Predictive Evaluation

Michael Scriven
School of Behavioral and Organizational Sciences, Claremont Graduate University

There has been extensive discussion of the relation of evaluation to: (i) research; (ii) explanations (a.k.a. theory-driven, logic model, or realistic evaluation); and (iii) recommendations, from each of which it is logically distinct, although there are times when they do and should overlap in practice. Making the distinction is not a mere linguistic issue since it vitally affects practice and the whole process of learning and teaching evaluation, and responsibly providing evaluation services. It is now time to look more carefully at the ghost at the banquet, (iv) prediction.

The essential logical point is that, unlike the preceding three concepts, prediction is ***necessarily*** part of most kinds of evaluation, although to varying degrees. Let's abuse the notion of quantitative analysis by indicating the rough proportions in the following terms, which are provided simply to start the discussion:

Product Evaluation

33% of the value relies on prediction AND it carries a bar (i.e., a minimum level requirement), ***because*** products including software must have significant durability to be worth anything (hence the bar) and assessing durability involves prediction of resistance to probable hazards of abuse and circumstance.

Program Evaluation

About the same ***because*** programs must be sustainable in order to be of any value (unless you are merely doing historical evaluation).

Personnel Evaluation

At least 50% of this is evaluation for selection/promotion/raises/retention/tenure, often called ex ante evaluation, all of which requires a prediction of continued performance at near the previous level or better, a particularly hazardous prediction. Ex post evaluation, which includes evaluation for rewards for past service or achievements, is retrospective only.

Policy Analysis

Probably 75% of this is done as a guide to decisions and hence must be advice about what will happen in the future. The remaining 25% is ex post evaluation, for the record, not prediction-involving.

Proposal Evaluation

100% predictive, unless for teaching purposes only.

Portfolio Evaluation

100% predictive in the case of investment portfolios, whether the investment is time, money, or research effort; 0% in the case of the original artist's portfolio. But note that if the portfolio is being used as a basis for hiring a commercial artist, rather than appraising their work for mounting an exhibition, this immediately becomes personnel and hence predictive evaluation. Your (distant) future probably rests entirely on this kind of evaluation, since that's where your retirement is likely to be stashed.

So, while there are exceptions like performance evaluation, which are mainly ex

post, it's clear prediction is a large part of the most common kinds of evaluation.

So what? Why does this matter? Wasn't that always obvious?

It matters because prediction is a very tricky business, and evaluators are usually not well versed in the literature about it, for example, the weighty and highly relevant clinical versus statistical prediction literature. What we have tended to do is to jump rather too quickly from evidence about past performance to conclusions about future performance. Very understandable, given that past performance is the best single indicator of future performance; but best single indicator isn't the same as best basis for prediction. The best basis for prediction is the collection and combination of all available indicators. To give a simple example: suppose that a candidate for a job as researcher has a great track record in research, much better than that of any other candidate. Surely, assuming that s/he is OK on a check for criminal record or evidence of plagiarism, that's a good basis for prediction? NO! These days, someone is likely to say that we need evidence of collegiality or team working strengths, but that's often irrelevant or mere political correctness (not always, of course). However, those are not the key missing elements; we need to think along very different lines to find those.

The basic issue is, What other evidence improves our prediction accuracy besides past performance? Well, when does good performance ***turn bad***? Theoretically, we should also ask when bad performance turns good, but that can't be shown to happen quickly—there's always a period of good performance needed to show it has occurred, so it becomes a matter of how long a period of good work do we need to establish high quality potential for the future. Note that 'turn bad' here means, 'no longer doing good work in the area in which we need the research.' There are six cases worth comment.

1. The researcher shifts area of interest. The evidence that this is occurring can be found in close study of publication content and site during the previous year or two of work, best estimated by also looking at recent presentations and visiting lectures, if content is available.
2. The researcher burns out, ceases to be productive altogether. The main indicator is a major drop in productivity in the last year or two, only picked up if you look specifically for it, since overall achievement level for age may still be outstanding. (A secondary indicator, for this and some of the other primary indicators of deterioration, might be a drop in attendance/presentations at key professional association meetings.).
3. The researcher runs into an intellectual wall; gets stuck in a dead-end ideational street. Work output plateaus and content is now repetitive; requires detailed expert consideration of content of recent output.
4. The researcher encounters a no-fault traumatic impediment to productivity, for example, diagnosis of serious disease condition, severe family trauma (divorce, loss of child, incarceration of spouse, etc.). Various indicators, mostly involving an inappropriate invasion of privacy, but including full medical report, which is permissible.
5. The researcher acquires a habit that will seriously impede productivity, for example, drugs, alcohol. Best detected in social interaction, but that is not a reliable source; better handled, as by Australian universities, by contract condition that makes any appointment reversible within first year. In the business world, this and are often picked up by arranging social events that include the spouse/partner, which is probably ethically acceptable.
6. An indicator that is often used by non-professionals is age; since the evidence is clear that this indicator, although providing a strong correlation with deterioration in performance, is trend data only, subject to

huge interpersonal variation, this is not acceptable in dealing with the individual case.

This example from the personnel evaluation domain shows that explicit attention to the prediction element in the evaluation task can lead to major error-reduction. Given the multi-million dollar cash equivalent of a single tenure or other high-zoot research appointment, attention to these correction factors is worth far more than the time and trouble involved.

Of course, in the case of the usual kind of investment portfolio evaluation, everyone knows that it's a prediction game. But shift the example only slightly, to the evaluation of a *program* portfolio, of the kind that any foundation (or, for that matter, department or research unit) assembles, and we find that the conventional approach is the simplistic one of looking at the programs as entities to be evaluated in themselves, that is, as they are now, instead of as time-spanning entitities, for which it is important to separate off the contemporary analysis of merit from the predictive claims.

The hidden assumption is nearly always the same; the assumption that things will go on as they have been, or in a way that can be predicted by simple extrapolation. And, as pointed out in the personnel evaluation case above, while that's the best simple bet, it can almost always be improved by: (i) the general procedure of generating a list of confounding circumstances and looking carefully for the presence of each one; and, in cases that justify some extra time, by (ii) checking the research literature both for improvements to the list of confounds and for further guides to improved prediction in cases of this type (An obvious example of this is the clear evidence that whenever possible one should use a regression formula to assist or provide the prediction rather than expert judgment.).