WEIGHING OUR OPTIONS

An Empirical Approach to Selecting an Intermediate Outcome Measure

Adrienne Smith

UNC COLLEGE OF ARTS & SCIENCES
PUBLIC POLICY

EPIC EDUCATION POLICY INITIATIVE at CAROLINA

THE PROBLEM

More often stakeholders want to make bold, data-driven statements about the effectiveness of a program, even when programs are in their infancy. Knowing the challenge of detecting long-term effects early in a program's implementation, but also wanting to be responsive to stakeholder needs, evaluators turn to intermediate measures of program impacts.

How can evaluators be more confident about whether the instrument selected is measuring a construct that relates to the long-term outcome of interest?

THE

In this presentation I propose a method for empirically testing between several intermediate measures in the quest to understand whether a program has an effect on the long term outcome of interest.

The focus is on the method, but I will illustrate through an example…

CONTEXT

Reform → Desired Outcome

Teacher Preparation → Greater Student Learning

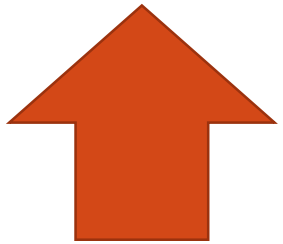
Such a long wait!

ON TRACK?

Used the program's theory of action to fill the gap so we can see if the reform is on-track

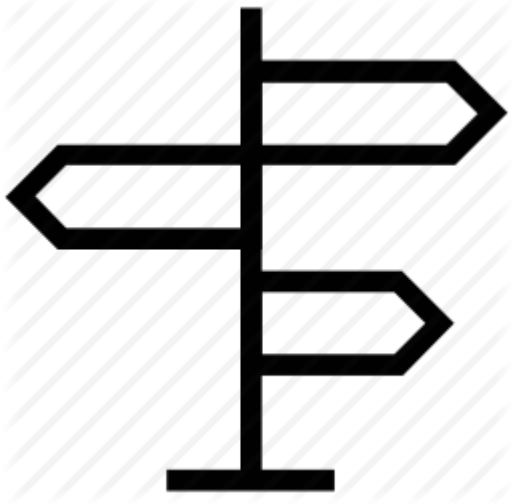
Teacher Prep Program → Higher Quality Teaching → Greater Student Learning

Reform



Now we ***just*** have to find a good measure for Teaching Quality!

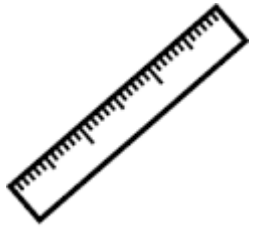
GUIDED BY THE THEORY OF ACTION



Again, using the program's theory of action, we identify classroom teaching practices that should be impacted by the program. Increases in these practices are hypothesized to lead to higher student achievement.

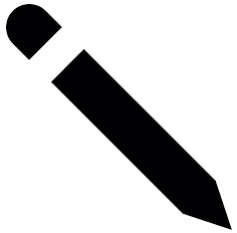
First we look for measures already available…..

CHOICES



Choice A: a well-validated measure

Choice B. a newer measure



Choice C. a project-developed measure

SHORT PILOT

Four years left in the study

Used all three for one year

Faithful to ideal measurement conditions

Once the data were collected and scored we turned our attention to the analysis…

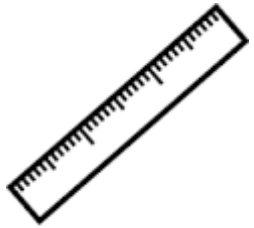
PREDICTIVE ANALYSIS – STEP 1

Used all cases (no need to differentiate between treatment and comparison groups)

Set up three regressions, one for each potential intermediate outcome measure

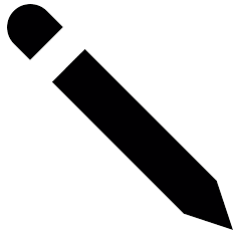
Regressed the long-term desired outcome (teacher effectiveness rating from student test scores) on set of dimensions for each intermediate outcome measure

PREDICTIVE ANALYSIS – STEP 1 CONT.



Choice A: Y = B0 + B1PositiveClimate + B2InstructionalDialogue +…error

Choice B: Y = B0 + B1StudentThinking + B2ClarityofDelivery +…error

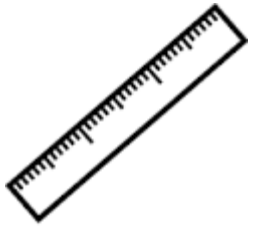


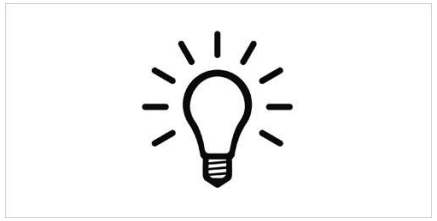
Choice C: Y = B0 + B1GraphicOrganizers + B2Think,Pair,Share +…error

PREDICTIVE ANALYSIS – STEP 1 CONT.

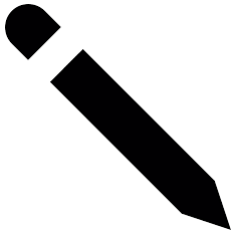
Compared the proportion of variance in the long-term desired outcome explained (R-squared) for each of the three regressions

C > A > B



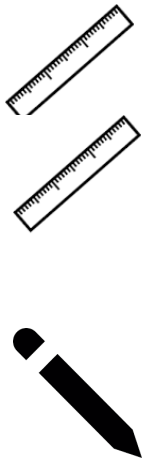


At this point the results favored the project-developed measure



PREDICTIVE ANALYSIS – STEP 2

Ran dimensions separately to see if any dimension was particularly powerful



Choice A: $Y = B0 + B1PositiveClimate + error$

Choice A: $Y = B0 + B1InstructionalDialogue + error$

.

.

.

Choice C: $Y = B0 + B1GraphicOrganizers + error$

Some R-squared values were larger than others, but nothing stood out

PREDICTIVE ANALYSIS – STEP 3

Ran them all together to see if they were measuring different aspects of Quality

Y = B0 + B1PositiveClimate + B2InstructionalDialogue + B3StudentThinking + B4ClarityofDelivery + B5GraphicOrganizers + B6Think,Pair,Share +…error

Very tiny increase from the R-squared for Choice C, which told us we were probably taping the same construct with all three measures

PREDICTIVE ANALYSIS – STEP 4

Repeated steps 1-3 with another outcome measure of interest (principal evaluation ratings)

Same pattern emerged

Step 1: C > A > B

Step 2: Some were larger than others, but nothing stood out

Step 3: Tiny change in R-squared over and above Choice C

REPORTING BACK TO CLIENT

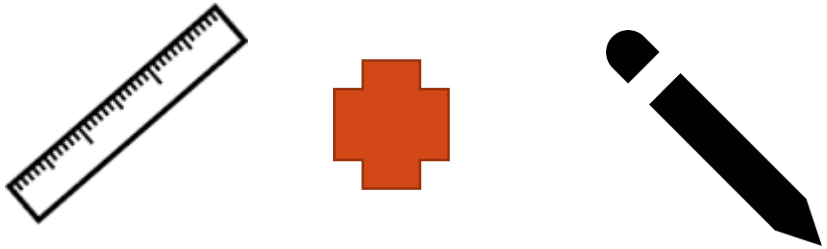
Presented findings from empirical analysis

Also had to consider cost, ease of collecting data, ease of scoring instrument, participant burden

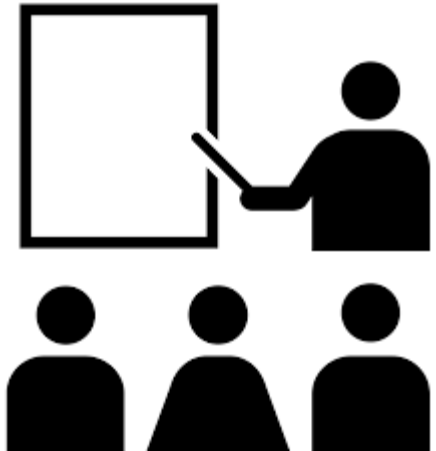
Client happy because in the future they could continue to measure intermediate outcome and intervene early

DECISION

Client chose to use A and C and dropped B



Client is training more individuals to be data collectors for A and C to help shore up their formative assessment of reform success



LINK TO THEME

The conference theme of **looking forward** calls us as evaluators to reflect upon our methods.

As our society is increasingly calling for **more rigorous evidence** we must reflect on the quality of our data collection tools and our processes for ensuring that the data obtained is measuring constructs of value.

By repurposing existing tools in our toolkit we can increase the level of confidence – both ours and the clients - in the intermediate outcomes we have identified.

CONTACT ME!

Adrienne Smith, PhD

Director of Research and Evaluation & Senior Research Scientist

Education Policy Initiative at Carolina (EPIC)

UNC Public Policy

314 Cloister Court

Chapel Hill, NC 27514

(o): 919.962.1178

(c): 919.616.1565

adrsmith@email.unc.edu

http://publicpolicy.unc.edu/epic-home