SELECTIVE PLANNING OF INTERFACE EVALUATIONS

William C. Mann
USC Information Sciences Institute

1 The Scope of Evaluations

The basic idea behind evaluation is a simple one: An object is produced and then subjected to trials of its performance. Observing the trials reveals things about the character of the object, and reasoning about those observations leads to statements about the "value" of the object, a collection of such statements being an "evaluation." An evaluation thus differs from a description, a critique or an estimate.

For our purposes here, the object is a database system with a natural language interface for users. Ideally, the trials are an instrumented variant of normal usage. The character of the users, their tasks, the data, and so forth are representative of the intended use of the system.

In thinking about evaluations we need to be clear about the intended scope. Is it the whole system that is to be evaluated, or just the natural language interface portion, or possibly both? The decision is crucial for planning the evaluation and understanding the results. As we will see, choice of the whole system as the scope of evaluation leads to very different designs than the choice of the interface module. It is unlikely that an evaluation which is supposed to cover both scopes will cover both well.

2 Different Plans for Different Consumers

We can't expect a single form or method of evaluation to be suitable for all uses. In planning to evaluate (or not to evaluate) it helps a great deal to identify the potential user of the evaluation.

There are some obvious principles:

1. If we can't identify the consumer of the evaluation, don't evaluate.

2. If something other than an evaluation meets the consumer's needs better, plan to use it instead.

Who are the potential consumers? Clearly they are not the same as the sponsors, who have often lost interest by the time an evaluation is timely. Instead, they are:

1. **Organizations that Might Use the System** --- These consumers need a good overview of what the system can do. Their evaluation must be holistic, not an evaluation of a module or of particular techniques. They need informal information, and possibly a formal system evaluation as well.

   However, they may do best with no evaluation at all. Communication theorists point out that there has never been a comprehensive effectiveness study of the telephone. Telephone service is sold without such evaluations.

2. **Public Observers of the Art** --- Scientists and the general public alike have shown a great interest in AI, and a legitimate concern over its social effects. The interest is especially great in natural language processing. However, nearly all of them are like observers of the recent space shuttle: They can understand liftoff, landing and some of the discussions of the heat of reentry, but the critical details are completely out of reach. Rather than carefully controlled evaluations, the public needs competent and honest interpretations of the action.

3. **The Implementers' Egos** --- Human self-acceptance and enjoyment of life are worthwhile goals, even for system designers and implementers. We all have ego needs. The trouble with using evaluations to meet them is that they can give only too little, too late. Praise and encouragement along the way would be not only more timely, but more efficient. Implementers who plan an evaluation as their vindication or grand demonstration will almost surely be frustrated. The evaluation can serve them no better than receiving an academic degree serves a student. If the process of getting it hasn't been enjoyable, the final certification won't help.

4. **The Cultural Imperative** --- There may be no potential consumers of the evaluation at all, but the scientific subculture may require one anyway. We seem to have escaped this one far more successfully than some fields of psychology, but we should still avoid evaluations performed out of social habit. Otherwise we will have something like a school graduation, a big, elaborate, expensive NO-OP.

5. **The Fixers** --- These people, almost inevitably some of the implementers, are interested in tuning up the system to meet the needs of real users. They must move from the implementation environment, driven by expectation and intuition, to a more realistic world in which those expectations are at least vulnerable.

   Such customers cannot be served by the sort of broad holistic performance test that may serve the public or the organization that is about to acquire the system. Instead, they need detailed, specific exercises of the sort that will support a causal model of how the system really functions. The best sort of evaluation will function as a tutor, providing lots of specific, well distributed, detailed information.

6. **The Research and Development Community** --- These are the AI and system development people from outside of the project. They are like the engineers for Ford who test Datsuns on the track. Like the implementers, they need rich detail to support causal models. Simple, holistic evaluations are entirely inadequate.

7. **The Inspector** --- There is another model of how evaluations function. Its premises differ grossly from those used above. In this model, the results of the evaluation, whatever they are, can be discarded because they have nothing to do with the real effects. The effects come from the threat of an evaluation, and they are like the threat of a military inspection. All of the valuable effects are complete before the inspection takes place.

   Of course, in a mature and stable culture, the inspected party learns to know what to expect, and the parties can develop the game to a high state of irrelevance. Perhaps in AI the inspector could still do some good.

Both the implementers and the researchers need a special kind of test, and for the same reason: **to support design.**[1] The value of evaluations for them is in its influence on future design activity.

There are two interesting patterns in the observations above. The first is on the differing needs of "insiders" and "outsiders."

- The "outsiders" (public observers, potential user organizations) need evaluations of the entire system, in relatively simple terms, well supplemented by informal interpretation and demonstration.
- The "insiders," researchers in the same field, fixers and implementers, need complex, detailed evaluations that lead to many separate insights about the system at hand. They are much more ready to cope with such complexity, and the value of their evaluation depends on having it.

These needs are so different, and their characteristics so contradictory, that we should expect that to serve both needs would require two different evaluations.

The second pattern concerns relative benefits. The benefits of evaluations for "insiders" are immediate, tangible and hard to obtain in any other way. They are potentially of great value, especially in directing design.

In contrast, the benefits of evaluations to "outsiders" are tenuous and arguable. The option of performing an evaluation is often dominated by better methods and the option of not evaluating is sometimes attractive. The significance of this contrast is this:

> **SYSTEM EVALUATION BENEFITS PRINCIPALLY THOSE WHO ARE WITHIN THE SYSTEM DEVELOPMENT FIELD: IMPLEMENTERS, RESEARCHERS, SYSTEM DESIGNERS AND OTHER MEMBERS OF THE TECHNICAL COMMUNITY.**[2]

It seems obvious that evaluations should therefore be planned principally for this community.

3 The Key Problem: Generalization

We have already noticed that evaluations can become very complex, with both good and bad effects. The complexity comes from the task: Useful systems are complex, the knowledge they contain is complex, users are complex and natural language is complex. Beyond all that, planning a test from which reliable conclusions can be drawn is itself a complex matter.

In the face of so much complexity, it is hopeless to try to span the full range of the phenomena of interest. One must sample in a many-dimensional space, hoping to focus attention where conclusions are both accessible and significant.

As a result, the outcomes of evaluations tend to be extremely conditional. The most defensible conclusions are the most conditional--they say "This is what happens with these users, these questions, this much system load..." Since those conditions will never cooccur again, such results are rather useless.

The key to doing better is in creating results which can be generalized. Evaluation plans are in tension between the possibility of creating highly credible but insignificant results on one hand and the possibility of creating broad, general results without a credible amount of support on the other.

I know no general solution to the problem of making evaluation results generalizable and significant. We can observe what others have done, even in this book, and proceed in a case by case manner. Focusing our attention on results for design will help.

Design proceeds from causal models of its subject matter. Evaluation results should therefore be interpreted in casual mode. There is a tendency, particularly when statistical results are involved, to avoid causal interpretations. This comes in part from the view that it is part of the nature of statistical models to not support causal interpretations.

Avoiding causal interpretation is formally defensible, but **entirely inappropriate.** If the evaluation is to have effects and value, causal interpretations will be made. They are inevitable in the normal course of successful activity. They must be made, and so these interpretations should be made by those best qualified to do so.

Who should make the first causal interpretation of an evaluation? Not the consumers of the evaluation, but the evaluators themselves. They are in the best position to do so, and the act of stating the interpretation is a kind of check on its plausibility.

By identifying the consumer, focusing on consequences for design, and providing causal interpretations of results, we can create valuable evaluations.

---

[1] Design here, as in most places, consists almost entirely of redesign.

[2] This is not to say that there are not legitimate, important needs among the "outsiders". Someone must select among commercially offered services, procure new computer systems and so forth. Unfortunately, the available evaluation technology does not even remotely approach a methodology for meeting such needs. For example, there is nothing comparable to compiler benchmarking methods for interactive natural language interfaces. It is not that "outsiders" don't have important needs; rather, we are poorly equipped to meet their needs.