Using the Program Evaluation Standards in Metaevaluation: Potential & Pitfalls

Lori A. Wingate

Annual Conference of the American Evaluation Association

November 11, 2010 | San Antonio

Can You Judge an Evaluation by its Report?
Overview

1. Context
2. Study design
3. Results
4. Implications: Pitfalls & Potential
**Context**

**NSF’s Advanced Technological Education program:**

“With an emphasis on two-year colleges, the ATE program focuses on the education of technicians for the high-technology fields that drive our nation’s economy.”

www.nsf.gov/ate

---

**EvaluATE’s Mission:**

- strengthen the ATE program’s evaluation knowledge base
- expand the use of exemplary evaluation practices
- support the continuous improvement of technician education throughout the nation.
Purposes of the Study:

1. Needs Assessment
   - Identify the strengths and weaknesses of ATE evaluations
   - 1 of 3 metaevaluation studies

2. Research
   - Investigate interrater agreement in metaevaluation
The Typical Metaevaluation

- Number of Metaevaluators: 1
- Criteria: Program Evaluation Standards
- Method: Document Review

Based on my review of 54 published metaevaluations

Raters & Reports

Students Practitioners Scholars

Projects:
- Project 1
- Project 2
- Project 3
- Project 4
- Project 5
- Project 6
- Project 7
- Project 8
- Project 9
- Project 10
Program Evaluation Standards (2nd edition)

- Utility: 7 standards
- Feasibility: 3 standards
- Accuracy: 12 standards
- Propriety: 8 standards

Metaevaluation Instrument

<table>
<thead>
<tr>
<th>Standard</th>
<th>1 Not addressed</th>
<th>2 Minimally addressed</th>
<th>3 Partially addressed</th>
<th>4 Mostly addressed</th>
<th>5 Fully addressed</th>
<th>Insufficient Information</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 Practical Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 Political Viability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3 Cost Effectiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Standard A10 Justified Conclusions

The conclusions reached in an evaluation should be explicitly justified, so that stakeholders can assess them.

Example

Measures

Intraclass correlation coefficient (ICC)

Proportion of variance in observed scores that is due to differences in true scores
Measures

Percent agreement

Exact agreement

Adjacent agreement

Results
Ratings

- F1 Practical Procedures
- U5 Report Clarity
- U3 Information Scope and Selection
- A4 Defensible Information Sources
- U6 Report Timeliness and Dissemination
- A1 Program Documentation
- P1 Service Orientation
- A3 Described Purposes and Procedures
- U1 Stakeholder Identification
- U7 Evaluation Impact
- P5 Complete and Fair Assessment
- A2 Context Analysis
- P4 Human Interactions
- P6 Disclosure of Findings
- A10 Justified Conclusions
- A11 Impartial Reporting
- A9 Analysis of Qualitative Information
- A8 Analysis of Quantitative Information
- A7 Systematic Information
- U4 Values Identification
- U2 Evaluator Credibility
- A5 Valid Information
- F3 Cost Effectiveness
- A6 Reliable Information
- F2 Political Viability
- P3 Rights of Human Subjects
- P8 Fiscal Responsibility
- P2 Formal Agreements
- P7 Conflict of Interest

Not Addressed

1 1.5 2 2.5 3 3.5 4 4.5 5 Fully Addressed
Interrater Agreement

Average ICC by Domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Single Rater</th>
<th>Group of Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>.30, .46</td>
<td></td>
</tr>
<tr>
<td>Utility</td>
<td>.17, .38</td>
<td></td>
</tr>
<tr>
<td>Propriety</td>
<td>.14, .31</td>
<td></td>
</tr>
<tr>
<td>Feasibility</td>
<td>.10, .26</td>
<td></td>
</tr>
</tbody>
</table>

Interrater Agreement

Percent Agreement

<table>
<thead>
<tr>
<th>Domain</th>
<th>Exact agreement</th>
<th>Adjacent agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>29%, 61%</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>24%, 60%</td>
<td></td>
</tr>
<tr>
<td>Propriety</td>
<td>22%, 50%</td>
<td></td>
</tr>
<tr>
<td>Feasibility</td>
<td>20%, 49%</td>
<td></td>
</tr>
</tbody>
</table>
Interrater Agreement

Adjacent Agreement by Group

Accuracy

- Students: 71%
- Practitioners: 67%
- Scholars: 59%

Feasibility

- Students: 53%
- Practitioners: 50%
- Scholars: 49%

Propriety

- Students: 53%
- Practitioners: 40%
- Scholars: 38%

Utility

- Students: 57%
- Practitioners: 59%
- Scholars: 60%

Metaevaluation Instrument

<table>
<thead>
<tr>
<th>Standard</th>
<th>1 Not addressed</th>
<th>2 Minimally addressed</th>
<th>3 Partially addressed</th>
<th>4 Mostly addressed</th>
<th>5 Fully addressed</th>
<th>Insufficient Information</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1 Practical Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 Political Viability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3 Cost Effectiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“It is impossible to render a reasonably well-informed judgment based on available information.”
Standards with the Highest Frequency of “Insufficient Information” Ratings

- A11 Impartial Reporting
- P6 Disclosure of Findings
- F3 Cost Effectiveness
- F2 Political Viability
- P4 Human Interactions
- U2 Evaluator Credibility
- P3 Rights of Human Subjects
- P7 Conflict of Interest
- P2 Formal Agreements
- P8 Fiscal Responsibility
- U6 Report Timeliness and Dissemination

Implications
Pitfalls

When the Standards are applied to reports:
- Poor interrater agreement
- Uncertainty about how adherence to the Standards should manifest
- Ambiguity of the rating term “Addressed”
- High degree of inference required

Potential

- The Program Evaluation Standards are widely known and accepted
- Metaevaluation can serve research and quality assurance purposes
Potential

Recommendations:
- Use the Standards as a springboard to developing specific metaevaluation tools and criteria
- Tailor criteria to context
- Establish reliability
- Avoid single-rater designs

Evaluator Accountability Standards (3)
- E1 Evaluation Documentation
  Evaluations should fully document their negotiated purposes and implemented designs, procedures, data, and outcomes.

www.jcsee.org
Thank You!

<table>
<thead>
<tr>
<th>Marvin Alkin</th>
<th>Tala Davidson</th>
<th>Mhora Newsom-Stewart</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mohammed Alyami</td>
<td>Jonathan Engelman</td>
<td>Lindsay Noakes</td>
</tr>
<tr>
<td>Ezechukwu Awgu</td>
<td>Jan Fields</td>
<td>Michael Patton</td>
</tr>
<tr>
<td>Lee Balcom</td>
<td>Jody Fitzpatrick</td>
<td>Jon Pedersen</td>
</tr>
<tr>
<td>Stacy Berkshire</td>
<td>Amy Germuth</td>
<td>Peter Saflund</td>
</tr>
<tr>
<td>Valerie Caracelli</td>
<td>Neal Grandgenett</td>
<td>James Sanders</td>
</tr>
<tr>
<td>Christina Christie</td>
<td>Carl Hanssen</td>
<td>Jean Sando</td>
</tr>
<tr>
<td>Jan Corbin</td>
<td>John Hattie</td>
<td>Faye Shafloot</td>
</tr>
<tr>
<td>Elaine Craft</td>
<td>Rodney Hopson</td>
<td>Sanjeev Sridharan</td>
</tr>
<tr>
<td>Lois-ellin Datta</td>
<td>Robbie McCarty</td>
<td>Carl Westine</td>
</tr>
</tbody>
</table>

Drs. Chris Coryn, Leslie Cooksy, & Arlen Gullickson

This material is based upon work supported by the National Science Foundation under Grant No. 0802245. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.