Performance Management as Evaluation

Prof. Jeffrey Smith
Department of Economics
University of Michigan

Workshop on Market Mechanisms in Active Labor Market Policy
October 24, 2008
IAB, Nürnberg

Introduction

Thanks!

English / German

Focus mainly on performance measurement among the trinity of topics for the workshop

Link performance measurement to program evaluation

Provide a strong and perhaps provocative critic to stimulate discussion

Defining performance management

Part of the “reinventing government” movement

Common to both the Clinton and Bush II administrations

Providing a “bottom line” where none exists

Use of short term measures to provide feedback and allocate rewards

James Q. Wilson and *Bureaucracy*: why is the government performing these tasks?

Example of a performance management system

US Job Training Partnership Act

Federal, state and local system

Performance measures largely outcome levels

Regression adjusted performance targets

Budgetary rewards

Other examples of performance management systems

US Workforce Investment Act

US Small Business Association loans

Canada ALMP

Canada CSLP

Motivations for performance management

1. Align agent actions with principal preferences rather than agent preferences

    Ex: serving the “hard-to-serve” versus those with large impacts

2. Increase agent effort

3. “Quick and dirty” evaluation

The literature focuses mainly on the first motivation. Why?

Alternative solutions to principal agent issues

Solutions different for preference alignment and effort

Close monitoring (for effort) or detailed regulations (for choices); but what about local knowledge as in Hayek (1945)?

Hire line workers who share the preferences of the principal in terms of whom to serve and how to serve them and/or effort

Develop a “professional” culture among line workers

These alternative solutions do not address the evaluation issue

Issues from the principal-agent literature

Rewards should relate to agent control over outcomes
Ex: ALMP and unemployment rate; CSLP and fraction with PSE

What is measured is rewarded; this matters when there are many dimensions to program goals

Measures should be closely related to the objects of interest

It should be easier to improve the measure by engaging in the desired activity than by engaging in strategic manipulation

Quick and dirty evaluation

Want to monitor in real time for rapid feedback

Want to monitor in real time for rapid course correction

Low cost: want to spend money on services not on evaluation

Simplicity increases caseworker buy-in

Simplicity increases public “understanding”

Ex: school and hospital “report cards” in the US

Serious evaluation

Estimate the discounted stream of impacts relative to the counterfactual of non-participation (or later participation?)

- Experiments
- Econometric evaluation via e.g. matching

Consider all relevant outcomes (and monetize as required)

Subtract off costs (including excess burden)

Issues with serious evaluation

Costly to produce; requires expensive econometricians etc.

Long term impacts take a long time to construct

Cannot provide rapid feedback

Non-experimental methods difficult for policy-makers

Competing non-experimental estimates with different implications

Cost measures complex due to fixed costs, other funding streams

Performance measures: short-run outcome levels

Almost universally used in performance systems

Ex: employment 13 weeks after termination

Outcomes may not equal impacts

Short run != long run

Ignores costs

Can be manipulated via cream-skimming or timing of termination

Same points hold with counts, plus incentive to serve too many

Before-after differences

Now used in US WIA system

Before-after differences may not equal impacts

Short run != long run

Ignores costs

Can be manipulated via selective enrollment

Misleading positive values due to Ashenfelter’s dip

Impacts and performance measures: conceptual issues

Correlate estimated impacts with observed values of performance measures at the individual level

Estimated impacts from subgroup variation in experiments (a regression of outcomes on X, D and interactions between X and D)

Issue: low variation in impacts among subgroups
Issue: subgroup impacts often imprecisely measured
Issue: impacts may vary on other dimensions
Issue: does not capture effects on effort (performance measurement increases impacts for all groups)

Impacts and performance measures: evidence

National Job Training Partnership Act Study
Barnow (1999), Heckman, Heinrich and Smith (2002)
No consistent relationship between performance measures and experimental impacts

National Job Corps Study
Burghardt and Schochet (2001)
No relationship between center-level impacts and center-level measured performance

Other literature – see the Barnow and Smith survey – finds no consistent effect (with one exception)

Sample sizes and the limits of performance management

Cells defined by site, subgroup and service type may be quite small

Even cells defined by only one or two of these may be small

Some outcomes, such as earnings and wages, have high variances

These facts imply that performance measures at disaggregate may have a large noise component

But, disaggregate measures are most useful for management

May be over-selling what these measures can do; see lit on classroom-level test score averages in the US

Serious impacts in the short run?

MTI = Medium Term Indicators in Canada

- Years of effort (consultant full employment)
- Matching in real time
- Can serious evaluation really be automated?
- Tradeoff between sophistication and automation / clarity

On-going randomization mit Treffer?

Can customer satisfaction measures do the job?

Some measures are like participant “self-evaluations”

Widely used in evaluations; given top billing when the econometric estimates are weak

Included in some performance management systems e.g. US WIA

Customers potentially include clients and firms

Customers have information that evaluators do not

Good optics and nominal satisfaction illusion

Example questions

Did the program help you get a job?

Would you recommend the program to your best friend?

How satisfied were you with the services you received from the program?

Evidence on customer satisfaction measures in ALMP

There is almost no evidence linking questions to impacts

There is no evidence comparing alternative questions

Are individuals good evaluators?

More precisely, can individuals construct the required counterfactual?

The story of the ineffective study intervention

Econometric evidence: where it comes from

Key issue: obtaining person-specific impacts

Smith, Whalley and Wilcox (2006) and related work with Tanya Byker and Sebastian Calonico

Compare person-specific impacts based on sub-group variation in experimental impacts to self-evaluation measures

Data from US JTPA Study, Connecticut “Jobs First” and National Supported Work Demonstration

Parallel analysis uses person-specific impacts from quantile treatment effects framework under rank preservation

Econometric evidence: findings

JTPA
No correlation between impacts and self-evaluations
Self-evaluations predicted by outcomes, changes and sites

Connecticut Jobs First (work in progress)
Some evidence of a weak positive correlation

National Supported Work (work in progress)
Some evidence of a weak positive correlation

Summary: existing customer satisfaction questions are not a good proxy for impacts

Alternative customer satisfaction measures

Existing evidence: bad idea or bad questions?

Would questions that ask directly about counterfactuals do better?

Ex: If you have not participated in the program, what is the probability that you would be employed today?

Recent survey methods work on how to ask about probabilities would help here

What are customer satisfaction measures good for?

Ask about things the customer directly experiences

Ex: friendly staff, extra visits, waiting time, mistakes

Do not ask about things that require a counterfactual (either implicitly or explicitly)

Use customer service measures to make relative comparisons and so avoid nominal satisfaction illusion

Think about regression adjustment for characteristics of eligible population and local economic environment?

Concluding thoughts

What we do not know:

What problem performance management should solve?
Are there alternatives that would do a better job?
What are the effects of current performance systems?
How to stop governments from misleading the public?

What we do know:

Performance measures a very poor proxy for impacts (schlecht!)
Quick and dirty evaluation may be worse than nothing
Customer satisfaction methods not (now) a solution