FY22 Budget Training

using and generating evidence

David_Yokum@brown.edu

August 27, 2020
Rhode Island

Director, The Policy Lab
thepolicylab.brown.edu
learning objectives

• How do you assess the existing evidence base?
  – Methods 101 crash course
  – The Evidence Scale

• When and how might you invest in generating new evidence?

  • Recall other resources:
    – Budget Instructions
    – Decision Package Template
    – Examples (good and bad)
    – Website Trainer
    – Office Hours
what evidence do you need?
3. Opportunity Statement

In this section, clearly explain the problem that exists today and the opportunity that your request presents to capitalize on. The best opportunity statements thoroughly explain, with as much detail as possible: (1) where we are today; (2) where we want to be in the future; and (3) why there is the gap between where we are and where we want to be. The best opportunity statements also quantify key variables wherever possible.

4. Proposed Intervention & Theory of Change

Provide a detailed description of the initiative you are proposing to respond to the above-described problem/capitalize on the opportunity. Your narrative here should clearly describe how your intervention, if funded, could close the gap described above and achieve the desired future state.
types of metrics

<table>
<thead>
<tr>
<th>inputs</th>
<th>outputs</th>
<th>outcomes</th>
<th>impacts</th>
</tr>
</thead>
</table>

## Types of Metrics

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Outputs</th>
<th>Outcomes</th>
<th>Impacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff, facilities, materials, etc.</td>
<td># people served</td>
<td># employed</td>
<td>Increase in employment or wages CAUSED BY the program</td>
</tr>
<tr>
<td></td>
<td># job apps submitted</td>
<td>average wages</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## aims of evaluation

<table>
<thead>
<tr>
<th>inputs</th>
<th>outputs</th>
<th>outcomes</th>
<th>impacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>staff</td>
<td># people served</td>
<td># employed average wages</td>
<td>Increase in employment or wages CAUSED BY the program</td>
</tr>
<tr>
<td>facilities</td>
<td># job apps submitted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>materials</td>
<td>etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**implementation**

**outcome**

**impact**
impact / causal claims

unemployment

jobs program

time
Causal Claims

Jobs program

Unemployment

Time
need a counterfactual

- no program
- jobs program
who benefited the most from the jobs program?

![Bar chart showing employment percentages for different groups.](chart.png)

- **Youth**: 63%
- **Adult men**: 57%
- **Adult women**: 49%

Aack, I can’t find the publication citation for this data!
who benefited the most from the jobs program?

<table>
<thead>
<tr>
<th></th>
<th>Program group</th>
<th>Control group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Youth</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>Adult men</td>
<td>57</td>
<td>53</td>
</tr>
<tr>
<td>Adult women</td>
<td>49</td>
<td>41</td>
</tr>
</tbody>
</table>
finding the evidence you need
6. Evidence Scale Ranking

Please rank the proposed initiative’s current level of evidentiary support on a scale from 0-5, based on the RI Evidence Scale, with one being the least evidentiary support and five being the most evidentiary support. You can use tools like the Pew Results First Clearinghouse and the Social Programs That Work database to determine whether the initiative you are proposing has been rigorously evaluated in other jurisdictions. The Office of Management & Budget understands that the majority of agency requests will likely not be in the top evidence tiers at the point of submittal, and you should certainly feel free to submit requests that are “theory-based” rather than evidence based. Please note that “theory-based” submissions should include a robust and compelling measurement and evaluation plan in the Performance Measurement section.

<table>
<thead>
<tr>
<th>Evidence of Insufficient Impact or Unintended Effects</th>
<th>Theory-Based</th>
<th>Promising</th>
<th>Proven Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
7. **Description of Evidence Base**

Describe the justification for your evidence scale ranking. What evidence exists that makes you think that the proposed initiative will work? Where is there uncertainty of effectiveness? It is helpful to include citations, links, or attachments of relevant evidence source(s).
evidence clearinghouses

https://www.youtube.com/watch?v=AFuOslr2cSw&feature=youtu.be
11 Tips for Doing Desk Research
WARNING: study quality

- Too few people were studied
- The people studied are not representative of the population you care about
- Causal claims made, but no counterfactual
- Effect sizes are not talked about clearly
- Fishy handling of the data
WARNING: study quality

- Too few people were studied
- The people studied are not representative of the population you care about
- Causal claims made, but no counterfactual
- Effect sizes are not talked about clearly
- Fishy handling of the data
• File drawer problem
• \( P \)-hacking

See generally, Lindsay, Simons, & Lilienfeld, “Research Preregistration 101,” APS Observer (December 2016).

Science Isn’t Broken
It’s just a hell of a lot harder than we give it credit for
pre-analysis plans help avoid fishiness

See https://osf.io/yjyng/ for the publically pre-registered analysis plan. See also https://osf.io/q6c45/ for PVD Talks example.
methods crash course
9. **What methodologies do you currently use for program evaluation?** Check all that apply.

- [ ] No evaluation is done
- [ ] Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
- [ ] Measurement of how many people use the service or initiative
- [ ] Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
- [ ] Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
- [ ] Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
- [ ] Measurement of causal impact with a randomized control trial (RCT)
- [ ] Other:
9. What methodologies **do you currently use** for program evaluation? Check all that apply.

- [ ] No evaluation is done
- [ ] Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
- [ ] Measurement of how many people use the service or initiative
- [ ] Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
- [ ] Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
- [ ] Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
- [ ] Measurement of causal impact with a randomized control trial (RCT)
- [ ] Other:
9. **What methodologies do you currently use for program evaluation?** Check all that apply.

- [ ] No evaluation is done
- [ ] Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
- [ ] Measurement of how many people use the service or initiative
- [ ] Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
- [ ] Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
- [ ] Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
- [ ] Measurement of causal impact with a randomized control trial (RCT)
- [ ] Other:
9. What methodologies do you currently use for program evaluation? Check all that apply.

- No evaluation is done
- Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
- Measurement of how many people use the service or initiative
- Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
- Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
- Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
- Measurement of causal impact with a randomized control trial (RCT)
- Other:
9. What methodologies do you currently use for program evaluation? Check all that apply.

- No evaluation is done
- Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
- Measurement of how many people use the service or initiative
- Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
- Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
- Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
- Measurement of causal impact with a randomized control trial (RCT)

Other:
1. People may answer in ways they think you want, rather than what they really think.
2. People can’t report unconscious causes of behavior.
3. Memory is faulty.
4. Often not a sufficiently representative sample.

wait, can’t folk just self-report impact?

See e.g. GSA Office of Evaluation Sciences, “Reducing Self-reporting Errors by IFF Form Users.”
**Figure 1**
*Trends in Self-Reported Happiness, 1971–1973*

*Note.* Estimates are derived from sample surveys of noninstitutionalized population of the continental United States, aged 18 and over. Error bars demark ±1 standard error around sample estimate. NORC = National Opinion Research Center; SRC = Survey Research Center. Questions were "Taken all together, how would you say things are these days—would you say that you are very happy, pretty happy, or not too happy?" (NORC) and "Taking all things together, how would you say things are these days—would you say you're very happy, pretty happy, or not too happy these days?" (SRC). From "Why Do Surveys Disagree? Some Preliminary Hypotheses and Some Disagreeable Examples" (p. 166) by C. F. Turner. 1984, in C. F. Turner and E. Martin, *Surveying Subjective Phenomena*, New York: Russell Sage Foundation. Copyright 1984 by the Russell Sage Foundation. Reprinted by permission.

*family context*

*work context*
causal methodologies

1. randomized controlled trials (RCTs)
2. “natural” experiments
3. pre-post comparisons
4. multiple regression / matching
5. instrumental variables
6. regression discontinuity

See also, “10 Strategies for figuring out if X caused Y,” EGAP Methods Guide.
causal methodologies

1. randomized controlled trials (RCTs)
2. “natural” experiments
3. pre-post comparisons
4. multiple regression / matching
5. instrumental variables
6. regression discontinuity

See also, “10 Strategies for figuring out if X caused Y,” EGAP Methods Guide.
RCTs

Population is split into 2 groups by random lot

Outcomes for both groups are measured

= looking for work

= found work
body-worn camera study

bwc.thelab.dc.gov
causal methodologies

1. randomized controlled trials (RCTs)
2. “natural” experiments
3. pre-post comparisons
4. multiple regression / matching
5. instrumental variables
6. regression discontinuity

See also, “10 Strategies for figuring out if X caused Y,” EGAP Methods Guide.
1. randomized controlled trials (RCTs)
2. “natural” experiments
3. **pre-post comparisons**
4. multiple regression / matching
5. instrumental variables
6. regression discontinuity

See also, “[10 Strategies for figuring out if X caused Y](#),” EGAP Methods Guide.
causal methodologies

1. randomized controlled trials (RCTs)
2. “natural” experiments
3. pre-post comparisons
4. multiple regression / matching
5. instrumental variables
6. regression discontinuity

See also, “10 Strategies for figuring out if X caused Y,” EGAP Methods Guide.
The regression was significant, ($R^2 = .05$), $F(9, 1775) = 12.96, p < .001$. But the effect of language condition was marginal and had a negative rather than the expected positive slope, $\beta = -.039$, heteroskedasticity-consistent (HC) SE = -.23, $p = .088$. 

### Model Fit:

<table>
<thead>
<tr>
<th>R-sq</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0545</td>
<td>12.9651</td>
<td>9.0000</td>
<td>1775.0000</td>
<td>.0000</td>
</tr>
</tbody>
</table>

### Heteroskedasticity-Consistent Regression Results

| Coeff  | SE(HC)  | t  | P>|t| |
|--------|---------|----|-----|
| Constant | .4749  | .0374 | 12.7038 | .0000 |
| language | -.0390 | .0229 | -1.7050 | .0884 |
| saw_scen | -.2057 | .0476 | -4.3266 | .0000 |
| saw_sc_1 | -.0421 | .0495 | -8.503 | .3953 |
| saw_sc_2 | .0309  | .0507 | 6.093  | .5424 |
| saw_sc_3 | .0243  | .0503 | .4829 | .6292 |
| saw_sc_4 | -.0159 | .0500 | -3.190 | .7498 |
| saw_sc_5 | -.1196 | .0484 | -2.4727 | .0135 |
| saw_sc_6 | -.0122 | .0498 | -2.457 | .8860 |
| saw_sc_7 | .2393  | .0477 | 5.0122 | .0000 |

### Covariance Matrix of Parameter Estimates

<table>
<thead>
<tr>
<th>Constant</th>
<th>language</th>
<th>saw_scen</th>
<th>saw_sc_1</th>
<th>saw_sc_2</th>
<th>saw_sc_3</th>
<th>saw_sc_4</th>
<th>saw_sc_5</th>
<th>saw_sc_6</th>
<th>saw_sc_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0014</td>
<td>-.0003</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
<td>-.0013</td>
</tr>
<tr>
<td>-.0003</td>
<td>.0005</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
<td>.0000</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0023</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0012</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
<tr>
<td>-.0013</td>
<td>.0000</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
<td>.0013</td>
</tr>
</tbody>
</table>

------ END MATRIX ------
Regression

\[ \hat{y}_i = a + b_1 x_1 \]

- Weight = 115 + 8.6(Height_Inches)
  - all else equal, each additional inch of height predicts an additional 8.6 pounds of weight
Regression

Scatterplot of weight vs height

weight

height
Regression

the basics

 Scatterplot of weight vs height

\[ \hat{y}_i = a + b_{yx} x_i \]
Regression

• An interactive app, to get an intuitive feel for the “best fit” line:
  – http://www.shodor.org/interactivate/activities/Regression/
Regression

visualizing one predictor

\[ \hat{y}_i = a + b_{yx}x_i \]
Regression

visualizing two predictors

Image from: http://geography.uoregon.edu/bartlein/courses/geog495/lectures/lec13.htm
Regression visualizing three plus predictors
Regression

two tips

1. **Look** at the data.

2. Ask yourself whether **all** the relevant predictors are included in the model.
the evidence scale
an evidence continuum

6. Evidence Scale Ranking

Please rank the proposed initiative’s current level of evidentiary support on a scale from 0-5, based on the [RI Evidence Scale](https://www.northcentralscience.org/policy-and-practice/research-integrity/ri-evidence-scale), with one being the least evidentiary support and five being the most evidentiary support. You can use tools like the [Pew Results First Clearinghouse](https://www.pewtrusts.org/en) and the [Social Programs That Work](https://sww.ssa.gov/OPP/) database to determine whether the initiative you are proposing has been rigorously evaluated in other jurisdictions. The Office of Management & Budget understands that the majority of agency requests will likely not be in the top evidence tiers at the point of submittal, and you should certainly feel free to submit requests that are “theory-based” rather than evidence based. Please note that “theory-based” submissions should include a robust and compelling measurement and evaluation plan in the Performance Measurement section.

<table>
<thead>
<tr>
<th>Evidence of Insufficient Impact or Unintended Effects</th>
<th>Theory-Based</th>
<th>Promising</th>
<th>Proven Effective</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Yokum
when and how should you generate more data?
what don’t you know, that matters?

<table>
<thead>
<tr>
<th>inputs</th>
<th>outputs</th>
<th>outcomes</th>
<th>impacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>staff</td>
<td># people served</td>
<td># employed</td>
<td>increase employment or wages CAUSED BY the program</td>
</tr>
<tr>
<td>facilities</td>
<td># job apps submitted</td>
<td>average wages</td>
<td></td>
</tr>
<tr>
<td>materials etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

implementation

outcome

impact
Forward Looking Opportunities to Develop an EvidenceBase

14. What methodologies will you use for program evaluation? Check all that apply.
   - [ ] No evaluation is planned
   - [ ] Measurement of the resources (e.g. staff, material expenses) required to deploy the service or initiative
   - [ ] Measurement of how many people use the service or initiative
   - [ ] Focus groups, surveys, or other qualitative methods that ask people about their experience with the service/initiative
   - [ ] Measurement of outcomes at a point in time or over time (e.g. monthly or quarterly reports of student test scores, average wages, crime reports, park admissions, medical claims, etc.)
   - [ ] Measurement of causal impact with econometric methods (e.g. regression discontinuity, multiple regression, matched controls, instrumental variables)
   - [ ] Measurement of causal impact with a randomized control trial (RCT)
   - [ ] Other: 


what effect size(s)?
Opportunity:
low-cost RCTs / evaluation

11. Which of the following best describe the type(s) of data that you currently use for evaluation? Check all that apply.

☐ No data is collected
☐ Qualitative data is collected (e.g. participant demographic information)
☐ Quantitative data is collected (e.g. number of participants)
☐ Aggregate-level data is collected (e.g. % of students who qualify for free/reduced lunch)
☐ Individual-level data is collected (e.g. household income amount for an individual student)

12. Which of the following best describes the source(s) of data that you currently use for evaluation? Check all that apply.

☐ No data is collected
☐ We use existing data that has been collected by another state agency, the federal government, a private entity, or another source
☐ We use existing data that has been collected by our agency for a different purpose
☐ We collect initiative-specific data
basic tips to assess data quality

1. Look at your raw data
   a. What’s missing?
   b. Do the entries make sense?
   c. Do formats change over time?
2. Create a data dictionary
3. Do field work to understand how data were generated.
   a. Pro Tip: Map it to your process map
4. Look at visualizations of your data
5. Compute basic descriptive statistics
partnership opportunities

18. Have you identified research partners to help evaluate the initiative? If so, please describe who. If not, what type of research partnership, if any, would be helpful?

[Blank space]

19. Do you want to speak with a methods expert for a consultation on what evaluation methods might be best for learning about and optimizing the performance of your initiative?

- No, we do not need a methods expert
- No, we have already engaged a methods expert
- Maybe, it might be useful, we’re open to it
- Yes, that would be helpful
resources

- Budget Instructions
- Decision Package Template
- Examples (good and bad)
- Website Trainer
- **Office Hours**

- *Your feedback, much appreciated!*
discussion
connect with The Policy Lab

• Explore at thepolicylab.brown.edu, and sign up for our listserv for updates, events, and ways to collaborate.

• Check out the podcast at thirtythousandleagues.com.