, 1953, Section 43)

By this, he meant that words are not defined by reference to the things they identify and distinguish nor by the mental associations they evoke but by how they are used. He argued that people do not need a formal definition of a word to use it successfully within a context where the word has socially understood meaning. Common usage precedes formal definition and can obviate any necessity for fixing the meaning of a word in a standardized definition. Words do not begin with definitions. They begin with usage. Definitions emerge from being used within the culture and society in which they are used. To understand how language works for most cases, Wittgenstein explained, we have to observe how it functions in a particular social situation. People can and do have conversations about objects and ideas without operational specificity or generating a definition that consists of necessary and sufficient conditions. Applied to evaluation, Wittgenstein’s philosophy of language posits (or at least hypothesizes) that stakeholders and intended users generally understand what we mean when we ask how an evaluation has been used. We can have conversations about use without operational specificity or generating a definition that consists of necessary and sufficient conditions.

Criteria for Reviewing Evaluation Use Theories

Let us now move from accepting ambiguity and being thoughtful, open, and modest about definitions to being likewise thoughtful and open to judgments about the state of evaluation utilization theory. Theories aim to explain and predict both how and why some phenomenon unfolds as it does. Explanation and prediction are different, each challenging in its own way, and each tricky. Social psychologist Warren Thorngate (1976) posited that it is impossible for a scientific explanation to be simultaneously general, accurate, and simple. Any explanation that meets two of these criteria will violate the third.¹ Keep that caveat in mind as we review the state of evaluation use theory in general and utilization-focused evaluation (U-FE) theory in particular. King and Alkin (2019) draw on and integrate frameworks for reviewing evaluation use theories that result in the following five criteria:

1. Operational specificity: Explicit details are given about how to foster evaluation use for studies in specific settings.
2. Range of application: Explicit description is provided of where the theory is likely to increase use and where it is not likely to succeed.
3. Feasibility in practice: Practitioners can easily and routinely conduct the activities.
4. Discernible impact: The prescribed activities do, in fact, lead to increased use.
5. Reproducibility: Different practitioners can reproduce the same outcomes (i.e., use) at different times and places.

King and Alkin conclude that applying this set of criteria for evaluating use theories

reaffirms the fact that the field is a long way from having an empirical-grounded descriptive theory of evaluation use. On the one hand, at least scholars know what theory is needed ultimately. On the other hand, however, absent such a descriptive theory (or theories), evaluation practitioners may continue to adopt and adapt prescriptive theories with which they are familiar without knowing exactly what to do to increase the likelihood of potential use.

I take exception to this conclusion that evaluation lacks an empirically grounded descriptive theory of evaluation use. The problem is how the criteria are applied and interpreted. I'll consider the example of U-FE to illustrate how different interpretations of criteria lead to different conclusions.

Applying Theory Criteria to U-FE as an Example

In applying the criteria listed above to U-FE, King and Alkin conclude that U-FE meets four of the five criteria, the exception being discernible impact. When I apply these criteria to U-FE, I reach different conclusions.²With regard to the definition of evaluation use offered by Alkin and King (2017) and discussed above (Exhibit 2), they commented,

In fact, Patton's utilization-focused evaluation (Patton, 2008), first published in 1978, systematically attends to each component, identifying the primary intended users, detailing their hoped-for primary intended uses, and then interacting with them over the course of the evaluation to ensure (to the extent possible) that the process generates credible information that can result in meaningful use. (p. 440)

That judgment sets the stage to review whether and how U-FE meets (or does not meet) the criteria for theory. In what follows, I'll offer my review of each of the five criteria as applied to U-FE.

1. *Operational specificity*: "Explicit details are given about how to foster use for studies in specific settings."

Commentary: For reasons presented earlier, I'm skeptical of operational specificity, either definitionally or procedurally. U-FE offers a set of principles that guide contextual adaptation, but it is the very nature of contextual variability that operational specificity is neither appropriate nor possible. I would replace "operational specificity" with the following criterion for judging theories that provide guidance under conditions of complexity and great contextual variation: *Guiding principles for contextual adaptation*. Guiding principles must meet the GUIDE criteria for principles-focused evaluation (Patton, 2018), which U-FE does (more on this below).

2. *Range of application*: “Explicit description is provided of where the theory is likely to increase use and where it is not likely to succeed.”

Commentary: The keyword in this criterion is “likely,” that the theory offers guidance on increasing the likelihood (probability) of use. Remember this distinction because it becomes quite important below. I agree with King and Alkin that U-FE clearly meets this criterion.

3. *Feasibility in practice*: “Practitioners can easily and routinely conduct the activities.”

Commentary: I take exception to interpreting feasibility as meaning “easy” and “routine.” Feasible means doable. U-FE may be hard to do, as it often is, for situational adaptiveness and being active–reactive–interactive–adaptive in complex dynamic systems is challenging—but doable (feasible). U-FE principles provide guidance for feasibility none of which reduces to “easy” and “routine.” U-FE is rarely either. I know Alkin and King as experienced, accomplished, and effective practitioners. We have often discussed our experiences and clients. I don’t remember ever characterizing our work as “easy” and “routine.” In fact, we often seek out and take on the difficult and nonroutine because such situations are challenging, usually important, and test the outer limits of theory in practice.

4. *Discernible impact*: “The prescribed activities do, in fact, lead to increased use.”

Commentary: King and Alkin conclude that though U-FE mostly meets four criteria for a useful theory of use,

it certainly falls short on the remaining criterion: discernible impact. To say with confidence that prescribed U-FE activities consistently (i.e., always) lead to increased use is to ignore the contextual nature of evaluation settings where, for example, the elimination of a PIU, U-FE’s Achilles’ heel (Patton, 2008, pp. 566–567), can scuttle even the most masterful U-FE study. As noted previously, there simply are no guarantees. And if this most detailed of prescriptive theories cannot meet all five criteria for a meaningful theory of evaluation use, then other less detailed theories will surely fail to meet them.

I disagree with this analysis and conclusion on two counts. First, theories concerning humans offer probabilities not certainties. The criterion that a theory must “consistently (i.e., always)” lead to increased use would make such a proposition a law not a theory. How did “consistently” become a parenthetical “always?” Consistently and always are quite different levels of certainty. Consistently speaks to a high level of probability. Always denotes complete determinacy.

Secondly, they conclude that the U-FE caution that turnover in primary intended users threatens use makes the theory inconsistent and therefore fails to meet the criterion of discernible impact. To the contrary, I would suggest that the U-FE caution about potential loss of primary intended users exemplifies a probability-based *if-then* theoretical proposition, the implications of which provide clear guidance to practitioners. Here is U-FE theory regarding intended user loss and turnover stated as a theory-based proposition.

The smaller the number of primary intended users (one being the lower limit) and the longer the duration of the evaluation, the greater the risk that intended use by intended users will fail to occur due to loss or turnover of primary intended users. Practice implications for evaluators are as follows:

- Identify and engage with multiple intended users as appropriate and possibly to avoid over-dependence on a single user or small number of primary intended users.
- Anticipate the increased likelihood of intended user loss or turnover for evaluations of long duration.

- Be alert to and monitor any signs of potential or actual loss or turnover of intended users and adapt accordingly.
- Cultivate new (or renewed) intended uses by new intended users when loss or turnover occurs.

Far from failing to meet the criterion of discernible impact, U-FE identifies the factors that may affect impact (use) and provides guidance about mitigating those factors. That is the essence of theory in practice. Moreover, in this example, discernible impact overlaps with range of application: “Explicit description is provided of where the theory is likely to increase use and where it is not likely to succeed.”

5. *Reproducibility*: “Different practitioners can reproduce the same outcomes (i.e., use) at different times and places.”

Commentary: King and Alkin address the reproducibility of U-FE as follows:

Novices learning about U-FE often lament the fact that they lack the intellectual and interpersonal skills of Michael Quinn Patton. “I’ll never be as good as he is,” they moan. “His skill set can’t be reproduced!” But what **exactly** would it mean for different U-FE practitioners to produce the same outcomes (i.e., use) at different times and places? It seems possible to us that different U-FE proponents could effectively engage in the practice at different times and in different contexts and successfully foster use. The challenge for this criterion may well rest in what we mean by “use” and how we measure it. (p. 438)

Rejecting the criterion of reproducibility

King and Alkin come close, but stop short, of rejecting the *reproducibility* criterion. I take that step. In so doing, let me draw on the wisdom of pioneering evaluation theorist (recipient of the AEA Lazarsfeld Theory Award in 1988) Bob Stake, who has insightfully articulated an alternative criterion. In a thoughtful review, he examined how individual evaluators will inevitably engage in evaluation somewhat differently and achieve diverse results because methods alone do not determine findings, conclusions, and judgments. The evaluator’s own perspective and values come into play. Rather than urging the evaluation profession to become ever more diligent in procedures to ensure objectivity, neutrality, and reproducibility, he asserted the inevitability of diversity and, in the end, considered it not only unavoidable but probably a good thing—on one condition: that those who commission, fund, participate in, and conduct evaluations acknowledge diversity and its implications. He even offered an evaluation principle to that effect. Here, then, in full, because both the tone and substance deserve savoring, is Stake’s proposal for a diversity principle that can be adapted as a criterion for judging a theory.

[M]uch will differ from evaluator to evaluator. Most of us aspire to a professional practice by which—hypothetically—all evaluators evaluating a single evaluand would produce largely the same findings. But it is not an attainable aspiration, and to force it to happen would invite disaster. Evaluators cannot help but see some things differently. Some findings will be different. Hopefully not often completely at odds, but that too will happen. In the complex determination of program quality and accomplishment, there is no single reality we can capture. Reality is constructed by people, and people sometimes differ. When we agree on what we see, we tend to think we see correctly, but sometimes we do not. When we disagree on what we see, we tend to think one sees incorrectly, but sometimes both see correctly.

We have an evaluation practice that is influenced by the value commitments of the evaluator and a set of operating standards that imply we can attain a widely agreed-upon picture of merit and worth. Something has to give. It could be that we should more effectively constrain our value commitments and search harder for meta-evaluation consensus, but we clearly should develop our standards and

principles so that they deal better with the uncertainty and individuality of evaluating. One of the guiding principles should say something like:

It should be expected that any two competent evaluators, working together or apart, will seldom agree fully on criteria and standards, critical incidents, and experience and on the appropriateness of the evidence of merit and worth. The full use of validation, triangulation, and meta-evaluation is essential, but it will not eliminate uncertainty in the evaluation findings.

Evaluators should be encouraged to “have a life” and to “have a dream” so their interpretations are enriched by personal experience. Comprehensive, idiosyncratic interpretations are small steps toward saving the world. (Stake, 2004, p. 107, emphasis added).

In place of the reproducibility criterion (“Different practitioners can reproduce the same outcomes [i.e., use] at different times and places”), I would offer an adaptation of Stake’s diversity principle:

Diversity criterion (proposed as new): Evaluation of an evaluation takes into account and transparently documents the effects of the characteristics and credibility of the evaluator(s) on evaluation processes and outcomes (uses).

Developing Evaluation Theory

I want to pose an alternative approach to constructing evaluation theory. Alkin and King have reviewed theories of evaluation use, including U-FE, which I have just argued constitutes a theory that meets appropriate criteria for theory. But U-FE is not a stand-alone theoretical approach or framework. It is based on multiple, integrated theories of change from social sciences. As an interdisciplinary field, evaluation can and should draw on relevant, well-conceptualized, and empirically validated theories from diverse arenas of inquiry that can be integrated to support a comprehensive theory of evaluation use. Evaluators do not have sufficient research support to independently validate a utilization theory, but drawing on fields that do have such resources provides a theoretical foundation for understanding and guiding evaluation use. This makes it all the more important that evaluation scholars keep pace with developments in other fields that can contribute important insights into evaluation theory and practice.

U-FE, as an example, has from the beginning been built on multidisciplinary theoretical propositions and explanatory frameworks relevant to evaluation use. For example, evaluation use theory and practice is supported by sociological theories of power, which I have used to help me explain to intended users how and why their involvement in an U-FE is in their own best interest. Sociological theory explains why, how, and under what circumstances knowledge is power. In essence, use of evaluation will occur in direct proportion to its power-enhancing capability. Power-enhancing capability is determined as follows: *The power of evaluation varies directly with the degree to which the findings reduce the uncertainty of action for specific stakeholders.* This view of the relationship between evaluation and power is derived from the classic organizational theories of Crozier (1964) and Thompson (1967). Crozier found that power relationships develop around uncertainties. Every group tries to limit its dependence on others and, correspondingly, enlarge its own areas of discretion. Evaluation can reduce uncertainty and enhance power. These classic theory sources have been validated and updated in more recent scholarship (Hall, 2020; Patton, 2014, 2015, 2017, 2018a).

Other important contributions to evaluation use theories include diffusion of innovation theories (Rogers, 1962; Rogers & Shoemaker, 1971), social innovation and scaling theories (Westley & Antadze, 2010; Westley et al., 2006, 2017, 2011, 2013), and prospect theory about decision making (Kahneman, 2011). Several interdisciplinary fields of inquiry have provided insights into how we

manage situation recognition in the face of complexity. Simon's (1957, 1978) classic works on bounded rationality and satisficing reveal how we reduce complexity to a manageable few adequate possibilities. Contingency theory, from organizational sociology, emphasizes how organizational decision making in complex open systems requires ongoing adaptation; there can be no one best practice because what is appropriate is contingent on the kinds of tasks being done and the volatility and dynamism of the environment in which adaptive and contingent decisions are made (Morgan, 2006, pp. 42–45). The field of cognitive science has investigated neuronetwork learning algorithms that constitute shortcuts for making sense of complexity (Torres, 2018). Decision sciences have been identifying decision heuristics that cut through the messy, confusing, overwhelming chaos of the real world, so that we can avoid analysis paralysis and take action. We rely on routine *heuristics*—rules of thumb, standard operating procedures, practiced behaviors, and selective perception (Kahneman & Tversky, 2000). These theories explain variations in evaluation use, conditions and factors that affect use, and practices that can enhance use. Social psychology theory offers insights into human and organizational behavior that are relevant to evaluation use (Mark et al., 2011). Standpoint theory addresses issues of power and privilege in professional relationships (Hall, 2020).

Theory Knitting, Layering Theories, Theory Ladders

Leeuw and Donaldson (2015) have done a masterful job of reviewing approaches to integrating multiple theories.

Theory knitting is integrating parts of (at first sight unrelated or loosely coupled) theories and by doing so not only reduce 'theoretical segregation' but also increase the chances of accumulation . . . In theory knitting, one attempts to integrate previous theories into a single higher order theory rather than to segregate a new theory from previous ones.

But, they caution, not all the types of theory can be knitted together, only "those that are more or less similar, in terms of their type of content and orientation" (p. 474). Theories relevant to evaluation use are excellent candidates for theory knitting. They also examine layering theories and theory ladders, based on Westthorp (2012), to address a

well-known problem in evaluation: that interventions work for some but not for others, or work for some for a long time and for others for only a very short time . . . The very concept of *theory layering* not only makes it possible to understand complex phenomena and to predict what will work for whom, but it also helps the evaluator to find variables that are in need of measurement at different system levels. (p. 475, emphasis in the original)

Again, however, they caution that, "As is the case with regard to theory knitting, layering theories can only apply to theories of a similar orientation and type of content" (p. 475). They also mention the strategy of nesting systems and subsystems of "mechanism-based explanations" into a theory hierarchy (Lieberman, 2005; Marra, 2011).

Theory knitting is the explicit and purposeful integration of social science theory in program design and program theories. The unit of analysis for theory knitting is a program or project. Transdisciplinary theory is generated when members of different disciplines use "a shared conceptual framework drawing together discipline-specific theories, concepts, and approaches to address a common problem" (Slatin et al., 2004, p. 62). The focus for evaluation use theory is the relevant set of integrated theories that illuminate and explain variations in utilization, what amounts to theory synthesis (Lemire, Whynot, & Montague, 2019).

A Cornucopia of Theories

Communications theory points to the importance of crafting messages to specific audiences. The overarching principle of U-FE to focus on intended use by intended users epitomizes and is undergirded by communications theory. Systems theory guides us to look at interrelations, perspectives, boundaries, and dynamics. Complexity theory directs us to watch for nonlinearities, emergence, cocreation, and adaptation in the face of changing conditions. Diffusion of innovations theory identifies how the characteristics of the thing to be adopted (evaluation findings) are affected by the characteristics of the innovation (e.g., understandability, perceived relevance, and incentives) and the credibility and status of those promoting adoption of an innovation (aka evaluators).

It strikes me, then, that evaluation use theory can best be constructed as a quilt of relevant applied social science theories rather than approached as a distinct phenomenon, requiring its own theory. Hall's (2020) layering and integration of standpoint theories constitutes an exemplar in this regard.

Conclusion

I want to end where I began: applauding and appreciating the important contribution Alkin and King have made to evaluation scholarship. I want to reiterate that, taken together, their three articles are an unprecedented *tour de force*. Their overview and synthesis are comprehensive, insightful, and generative.

I agree completely with their closing call for future research to "pay close attention to the evolving contexts of evaluation use We believe that describing and documenting context in careful detail is imperative." Where we disagree is their corresponding (and I believe contradictory) call for "a common definition [of use] and outcomes." Treating evaluation generally and evaluation use specifically as thick sensitizing concepts invites ongoing dialogue about the diverse nature of evaluation and, correspondingly, the diverse nature of evaluation use. Wittgenstein's philosophy of language is informative and insightful in this regard. He wrote:

Hegel seems to me to be always wanting to say that things which look different are really the same, whereas my interest is in showing that things which look the same are really different. I was thinking of using as a motto for my book a quotation from King Lear: "I'll teach you differences."

Wittgenstein's "motto" directs our attention to particularities rather than generalities. Customizing not standardizing has been the core of my own 50 years of U-FE practice. That same appreciation of case-by-case, situation-by-situation, circumstance-by-circumstance, study-by-study, evaluation-by-evaluation accumulation of knowledge, understanding, and wisdom was articulated 40 years ago by pioneering evaluation thought leader Lee J. Cronbach in positing and posting, as part of the Stanford Evaluation Consortium, *95 Theses Toward Reforming Program Evaluation*. I share you with two of my favorites while urging you to visit (or revisit) the entire list.

4. An evaluation of a particular program is only an episode in the continuing evolution of thought about a problem area.

65. Each evaluation study analyzes a spoonful dipped from a sea of uncertainties.

(Cronbach & Associates, 1980)

Rather than future research on evaluation being aimed at classifying evaluations in standardized categories of use and influence, a sensitizing concept approach to research on evaluation would study and elaborate how evaluation use and influence varies by context, purpose, intended and unintended users, and intended and unintended uses. The evaluation use framework developed by Alkin and presented in Exhibit 2 can guide research on evaluation when the elements of the

framework are treated as sensitizing dimensions and concepts, abandoning the notion of operationalizing those concepts and dimensions through standardized definitions and measurement scales. In essence, a pragmatic constructivist epistemology—emphasis on socially constructed perspectives and meanings within diverse contexts—would replace a positivist (standardized operationalization) orientation.

It is not just in seeking standardized definitions that positivism remains deeply influential in recommendations for future research on evaluation. Coryn et al. (2017) conducted a systematic review of research on evaluation published between 2005 and 2014. They found mostly descriptive studies and called on those conducting research on evaluation to move beyond descriptive studies “to more explicitly establish an evidence base of causal relationships in order to better inform evaluation practice.” This recommendation treats evaluation practice and evaluation use as discrete things that are “caused” by specific variables that can be controlled and predicted to inform evaluation practice and use. Framing research on evaluation as a search for linear causal models that support replication and standardization is out of sync with systems thinking and complexity theory. Understanding evaluation as part of complex dynamic systems will direct future research on evaluation to map interrelationships, capture diverse perspectives, document both linearities and nonlinearities, pursue both intended pathways and emergent ones, and examine the interplay between what was planned and done, what was planned and undone, what was unplanned and done, and what was omitted altogether, all in dynamic interaction, interdependence, and interrelationship. Evaluation use as a discrete thing to be operationally defined and causally explained will distort not illuminate, will confine not expand our understanding, will create the illusion of certainty when wisdom lies in engaging ambiguities, and is a lingering manifestation of positivist research hubris, when an ever more turbulent and uncertain world invites us to go gently into future inquiries, being open, modest, and thoughtful. Evaluation use theory, informed by and knitted together from more general social science theories and philosophical understandings, will offer all the explanatory and predictive power possible in complex dynamic systems.

Acknowledgment

I acknowledge and thank Justus Randolph, Associate Editor, *American Journal of Evaluation*, for suggesting the relevance of Ludwig Wittgenstein during the review process.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Michael Quinn Patton  <https://orcid.org/0000-0002-4706-2941>

Notes

1. I am indebted to *AJE* editor, George Julnes, for this observation and reference.
2. As a matter of full disclosure, and for those who may not know, I am the originator and propagator of utilization-focused evaluation (Patton, 1978, 1986, 1997, 2008, 2015) and therefore approach this issue from a perspective of advocacy, having a stake in the outcome of the discussion and debate.

References

- Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* (Vol. 76). Sage Library of Social Research.
- Alkin, M. C., & King, J. A. (2016). The historical development of evaluation use. *American Journal of Evaluation*, 37(4), 568–579.
- Alkin, M. C., & King, J. A. (2017). Definitions and factors associated with evaluation use and misuse. *American Journal of Evaluation*, 38(3), 434–450.
- American Evaluation Association. (2014). *What is evaluation?* <https://www.eval.org/d/do/492>
- Amo, C., & Cousins, J. B. (2007). Going through the process: An examination of the operationalization of process use in empirical research on evaluation. *New Directions for Evaluation*, 116, 5–26.
- Blumer, H. (1954). What is wrong with social theory? *American Sociological Review*, 19, 3–10.
- Coryn, C. L. S., Wilson, L. N., Westine, C. D., Hobson, K. A., Ozeki, S., Fiekowsky, E. L., Greenman, G. D., II, & Schroter, D. C. (2017). A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation*, 38(3), 329–347.
- Cousins, J. B. (2007). Process use in theory, research, and practice. *New Directions for Evaluation*, 116(Winter), 1–112.
- Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. Jossey-Bass.
- Crozier, M. (1964). *The bureaucratic phenomenon*. University of Chicago Press.
- Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures: Selected essays* (pp. 3–30). Basic Books.
- Gullickson, A. M. (2020). The whole elephant: Defining evaluation. *Evaluation and Program Planning*, 79(April), 1–9. Article number: 101787.
- Hall, J. N. (2020). The other side of inequality: Using standpoint theories to examine the privilege of the evaluation profession and individual evaluators. *American Journal of Evaluation*, 41(1), 34–53.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge University Press.
- King, J. A., & Alkin, M. C. (2019). The Centrality of use: Theories of evaluation use and influence and thoughts on the first 50 years of use research. *American Journal of Evaluation*, 40(3), 431–458.
- Kirchin, S. (2019). *Thick evaluation*. Oxford University Press. http://book-epub.com/book/thick-evaluation/#book_description
- Kyle, B. G. (n.d.). Thick concepts. *Internet Encyclopedia of Philosophy*. <https://www.iep.utm.edu/thick-co/#SH5a>
- Leeuw, F. L., & Donaldson, S. I. (2015). Theory in evaluation: Reducing confusion and encouraging debate. *Evaluation*, 21(4), 467–480.
- Lemire, S., Whynot, J., & Montague, S. (2019). How we model matters: a manifesto for the next generation of program theorizing. *Canadian Journal of Program Evaluation*, 33(3), 414–433.
- Lieberman, E. (2005). Nested analysis as a mixed-method strategy for comparative research. *American Political Science Review*, 99(2), 435–452.
- Mark, M. M., Donaldson, S. I., & Campbell, B. (Eds.). (2010). *Social psychology and evaluation*. Guilford.
- Marra, M. (2011). Micro, meso and macro dimensions of change: A new agenda for the evaluation of structural policies. In K. Forss, M. Marra, & R. Schwartz (Eds.), *Evaluating the complex: Attribution, contribution and beyond* (pp. 97–122). Transaction.
- Mark, M. M., Donaldson, S. I., & Campbell, B. (Eds.). (2011). *Social psychology and evaluation*. Guilford.
- Minnich, E. K., & Patton, M. Q. (2020). *Thought work: Thinking, action, and the fate of the world*. Rowland & Littlefield.
- Morgan, G. (2006). *Images of organizations*. Sage.
- Mueller, C. W. (2004). Conceptualization, operationalization, and measurement. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. 161–165). Sage.

- Patton, M. Q. (1978). *Utilization-focused evaluation* (1st ed.). Sage.
- Patton, M. Q. (1986). *Utilization-focused evaluation* (2nd ed.). Sage.
- Patton, M. Q. (1997). *Utilization-focused evaluation* (3rd ed.). Sage.
- Patton, M. Q. (1998). Discovering process use. *Evaluation*, 4(2), 225–233.
- Patton, M. Q. (2007). Process use as a usefulism. *New Directions for Evaluation*, 116(Winter), 99–112.
- Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Sage.
- Patton, M. Q. (2011) *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford.
- Patton, M. Q. (2014). What brain sciences reveal about integrating theory and practice. *American Journal of Evaluation*, 35(2), 237–244.
- Patton, M. Q. (2015). *Essentials of utilization-focused evaluation* (4th ed.). Sage.
- Patton, M. Q. (2017). Principles of evaluation: Interpreting Freire. *New Directions for Evaluation*, 155(Fall), 49–78.
- Patton, M. Q. (2018a). Evaluation science. *American Journal of Evaluation*, 39(2), 183–200.
- Patton, M. Q. (2018b). *Facilitating evaluation*. Sage.
- Patton, M. Q. (2020). *Blue marble evaluation: Premises and principles*. Guilford.
- Poth, C., Lamarche, M. K., Yapp, A., Sulla, E., & Chisamore, C. (2014). Towards a definition of evaluation within the Canadian context: Who knew this would be so difficult? *Canadian Journal of Program Evaluation*, 29(1, Spring), 87–103.
- Rogers, E. M. (1962). *Diffusion of innovation*. Free Press.
- Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of innovation*. Free Press.
- Ryle, G. (1971). “The thinking of thoughts: What is ‘Le Penseur’ doing?” In *Collected Papers* (Vol. 2, pp. 480–483). Routledge.
- Schwandt, T. (2001). *Dictionary of qualitative inquiry* (2nd Rev. ed.). Sage.
- Simon, H. (1957). *Administrative behavior*. Macmillan.
- Simon, H. (1978). On how we decide what to do. *Bell Journal of Economics*, 9, 494–507.
- Slatin, C., Galizzi, M., Melillo, K. D., & Mawn, B. (2004). Conducting interdisciplinary research to promote healthy and safe employment in health care: Promises and pitfalls. *Public Health Reports*, 119(1), 60–72. <https://doi.org/10.1177/003335490411900112>
- Stake, R. E. (2004). How far dare an evaluator go toward saving the world? *American Journal of Evaluation*, 25(1), 103–107.
- Thompson, J. D. (1967). *Organizations in action*. McGraw-Hill.
- Thorngate, W. (1976). Possible limits on a science of social behavior (Chapter 5). In L. H. Strickland, F. E. Aboud, & K. J. Gergen (Eds.), *Social psychology in transition*. Plenum Press.
- Torres, J. (2018). *Learning process of a neural network how do artificial neural networks learn? Towards data science blog*. <https://towardsdatascience.com/how-do-artificial-neural-networks-learn-773e46399fc7>
- Vo, A. T., & Archibald, T. (2018). (Eds.), Evaluative thinking. *New Directions for Evaluation*, No. 158.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$.” *The American Statistician*, 73(Suppl. 1), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>
- Wei-Hass, M. (2019, October 28). *Controversial new study pinpoints where all modern humans arose*. National Geographic. <https://www.nationalgeographic.com/science/2019/10/controversial-study-pinpoints-birth-place-modern-humans/>
- Westhorp, G. (2012). Using complexity-consistent theory for evaluating complex systems. *Evaluation*, 18(4), 405–420.
- Westley, F., & Antadze, N. (2010). Making a difference: Strategies for scaling social innovation for greater impact. *The Innovation Journal: The Public Sector Innovation Journal*, 15(2), article 2.
- Westley, F., McGowan, K., & Tjornbo, O. (Eds.). (2017). *The evolution of social innovation: Building resilience through transitions*. Edward Elgar.

- Westley, F., Olsson, P., Folke, C., Homer-Dixon, T., Vredenburg, H., Loorbach, D., Thompson, J., Nilsson, M., Lambin, E., Sendzimir, J., Banerjee, B., Galaz, V., & Van der Leeuw, S. (2011, November). Tipping toward sustainability: Emerging pathways of transformation. *AMBIO: A Journal of the Human Environment*, 40(7), 762–780.
- Westley, F., Tjornbo, O., Olsson, P., Folke, C., Crona, B., Schultz, L., & Orijan Bodin, O. (2013). A theory of transformative agency in linked social-ecological systems. *Ecology and Society*, 18(3), 27. <http://dx.doi.org/10.5751/ES-05072-180327>
- Westley, F., Zimmerman, B., & Patton, M. Q. (2006). *Getting to maybe: How the world is changed*. Penguin Random House.
- Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.
- Williams, M. (2004). Operationism/operationalism. In M. S. Lewis-Beck, A. Bryman, & T. Futing Liao (Eds.), *The Sage encyclopedia of social science research methods*. Sage.
- Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.; 1st English Ed.). MacMillan.