Evaluation of Educational Innovations for Students and Residents: Principles to Enhance Medical Education Research

Judy A. Shea, Ph.D.
Associate Dean Medical Education Research
Professor of Medicine
Division of General Internal Medicine
Penn Medicine
Overview

Mission of medical education – produce clinically competent physicians

Many players:
- Learners
- Educators
- Administrators
- Evaluators

Value similar endpoints
- Have different foci, priorities, agendas
Objectives

- To present a framework for thinking about medical education research
- To review four evaluation/research principles to help enhance the quality of medical education research: study design, data collection methods, data quantity, data interpretation
# Medical Education Research Framework

<table>
<thead>
<tr>
<th>Unit of Analyses</th>
<th>Focus of Research</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process</td>
</tr>
<tr>
<td>Person (student, resident, faculty)</td>
<td>x</td>
</tr>
<tr>
<td>Program (course, residency)</td>
<td>x</td>
</tr>
</tbody>
</table>
General research principles

Ask a good question
  » Literature/theory based

Design a good study

Select optimal data collection methods

Gather enough data
  » Psychometrics/reproducibility/power

Interpret the data appropriately
General research principles

Ask a good question
  » Literature/theory based

Design a good study

Select optimal data collection methods

Gather enough data
  » Psychometrics/statistics/reproducibility

Interpret the data appropriately
Principle #1: Design a Good Study
“Planning is Better Than Not Planning”

What does this imply?
- have a [good] research question
- have thought about how to answer it
- have measurable processes/outcomes

Keywords: a priori/planning/prospective
Designs

Correlational/cohort studies
   observational
   descriptive

Experimental studies
   intervention
   post-intervention observation
Non-experimental

Single group - post test only
G1: ---------------I---------------O

Single group – pre-post test
G1: O---------------I---------------O

Extremely common in medical education
Multiples threats to validity
Quasi experimental

Have a nonrandom control group
G1:________UC________O
G2:________I__________O

Probably underutilized in medical education
Experimental

Subjects are randomly assigned to treatment groups

Some level(s) of blinding

RG1: _______ UC _________ O
RG2: _______ I __________ O
Example for Principle #1

What if we taught students to give oral presentations and had them practice? Would they be better with practice than without?

<table>
<thead>
<tr>
<th>Unit of Analyses</th>
<th>Focus of Research</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process</td>
</tr>
<tr>
<td>Person</td>
<td>x</td>
</tr>
<tr>
<td>Program</td>
<td>x</td>
</tr>
</tbody>
</table>
RATIONALE: Communication skills are an essential component of professional competence. Little research has focused on medical student’s oral case presentation (OCP) skills.

OBJECTIVE: To develop OCP encounter cards to enable ratings of students’ OCP skills and examine the impact of an intervention on students’ presentation skills.
**DESIGN**: Medicine clerkship has 4 12-week blocks; each has a 6-week inpatient session; randomized based on whether begin with inpatient session within block

Block 1: [Diagram of block 1]

Block 2: [Diagram of block 2]

Block 3: [Diagram of block 3]

Block 4: [Diagram of block 4]

All (approximately 20) students within session get same treatment
TREATMENT: Everyone gets curriculum.

Treatment Group: complete 9 OCP cards:
- 3 from inpatient attendings
- 3 from inpatient residents
- 3 from outpatient attendings

OCP Cards:
- 9-point scale
- Rate 7 content dimensions (e.g., Hx of present illness, assessment, plan), organization, overall speaking ability
OUTCOME: Make oral presentation to faculty
  » Randomized to type of case
  » Examiner blinded to treatment status
ANALYSES:
Completion rate, mean OCP card ratings, completion times, and satisfaction ratings
Differences in ratings for different types of evaluators
Reliability, homogeneity
Treatment effect: did the intervention work
KEY FINDINGS:

Performance on OCP cards correlated with:
  - inpatient clinical evaluations ($r = .58$)
  - ratings of presentation skills ($r = .43$)
  - final grades ($r = .40$)

BUT

Final summative OCP performance was worse for intervention than control group
Principle #1: Design a Good Study
“Planning is Better Than Not Planning”

TAKE HOME MESSAGE: With some planning it is possible to implement an experimental design.
Principle #2: Select Optimal Data Collection Methods “Do the Right Thing”

What does this mean?

There are lots of ways to collect data
Quantitative/qualitative debate
How am I going to answer my question?
ask people questions
ask people about other people
watch people
Example for Principle #2:

What if I do a study on mentoring of residents? Everyone says it is important but what do we really know about it?

Medical Education Research Framework

<table>
<thead>
<tr>
<th>Unit of Analyses</th>
<th>Focus of Research</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process</td>
</tr>
<tr>
<td>Person</td>
<td>x</td>
</tr>
<tr>
<td>Program</td>
<td>x</td>
</tr>
</tbody>
</table>
RATIONALE: Mentoring is important. Most of what we know about mentoring is for faculty. The benefits might be expected to extend to residents in need of career and personal counseling.
RATIONALE: Mentoring is important. Most of what we know about mentoring is for faculty. The benefits might be expected to extend to residents in need of career and personal counseling.

OBJECTIVE: Assess program directors’ attitudes about mentoring, estimate prevalence of formal mentoring programs, describe characteristics of formal mentoring programs.
METHODS: Send mail survey to internal medicine program directors. Page 1 assesses program demographics and attitudes. Page 2 asks about structural features of formal mentoring programs.

KEY FINDINGS: Attitudes regarding mentoring were very favorable. Half had a formal program. Programs are largely unstructured, loosely monitored, and under evaluated.
<table>
<thead>
<tr>
<th>Attitude</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mentorship is an important tool for career/professional development</td>
<td>91%</td>
</tr>
<tr>
<td>Program Directors have a responsibility to encourage faculty to mentor residents</td>
<td>92%</td>
</tr>
<tr>
<td>It is important for a resident to have a mentor during training.</td>
<td>79%</td>
</tr>
<tr>
<td>Program Directors have a responsibility to encourage residents to identify a mentor</td>
<td>73%</td>
</tr>
<tr>
<td>Program Directors have a responsibility to identify mentors for residents</td>
<td>66%</td>
</tr>
<tr>
<td>Residency Programs should have structured mentoring programs for residents</td>
<td>61%</td>
</tr>
<tr>
<td>Features of a Mentoring Program:</td>
<td>Ideal %</td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Individual mentoring</td>
<td>89</td>
</tr>
<tr>
<td>Group mentoring</td>
<td>23</td>
</tr>
<tr>
<td>Peer mentoring</td>
<td>35</td>
</tr>
<tr>
<td>Regularly scheduled meetings</td>
<td>57</td>
</tr>
<tr>
<td>Evaluation by residents</td>
<td>36</td>
</tr>
<tr>
<td>Evaluation by mentors</td>
<td>30</td>
</tr>
<tr>
<td>A structured curriculum</td>
<td>14</td>
</tr>
</tbody>
</table>
Principle #2: Select Optimal Data Collection Methods
“Do the Right Thing”

TAKE HOME MESSAGE:
Process assessment is useful for learning about features of programs - what they do/offer (caution: what people say they do may be different from what they actually do). Triangulation is helpful.
Principle #3: Gather Enough Data
“More Is Better Than Less”

What does this mean?
- Need an adequate number of observations
- Case specificity
- Sampling/bias
- Capture more signal than noise
How do we evaluate?

<table>
<thead>
<tr>
<th></th>
<th>Knowledge</th>
<th>Skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ test</td>
<td>++</td>
<td>-</td>
</tr>
<tr>
<td>SP based test</td>
<td>--</td>
<td>++</td>
</tr>
<tr>
<td>Teachers/attendings</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Chart reviews</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Where is the sampling error?

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MCQ test</td>
<td>Wrong content, not enough items</td>
</tr>
<tr>
<td>SP based test</td>
<td>SP training/specificity</td>
</tr>
<tr>
<td>Teachers/attendings</td>
<td>Different ideas/ “hawks” and “doves”</td>
</tr>
<tr>
<td>Chart reviews</td>
<td>A lot does not get recorded</td>
</tr>
</tbody>
</table>

***Need a lot of observations***
Example for Principle 3:

Example for Principle 3:

Everyone knows students are not being watched and given feedback. What if we designed a process to help students get feedback about their history and physical examination skills?

## Medical Education Research Framework

<table>
<thead>
<tr>
<th>Unit of Analyses</th>
<th>Focus of Research</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Person</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Program</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
RATIONALE: Medical students must acquire core clinical skills. Many students complete training without being observed performing a history or physical exam. A feasible tool that promotes observation is needed.

OBJECTIVE: To determine the feasibility, reliability and validity of the mCEX when used in a medicine core clerkship.
<table>
<thead>
<tr>
<th>Skill</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical Interviewing Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Physical Examination Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Humanistic Qualities/Professionalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Clinical Judgment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Counseling Skills</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Organization/Efficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Overall Clinical Competence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Mini-CEX Time: Observing: _____ min    Feedback: _______min

Evaluator Satisfaction with Mini-CEX
LOW 1 2 3 4 5 6 7 8 9 HIGH

Student Satisfaction with Mini-CEX
LOW 1 2 3 4 5 6 7 8 9 HIGH
METHODS:
Participants: 121 medicine clerkship students (1/02-9/02)
Requirement: complete 9 mCEX during clerkship
mCEX booklets: ratings in 7 domains on 9-point scale
  » Record time and satisfaction with encounter

ANALYSES:
  » Feasibility: number of forms, time and satisfaction
  » Reproducibility: student x item x rater D-study
  » Validity: Pearson correlations and ANOVAs
KEY FINDINGS:
Feasibility

Completion rates: 89% of all forms (n=1,297)
Mean # forms/student: 7.9 (range 2-10, median 8)
Median observation time: 15 minutes
Median feedback time: 5 minutes
Evaluator Satisfaction

<table>
<thead>
<tr>
<th></th>
<th>Inpatient Faculty</th>
<th>Outpatient Faculty</th>
<th>Residents</th>
<th>p &lt; .0001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Satisfaction</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

p < .0001
Reproducibility: How Many Do We Need?

<table>
<thead>
<tr>
<th>Number of Forms</th>
<th>Reproducibility Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>.58</td>
</tr>
<tr>
<td>6</td>
<td>.68</td>
</tr>
<tr>
<td>8</td>
<td>.74</td>
</tr>
</tbody>
</table>
Principle #3: Gather Enough Data
"More Is Better Than Less"

TAKE HOME MESSAGE:
Multiple observations are needed to get a reliable estimate of performance.
Principle #4:
Interpret the Data Appropriately
“What Does it Really Mean”

What does this mean?
make a correct interpretation
consider alternative explanations
know the limits of your data
How to Study Validity

From the books….
content
concurrent
construct
consequential

Do the data behave as expected? (confirmatory)
Can we make sense of the results? (exploratory)
Example Principle 4:

Do residents and students agree on who is a good teacher? What if we used our existing evaluation data and ‘looked back’ to see?

# Medical Education Research Framework

<table>
<thead>
<tr>
<th>Unit of Analyses</th>
<th>Focus of Research</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Process</td>
</tr>
<tr>
<td>Person</td>
<td>x</td>
</tr>
<tr>
<td>Program</td>
<td>x</td>
</tr>
</tbody>
</table>
RATIONALE: Medical students and residents routinely evaluate teachers and attendings. The data have consequences for retention, promotion, and salary.

OBJECTIVE: To examine differences in students’ and residents’ ratings of the same clinical faculty and see how scores varied over the course of a year.
METHODS: Retrieved data from evaluation database for all residents and students in medicine clerkships, electives and residency.

KEY FINDINGS: Students gave more favorable ratings than residents. Over the year, students became less critical and residents became more critical. Winners of teaching awards had better ratings.
**Example of Quarter Data**

<table>
<thead>
<tr>
<th></th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>2.7</td>
<td>2.6</td>
<td>2.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Residents</td>
<td>1.8</td>
<td>2.1</td>
<td>2.4</td>
<td>2.7</td>
</tr>
</tbody>
</table>

On a scale where 1 = almost always and 4 = never
### But How Do the Data Really Line Up?

<table>
<thead>
<tr>
<th></th>
<th>Jan-Mar</th>
<th>Apr-Jun</th>
<th>Jul-Sep</th>
<th>Oct-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>2.7</td>
<td>2.6</td>
<td>2.1</td>
<td>1.7</td>
</tr>
<tr>
<td>Residents</td>
<td>2.4</td>
<td>2.7</td>
<td>1.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

On a scale where 1 = almost always and 4 = never
What are the Consequences?

![Graph showing the consequences for students and residents from Jan-Mar to Oct-Dec.](image)
Principle #4: Interpret the Data Appropriately
“What Does it Really Mean”

TAKE HOME MESSAGE:
Don’t just analyze the data - think about the consequences of using the data.
Summary

Medical education framework
  unit of analysis
  process and/or outcomes
Four research principles
  enhance quality of medical education research and in turn education
Things I Did Not Talk About

Qualitative studies
Unique challenges:
- real deadlines and timelines
- small samples
- many “competing” curricula
Abundant opportunity and need to conduct medical education research