Critical Thinking and Program Evaluation

David R. Narum and Scott K. Pigg, Wisconsin Energy Conservation Corporation

The evaluation of energy efficiency programs requires both critical thinking and technical analysis, yet evaluations that emphasize technical analyses can often overlook basic critical thinking concepts. Essentially, critical thinking forces us to ask whether the conclusions (results) follow logically from the premises (analyses). Critical thinking is a vital component of evaluation and evaluation review as it can help identify areas where the basic validity of evaluation approaches and conclusions may be threatened.

Introduction

An evaluator is trying to determine whether the comparison group he is using represents an adequate baseline against which to gauge the program effect on the treatment group. A t-test on various characteristics of the two groups indicates statistically insignificant differences. Does this mean the baseline is valid?

In estimating the total energy savings for a commercial rebate program, an evaluator chooses the median percent savings. Why might this be wrong?

An evaluator has a sample of residential customers who received furnace rebates and a random sample of non-participant households from the same utility service territory. What's the best approach for estimating the net impact of the rebates on energy savings?

In *Elementary Lessons in Logic* (1957, p. 11), Warren Jevons writes that "Logic may be most briefly defined as the Science of Reasoning. It is more commonly defined, however, as the Science of the Laws of Thought . . . .By a Law of Thought we mean a uniformity or agreement which exists in the modes in which all persons think and reason, so long as they do not make what we call mistakes, or fall into self-contradiction and fallacy . . . there are modes in which all persons do uniformly think and reason, and must think and reason."

Jevons' main point in this passage is that, as there are immutable laws of mathematics, so too are there laws regarding the way we think, the breach of which leads to "self-contradiction and fallacy." Traditionally, logic has been called the "science of the sciences," inasmuch as logical principles and critical thinking must form the basis of all knowledge (for example, geology, biology, sociology, physiology, anthropology, and so on). Critical thinking uses the laws of logic and develops rules that can help us reason and identify false reasoning.

The profession of program evaluation relies on reason, but there can be a tendency to use methods that worked in a previous situation (or that worked for someone else) in new and perhaps quite different evaluation settings. True, this is often a parsimonious way to use resources, but it is also possible to overlook various threats to the validity of certain approaches that derive from the particular qualities of the data we are analyzing. Much of program evaluation relies on the judgment of the evaluator, judgment that involves both logical and creative thinking. Critical thinking (logic) is a way of organizing those judgments as the evaluator progresses toward a conclusion. Further, critical thinking is a way to test conclusions for *validity*, which means that the results can be said to logically follow from the supporting data. (An attachment presents a more formal philosophical discussion of the forms of argumentation and of the concepts of validity and soundness.)

Knowing what tests conclusions must pass before they are considered valid can help evaluators and evaluation readers identify the information gaps and analytical inconsistencies that could pose a threat to the validity of an evaluation conclusion. Often, this threat arises when a fallacy is committed-when the data given to infer a result do not fully substantiate, and may in fact invalidate, that result. An understanding of logical fallacies helps to clarify our insights into the relationship between data and results when we make inferences in observational studies, and can reduce the chance that we might base our conclusions on faulty reasoning.

In this paper we examine ways of using critical thinking to respond to threats to the validity of program evaluations. We recognize that no evaluation is perfect—all evaluations have uncertainty and potential bias associated with them—and see critical thinking as a way to identify large and small sources of uncertainty and bias in evaluation studies. As evaluators need to think critically about their data, methods, and conclusions so that they can produce valid and useful work, so too do evaluation users need to think critically about evaluations to guard against basing decisions and conclusions on flawed evaluation studies.

An important purpose of this paper is to help evaluators conduct valid and practical evaluations that merge process and content—by being able to detect logical inconsistencies before too much time and money is expended. A second purpose is to provide a means for evaluation readers to gauge the extent to which evaluations are, in fact, valid and practical.

From the perspective of resource planning, critical thinking is necessary for program evaluation of DSM efforts, because the impact of DSM on energy usage is less amenable to measurement than supply side resources. While supply side production is easily metered, the effect of demand side efforts must always be inferred relative to what would have happened in the absence of those efforts-and it must be inferred within a dynamic environment of other factors that influence energy usage. It is therefore vital that program evaluation be able to correctly isolate that portion of an observed change that is in fact caused by the DSM effort. But we do *not* mean to suggest that DSM is therefore a more risky endeavor than meeting energy demand through the supply side: planning for either side relies on data and analyses that are fraught with uncertainty. We have simply focused this paper on reducing the risk of inaccurate results from DSM program evaluations.

The paper is structured as follows: 1) a discussion of the basics of critical thinking, including general components, validity and practicality, and a review of some general fallacies that can arise in program evaluation, 2) a closer analysis of the three examples from the beginning, and 3) a brief note on some of the differences between vertical (critical) and lateral (creative) thinking.

Basics of Critical Thinking

The study of critical thinking is wide-ranging and applicable to many different areas of human inquiry. In this section we touch briefly on the more basic elements of critical thinking related to program evaluation. While what we present here can only scratch the surface of the field of critical thinking, we try to emphasize the more essential issues. The first part of this section is a brief overview of the components of critical thinking. Secondly, we discuss the issues of validity and practicality in the context of program evaluation. Lastly, we provide a review of some general fallacies, or breakdowns in critical thinking, that can occur in the process of program evaluation.

The Components of Critical Thinking

The process of critical thinking has been formalized by Toulmin et al. (1979) and broken down into six components: 1) *claim*— the position, conclusion, or result, 2) *grounds*— the foundation for the claim, such as information, experimental observations, common knowledge, statistical data, 3) *warrants*— the justification for moving from the grounds to the claims, such as the applicability and choice of the statistical measure or model, 4) *backing*— is the warrant justified? the general background of warrants (e.g. statistical theory), 5) *modal qualifiers*— the reliability and precision of results, and 6) *rebuttals*— possible bases for invalidating the conclusion.

For example, given general statistical theory (Backing), and its regression analysis component (Warrant), some observed data (Grounds), assuming predetermined levels for significance are met (Modal Qualifiers), support a result (Claim). A possible Rebuttal might be that the data are unrepresentative or nonlinear. Evaluators usually will not concern themselves with Backing, but the other components should be addressed in the analysis. Identification of fallacies is considered under possible Rebuttals to the analysis. Evaluators want to identify these and account for them (to the extent possible) both before and during the analysis, and definitely before the results are presented.

We do not present these components with the expectation that evaluators will explicitly use the terminology (there is enough jargon in this field as it is). But these basic components do exist in any form of argumentation-even those forms applied in program evaluation-and an understanding of them may help evaluators and evaluation readers conceptualize the broader framework in which the overall evaluation process rests.

Validity (and Practicality)

Most program evaluation is conducted in the pursuit of results (claims) that are valid, but that are practical as well (meaning useful). That is, not only should the results be based on valid reasoning, but they should also reflect reality in such a way that they have a practical usability, particularly if we are too reshape that reality based on those results.

As noted above (and discussed in the attachment), a valid argument is one in which the claims can be shown to logically follow from the grounds, as justified by the warrants. In the language of program evaluation, this would be stated as "the results follow from the data, as manipulated by statistical (or other) methods." It is important to remember that the fact that an argument is valid does not also mean it is *true!* Falsified data can still lead to a valid conclusion.

In their book *Quasi-Experimentation,* Cook and Campbell (1979) list four fundamental questions that evaluators should ask in conducting research: 1) is there a relationship between the variables in the study? 2) is the relationship causal? 3) if it is causal, what are the particular cause and effect factors? and 4) are the results generalizable? This is the purpose of program evaluation in a nutshell—to look for a causal relationship between a program intervention and a program effect, and to assert whether the observed effect (if one is observed) can be generalized to the program population, and perhaps to other populations.

Cook and Campbell (p. 37) define the difference between the internal and external validity of an evaluation. *Internal validity* is "the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of a cause." In essence, internal validity refers to the internal consistency of an evaluation—whether the grounds offered lead to and substantiate the claims drawn from the data at hand.

The challenge for internal validity is to correctly separate out causal factors so that one can measure what one is interested in. The biggest threats to an evaluation in this sense are extraneous factors that are not controllable (there are likely to be many such threats in the quasi-experimental setting of most program evaluations, as opposed to a controlled experimental setting). Further, one should note that the absence of a *detected* causal relationship is not the same as the absence of any relationship. In many cases, as we discuss below, the difficulty in positing a relationship is owing to insufficient data or misspecified analytical techniques that do not substantiate a move from the claims to the grounds.

The authors define *external validity* as "the approximate validity with which we infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings and times." In essence, this refers to the ability of evaluators to extrapolate sample results (claims) to a larger population, or to other populations not necessarily associated with the study sample; this commonly involves issues of study group *representativeness* and program changes over time.

In addition to the ever-present concerns with validity, evaluators also strive to make their evaluations practical, or useful (or so we would hope). In his book *Practical Evaluation,* Michael Patton discusses the importance of practicality in evaluation, by which he means that evaluators must have a sense of the purposes to which the evaluation will be put, and the analytical tools and abilities at their disposal. If evaluation resources are constrained, an attempt should be made to do what can be done to meet the goals of the evaluation given those constraints, not necessarily to get as much data as possible to run a desired model. As he writes (p. 19):

> "It's relatively easy to generate a great deal of information with sophisticated evaluations made possible by fairly ample resources. It's also relatively easy to design an extremely simple evaluation with very limited resources, one that generates a certain minimum amount of acceptable information. What is more difficult is to generate a great deal of really useful information with extremely scarce resources. The latter challenge seems also to be the most typical."

Evaluators and evaluation readers are typically concerned with the *significance* of the results. That is, because most evaluations rely on inductive reasoning, or observation of population samples, there is no way to be absolutely certain about a given result. Significance levels, or the range with which we can rely on results (usually expressed through statistical confidence intervals), are the way statistics deals with its basis in inductive reasoning—they are the modal qualifiers that frame the range in which we can accept a given claim.

Bear in mind that the usual significance levels and confidence intervals are predicated on the assumption that the sample is randomly drawn from the population; that is, the only error between the sample result and the (unknown) population result is due to error that could be expected from one random sample to another. Yet evaluations are seldom conducted in such a world, and more often must confront confounding and often hard to detect sources of potential bias. As Cochran (1983, p. 11) notes,

> "[T]he investigator may do well to adopt the attitude that, in general, estimates of the effect of a treatment or program from observational studies are likely to be biased. The number of variables affecting y on which control can be attempted is limited, and the controls may be only partially

effective on these variables. One consequence of this vulnerability to bias is that the results of observational studies are open to dispute. Such disputes, often voluminous, contribute little to understanding. One critic may believe that the failure to adjust for *x4* made the results useless, while the investigator may believe that there is little risk of bias from *x4."*

General Types of Fallacies in Program Evaluation

One role for critical thinking in program evaluation is to form judgments about the magnitude and direction of potential sources of bias, which left unresolved could otherwise lead to fallacious conclusions. Essentially, a fallacy occurs when an invalid argument is mistaken for a valid argument. Fallacies in reasoning are usually accidental, such as neglecting to confirm a key assumption of a model that would invalidate the result. But fallacies can also be deliberately used to mislead.

For example, it would be fallacious to suppress certain data points to produce a *desired* result. It is more likely, however, that the suppression of data is an unconscious response to having prior expectations of what the results should be. If the initial analysis shows results that are comparable to these expectations, no further analysis may be conducted to test the robustness of this answer. But if the initial results are very different from what we expected, the data are tweaked (meaning that certain data points may be omitted or weighted somehow) in the hope of moving the number closer to the expected result. Then ex-post arguments are made to show why that tweaking was justified. There is a difference between looking for a causal relationship one *knows is* there, as opposed to a relationship one *hypothesizes is* there. In the latter case, an evaluator may more readily accept a finding of no causality.

In many cases, fallacies are committed simply because of an omission of caveats. Because all evaluation is uncertain to some degree (that is the nature of inductive reasoning—there is no certainty), a failure to note the potential pitfalls of a particular analysis or chain of reasoning in an evaluation can imply that the results are more certain than they are. While we all wish we could provide greater certainty (as do the clients for whom evaluations are conducted), it is not always possible.

There is an old story about the man who was seen one night crawling around on his knees at the base of a streetlight. A passerby asked him what he was doing. "I'm looking for my keys!" he replied. The passerby then asked, "Did you lose them here?" To which the man said "No, but it's the only place I can see anything. " This example of fallacious reasoning seems silly, but actually occurs frequently in the realm of DSM program evaluation.

For example, if the goal of impact evaluation is an estimate of net lifecycle benefits, we need information on first-year impacts, program costs, measure persistence/ reliability, measure life, utility avoided costs, free ridership, program spillover, and so on. To concentrate exclusively on first-year impacts and program costs (because that is what we have data on) while ignoring others risks producing misleading results. Evaluators and evaluation readers must have a sense of the type and extent of the factors that affect overall program performance. Good evaluations target those factors that are both important and uncertain.

Another typical fallacy is defined generally as "hasty generalization. " This fallacy occurs when an argument is presented in which the premises point toward the conclusion, but in themselves are insufficient to establish that conclusion. In the context of program evaluation, hasty generalization can occur when there are small samples from which to draw a conclusion about a population and a conclusion is made anyway (this is particularly troublesome when the samples are not random). If one goal of a study is to be able to generalize to a larger population (and perhaps to other populations), the sample from which the conclusions are made must be of sufficient size and representativeness to warrant such an extrapolation.

While randomization can alleviate the concern with small samples to some extent, it is not always possible to randomize. Evaluation data collection and analysis is usually performed on a subgroup of the population of interest, and in some cases a census of the population is attempted. The statistical methods leading to the results (and measures of confidence in those results) are usually based on the assumption that the data represents a random sample of the population. But cases inevitably fall out for one reason or the other, which can lead to a biased sample. Further, claiming very precise results for a study that covers a large fraction of a small population of past participants when the results are used to generalize to future participants can impose bias if the characteristics of the program or the program population are changing over time.

Addressing potential problems of hasty generalization is not particularly difficult. Essentially, the critical evaluator and evaluation reader should try to discern potential sources of bias from hasty generalization and estimate: 1) whether they are large enough to substantially alter the conclusions; and 2) the direction in which they might bias the results. If the sample is not random, this should be

documented, and, if possible, generalization to the larger population should be made via adjustments that properly account for differences between the sample and the population. If the program or population has changed over time, the evaluator should both try to measure the impacts as well as identify where the sample data is of most relevance to the existing program (which are often not the same goal).

Back to the Examples

In this section we return to the examples presented at the beginning of the paper. We explore them in the context of the preceding discussion.

Example 1. Is the Baseline Valid?

An evaluator is trying to determine whether the comparison group he is using represents an adequate baseline against which to gauge the program effect on the treatment group. A t-test on various characteristics of the two groups indicates statistically insignificant differences. Does this mean the baseline is valid?

The components of this chain of reasoning are: 1) Claim = a valid baseline, 2) Grounds = treatment and comparison group data, 3) Warrant = t-tests, 4) Modal Qualifiers = significance levels, 5) Rebuttals = Type II error.

If an evaluator wants to know whether a treatment group is different from a comparison group, she may perform a t-test on the difference of the means of some indicator variables (e. g., facility age, heated square footage). What if the test on the difference of the means is statistically insignificant? Can we stop worrying about whether the comparison group can be used as a proxy for the treatment group? No. Here, the potential danger of bias comes if the populations *are* different, and we have applied a test that is designed to throw up a flag only if it is *very certain* they are different. In statistical terms, we have minimized Type I error (erroneously concluding the populations are different when they are not), but are exposed to Type II error (erroneously concluding the populations are not different when they are). A more appropriate analysis would be to assess the probability that the observed difference might be due to chance, and also to assess the *practical* significance of the difference on the results (is it trivial?).

Example 2: Absolute versus Relative Impacts

In estimating the total energy savings for a commercial rebate program, an evaluator chooses the median percent savings. Why might this be wrong?

The components of this chain of reasoning are: 1) Claim = median percent savings as representing the total program, 2) Grounds = the distribution of data, 3) Warrant = measures of central tendency, 4) Modal Qualifiers = standard errors/deviations, 5) Rebuttals = biased estimator.

The problem here is one of choosing the proper measure of central tendency, the warrant for our claim. The median percent savings is a robust measure of the average *relative* impact of a program on energy usage, but it may not be an appropriate statistic for assessing the total savings from the program if there is a large variation in relative and absolute savings across facilities.

Table 1 illustrates this concept with a hypothetical population of 11 facilities for which we (miraculously) also know the program-induced energy savings. The population is characterized by ten facilities with similar savings in the range of 1%-8%, and one huge facility (#11) with 15% savings. The median percent savings mostly represents the

Table 1. Mean and Median Savings Comparison

	Usage	Savings	
Facility	kWh	kWh	%
1	5,000	200	4.0%
2	5,000	150	3.0%
3	7,000	475	6.8%
4	7,000	450	6.4%
5	8,000	100	1.3%
6	8,000	275	3.4%
7	9,000	625	6.9%
8	10,000	800	8.0%
9	12,000	225	1.9%
10	18,000	700	3.9%
11	200,000	30,000	15.0%
Total	289,000	34,000	
Mean	26,273	3,091	5.5%
Median	8,000	450	4.0%

Estimates of total savings:

Using mean absolute: 3,091 * 11 = 34,001 kWh

Using mean percent: 5.5% * 289,000 = 15,895 kWh

Using median percent: 4.0% * 289,000 = 11,560 kWh

relative savings for the small facilities; it underestimates the relative savings in the large facility, which represents almost 90% of the program impacts.

While this example is highly simplified, the problem is very common in commercial and industrial programs, which often have extreme variation in facility and DSM project size. There is a tension between measuring absolute impacts (which exposes the analysis to extreme heterogeneity) and measuring relative impacts (which will underrepresent large projects).

The fallacy that can occur in these situations is known as "composition and division." The fallacy of composition occurs when we assume that the same thing holds for an entire group that holds for its component parts. The fallacy of division occurs when we claim that what holds for the parts of something also holds for the whole. If the goal of an analysis is to assess the *total* savings due to a program, then one should use methods that properly account for the size of the facilities and conservation measures. The evaluator should keep clear the distinction between analyses intended to estimate total program impacts, versus those intended to elucidate relative impacts in the typical facility.

Related Fallacies. Another related type of fallacious reasoning is if relevant data are somehow omitted. The potential problem may be owing in part to subjective researcher bias—that is, data are (perhaps subconsciously) manipulated to be more in line with prior expectations, and data that were excluded are judged to have been anomalous somehow. That is, the "unrepresentative" data points may be omitted to the extent they significantly deviate from what is seen as the norm (be it mean, median, or whatever). Another way an evaluator may inappropriately censor data is if, for example, billing data is excluded when the occupants are on vacation, or if there is turnover in occupancy. If these are normal occurrences in a population, to exclude them from the analysis will introduce bias. Of course, tweaking data by removing data points is a component of sensitivity analysis, which can validate appropriate models as well as identify tenuous or invalid ones.

The most straightforward way to address the threat of inappropriately suppressing data is to fully document all of the available data and the reasons why data were or were not used in the analysis. If influential data points or outliers are excluded, the reasons why they were excluded should be documented. So too should the report discuss the impact of those points on the analysis; that is, if one influential data point drastically affects the regression line, evaluators should discuss that point, whether or not it is included in the final analysis. Sensitivity or error propagation analyses can help to identify which values contribute most to the uncertainty of the final results.

Lastly, one is guilty of the fallacy of "begging the question" if the results are somehow asserted in the premises (the premises being the data manipulation or models used to arrive at the results). For example, an evaluator may remove cases that individually did not show statistically significant savings on the grounds that their savings were "masked" by other influences on energy usage. While it is true (particularly with utility billing data) that other influences can mask the effect of the measures of interest, it is also plausible that the measures that were installed simply were not effective. To study only the cases that show savings begs the question of whether there are demonstrable savings from a program.

Example 3: Measuring Program-Induced Impacts

An evaluator has a sample of residential customers who received furnace rebates and a random sample of non-participant households from the same utility service territory. What's the best approach for estimating the net impact of the rebates on energy savings?

The components of this chain of reasoning are: 1) Claims = net program impacts, 2) Grounds = treatment and comparison group data, 3) Warrant = assumption of comparability of groups based on random sampling, 4) Modal Qualifier = measures of uncertainty, 5) Rebuttal = noncomparability of groups.

The evaluator in this case has a sample of households that received a heating system replacement, and a comparison group of untreated households randomly selected from the same population. A straightforward pre-post analysis is conducted in which the usage change for the control group is subtracted from that of the treatment group to get a net savings estimate. Does this capture the net impact of the program? Do we need to adjust this estimate for the free ridership rate found in the survey data?

The biggest threat to this approach is, obviously, that the nonparticipants are not representative of the participants-the well-known issue of self-selection bias. In addition to the general concerns about the self-selection of participants (which we will largely ignore here), there are some more fundamental threats that require some understanding of the motivations of participants.

Suppose all of the customers who received rebates would have replaced their heating systems at the same time anyway (perhaps because they had already failed), but that the program prompted them to upgrade to a higher efficiency model. In this case, the net impact of the program

is the efficiency gain between the high efficiency models that were installed and the less efficient models that would otherwise have been chosen. The proper baseline would be a group of customers who did replace their heating systems, but at the lower efficiency level. Instead, the comparison group comprises a random sample of customers, a only small proportion of whom probably replaced their heating system during the period of analysis.

It is likely that the "net" impacts we will measure between participants and non-participants are the full savings between the new heating systems and the old ones that failed. These savings will almost certainly be higher than the savings between a new high efficiency furnace and a new less efficient model, which we posit to be the true impact of the program. To claim the former as the net program-induced savings would be fallacious, unless our argument that participants were already in the market for a heating system replacement is invalid.

Of course, the reality is likely to be that some customers were completely motivated by the program, some were motivated to upgrade the efficiency level, and some customers may not have been motivated by the program at all (free riders). The challenge here is to separate the twin factors of the technological savings from replacing heating systems (regardless of what motivated the change) from the motivational aspects that determine how much the program was the cause of the observed savings. The best approach might be to deliberately remove customers who *did* replace their heating system from the comparison group, and thereby obtain a more pure measurement of the full savings between the old heating systems and the new ones. We might then apply other market research means to estimate how much of this observed savings is in fact due to the program.

This example perhaps best illustrates the opportunities for applying a critical thinking approach to evaluation. We would begin by laying out plausible program scenarios, such as a claim that participants would have purchased their new system anyway. Doing this *before* conducting the evaluation would help determine the data that need to be collected to either validate or refute these claims. After the data are collected, one can continue to speculate about possible sources of bias arising from either the data themselves or the methods used to estimate the average impact of the program, and estimate the possible direction and magnitude of the bias.

Not All Thinking Is Critical

Critical thinking is a way to apply logic to program evaluation, to think *vertically* from one step of reasoning to the next to produce valid conclusions. But because program evaluation and statistical analysis are largely inductive processes, there is often more than simple logic at work. Evaluators should always approach thinking in a logical or vertical framework, but along the way must also often think creatively, or *laterally*. Lateral thinking is essentially inductive reasoning. Unlike vertical thinking, which selects options, lateral thinking seeks to generate options. As de Bono (1970, p. 42) writes:

> "Vertical thinking is selection by exclusion. One works within a frame of reference and throws out what is not relevant. With lateral thinking one realizes that a pattern cannot be restructured from within itself but only as the result of some outside influence. So one welcomes outside influences for their provocative action. The more irrelevant such influences are the more chance there is of altering the established pattern. To look only for things that are relevant means perpetuating the current pattern. "

Table 2 shows some of the differences between vertical and lateral thinking.

Lateral thinking is a necessary component of and complementary to vertical thinking. It is a way of problem-solving through the generation of alternatives, through seeing new ideas where old ideas have shown nothing. It can be seen as fitting in with vertical thinking as exploratory data analysis (EDA) fits in with classical statistical analysis. The process of EDA is to look for resistant data patterns (and thus resistant analyses) by examining different displays of data (e. g., stem-leaf, boxplots, smoothing lines), the residuals from analyses, and re-expressions of the data (Velleman and Hoaglin 198 1). The point is that, while logic is important, so too are creativity and imagination.

In Sum

Patton (1982) places great importance on being grounded in the fundamentals of evaluation, which will allow an evaluator to think logically and creatively, to be situation-responsive, and to choose wisely from among various options in seeking to conduct evaluations that are "responsive, useful, accurate, understandable, and practical." To the extent possible, an evaluator should seek to merge the process of an evaluation with its content—that is, to conduct the evaluation in light of the ends to which it will be put.

The goal of critical thinking is to identify and handle major sources of bias and methodological error before the results of a program are presented, thereby reducing the vulnerability of evaluations to disputes over validity. The examples presented earlier are only illustrative of the way

Table 2. Vertical and Lateral Thinking

Vertical	Lateral
• Fixed categories	• No fixed categories
• Follow most likely paths	• Explore least likely paths
• Finite endeavor	• Probabilistic endeavor
• Promises minimum solution	• Increases chance of maximum solution but makes no promises
• Fixed patterns	• Changing patterns

critical thinking may help to forestall potential evaluation problems. We do not see it as necessary that the reasoning components be explicitly followed, but an implicit foundation of sound reasoning can help streamline evaluations and perhaps, by freeing resources unnecessarily spent on paths of bad reasoning, enhance the validity, usability and practicality of evaluations.

Acknowledgments

The authors would like to thank the reviewers and Ralph Prahl for their helpful comments.

Endnote

1. A t-test is a commonly used statistical test to assess whether the mean values of some a variable differ between two groups to a greater degree than might be expected to occur by chance.

References

de Bono, Edward. 1970. *Lateral Thinking: Creativity Step by Step.* Harper and Row, New York.

Cochran, William G. 1983, *Planning and Analysis of Observational Studies,* Wiley and Sons, New York.

Cook Thomas and Donald Campbell. 1979. *Quasi-Experimentation: Design and Analysis for Field Settings* Houghton-Mifflin, Boston.

Flew, Anthony. 1977. *Thinking Straight.* Prometheus, Buffalo .

Fogelin, Robert. 1978. *Understanding Arguments: An Introduction to Informal Logic.* Harcourt Brace Jovanovich, New York.

Jevons, Warren. 1957. *Elementary Lessons in Logic.* MacMillan, London.

Kahane, Howard. 1976. *Logic and Contemporary Rhetoric.* Wadsworth, Belmont.

Patton, Michael. 1982. *Practical Evaluation.* Sage Publications, New York.

Runkle, Gerald. 1978. *Good Thinking: An Introduction to Logic.* Holt Rinehart and Winston, New York.

Toulmin, Stephen, Richard Rieke and Allan Janik. 1979. *An Introduction to Reasoning.* MacMillan, New York.

Velleman, Paul and David Hoaglin. 1981. *Applications, Basics and Computing of Exploratory Data Analysis.* Duxbury Press, Boston.

Williams, Frederick. 1992. *Reasoning With Statistics.* Harcourt Brace Jovanovich, Fort Worth.

Attachment. Deductive and Inductive Reasoning

Deductive Reasoning

Deductive reasoning occurs when a general truth is proved (deduced) from particular or previously known truths. A deductive syllogism takes the form of two premises leading to one conclusion. For example: "All evaluators are human (major premise), Jeff Schlegel is an evaluator (minor premise), *therefore* Jeff Schlegel is human (conclusion). " This (and any) argument has three components, truth, validity, and soundness. *Truth* applies to the premises. In the above example the premises are true, but does that mean the conclusion is true as well? Only if the argument is valid can we say that the conclusion is true, and only when an argument is valid and the premises are true is it a *sound* argument.

To illustrate validity we can posit an invalid syllogism: "All Olympic athletes are human (major premise), Jeff Schlegel is a human (minor premise), *therefore* Jeff Schlegel is an Olympic athlete (conclusion). " In this syllogism, "Olympic athletes" is the *major term,* "Jeff Schlegel" is the *minor term,* and "human" is the *middle term.* The premises are true but the conclusion is not because the argument form is invalid. We can tell if an argument is valid or not by applying the rules of syllogistic inference (Runkle 1978): 1) at least one of the premises must be affirmative; 2) if a premise is negative the conclusion must be negative, and if the conclusion is negative a premise must be negative; 3) the middle term must be distributed at least once; and 4) any term distributed in the conclusion must also be distributed in a premise,

To be distributed means that a term applies to all members of a class of things denoted by the term in question. In the above argument, the middle term "human" is not distributed in the major premise--that is, nothing has been said in the major premise that is true of all humans. The subject class (Olympic athletes) *is* distributed because all of them *are* human, but not all humans are Olympic athletes. Therefore, to conclude that Jeff Schlegel is an Olympic athlete is not valid (nor likely, anyway), so the argument is not sound.

Inductive Reasoning

In inductive reasoning one still reasons from premises to a conclusion, but the difference is that the middle term is usually based on observation rather than on a known fact or truth. Statistics is fundamentally inductive reasoning. That is, unless we have all members of a population accounted for, we need to rely on certain statistical laws (e. g., of randomness, probability) to generalize from the sample to the broader population, Unlike deductive inference, which holds that in a valid argument the conclusion *must* be true if the premises are, in inductive argumentation we can only get to the point of stating that a conclusion is probably true, even if we grant the truth of the premises. The strength of the evidence supporting the premises determines the strength of the conclusion,