

**Evaluating STEM  
Teaching-and-Learning Innovations  
at Multiple Scales**

Kirk Knestis, PhD

**Hezel**  
ASSOCIATES

# Hezel Associates, Syracuse NY

## Specialists in STEM and Workforce Education

- 10-person for-profit contracting firm
- 11 ea. U.S. Department of Labor workforce development education projects
- 3 ea. U.S. Department of Education grant subcontracts
- 13 ea. National Science Foundation research or evaluation subcontracts

# Today's Conversation

## Evaluating STEM Teaching-and-Learning Innovations at Multiple Scales

- Challenges of Multilevel, Multisite Evaluation
- Research & Development Orientation
- Studying Innovations versus Programs
- Implications for the NSF INCLUDES Program

# Today's Conversation

## **Evaluating STEM Teaching-and-Learning Innovations at Multiple Scales**

- Challenges of Multilevel, Multisite Evaluation
- Research & Development Orientation
- Studying Innovations versus Programs
- Implications for the NSF INCLUDES Program

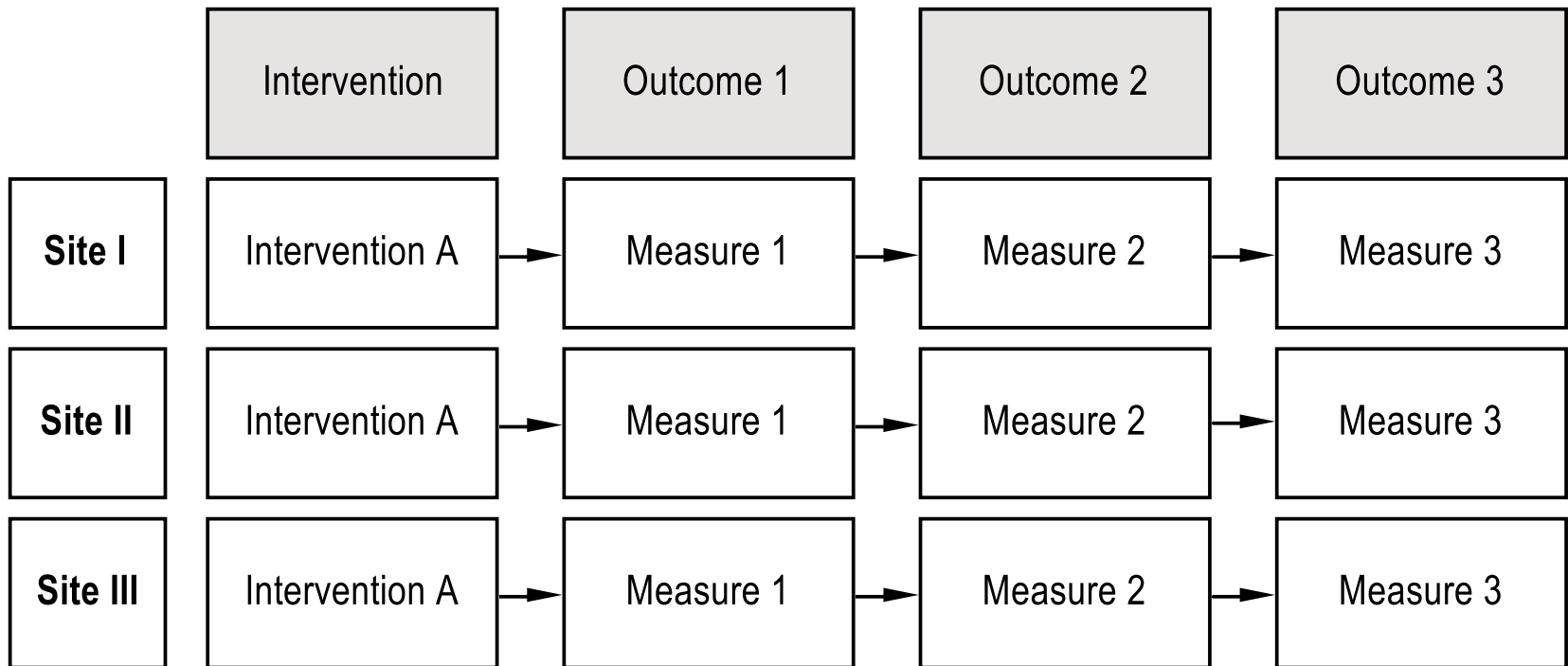
# Multilevel, Multisite Evaluation

## Variations

- Single interventions implemented at multiple sites
- Multiple interventions implemented at multiple sites with shared outcomes
- Nested models (e.g., projects within a program)
- Multilevel theoretical models

# Multilevel, Multisite Evaluation

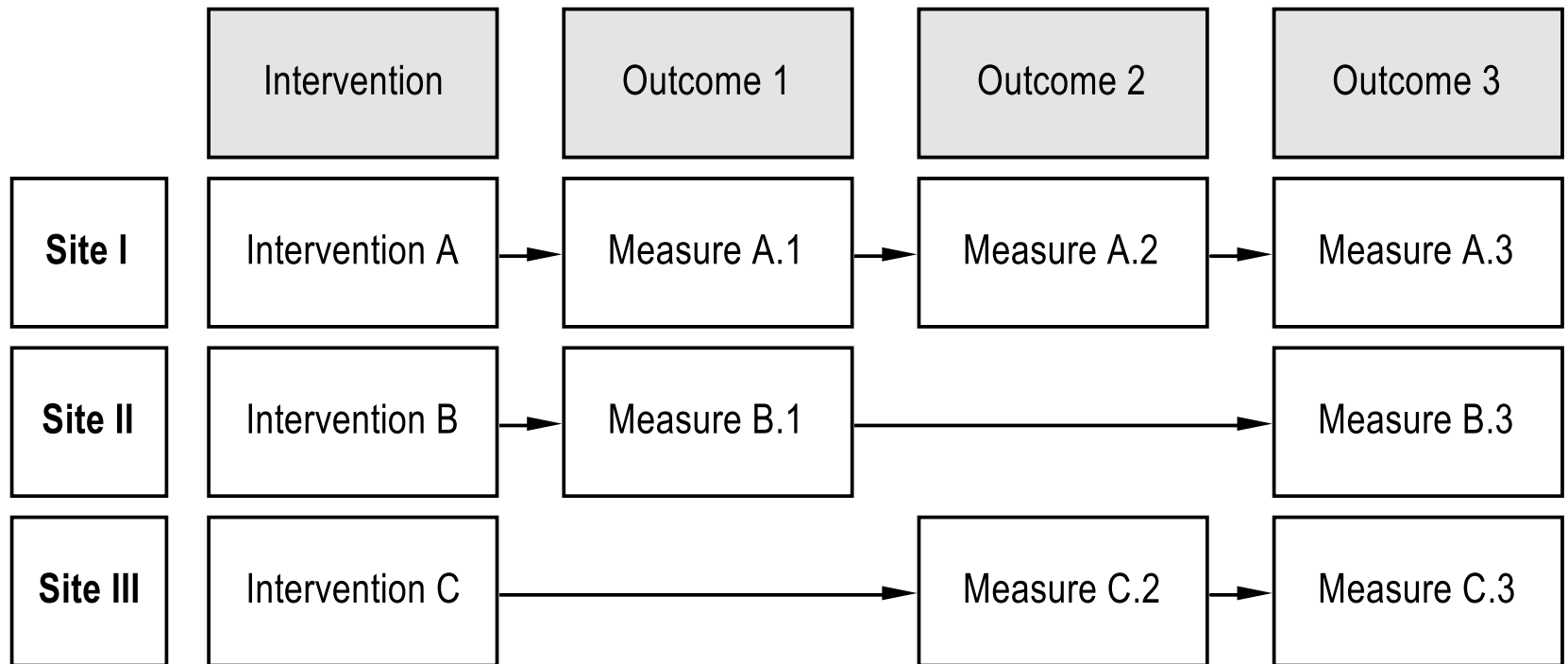
## Single Intervention, Multiple Site



*School district-wide implementation of an intervention*

# Multilevel, Multisite Evaluation

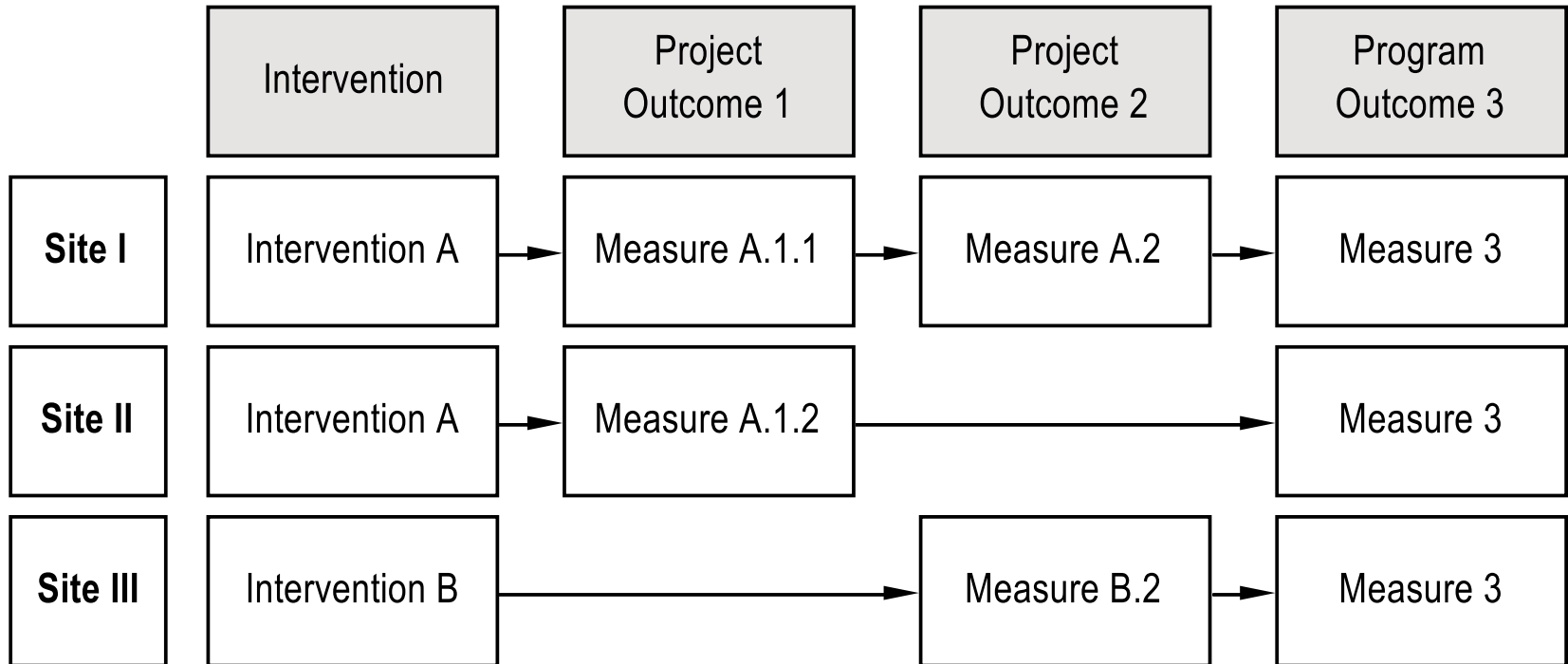
## Multiple Interventions, Shared Outcomes



*Typical NSF EHR project (e.g., ITEST)*

# Multilevel, Multisite Evaluation

## Nested Models, Projects within a Program

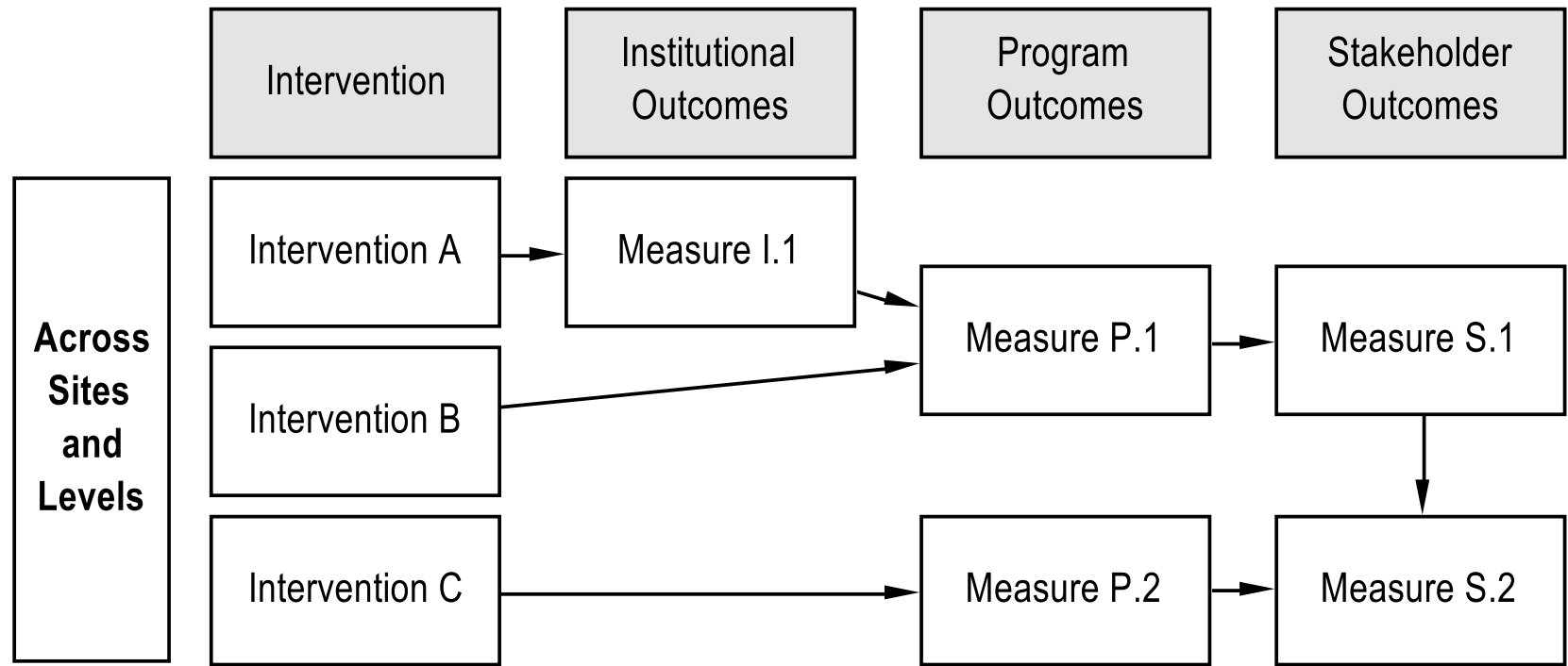


*Federal program with standardized outcome (e.g., GPRA)*



# Multilevel, Multisite Evaluation

## Multilevel Theoretical Models



*Messy logic with tiered outcomes: INCLUDES?*

# Today's Conversation

## **Evaluating STEM Teaching-and-Learning Innovations at Multiple Scales**

- Challenges of Multilevel, Multisite Evaluation
- **A Research & Development Orientation**
- Studying Innovations versus Programs
- Implications for the NSF INCLUDES Program

# Research & Development

## The Problem – *The “NSF Conundrum”*

Historically, distinctions between “research” and “evaluation” have been unclear or inconsistent

- Grantee **PIs** focused on delivery of program activities
- External **evaluators** often became *de facto* researchers, testing the PI’s innovation
- Quality of both research and evaluation suffered

# Research and Evaluation

## **One Response – *The Common Guidelines***

- Innovations should be conceived, improved, and adopted to achieve lasting education outcomes for stakeholders
- Emphasis on models rather than instances of implementation
- Learning from such work should advance collective understandings

# Research & Development

**R&D Type**



**Purposes**

6. Scale-up
5. Effectiveness
4. Efficacy
3. Design & Development
2. Early-Stage/Exploratory
1. Foundational

Iteratively improve the innovation's design; so its promise for impact

Advance collective understandings about teaching and learning

# Research & Development

**R&D Type**



**Purposes**

6. Scale-up
5. Effectiveness
4. Efficacy
3. Design & Development
2. Early-Stage/Exploratory
1. Foundational

(IES & NSF, 2013)

**Development**

*NSF Broader Impacts*

**Research**

*NSF Intellectual Merit*

# Today's Conversation

## **Evaluating STEM Teaching-and-Learning Innovations at Multiple Scales**

- Challenges of Multilevel, Multisite Evaluation
- Research & Development Orientation
- **Studying Innovations versus Programs**
- Implications for the NSF INCLUDES Program

# Innovations vs. Programs

**Research**

**Evaluation**



# Innovations vs. Programs

## Research & Development

Reframed as **Research and Development (R&D)**

*Structured study of the **innovation** in terms of its promise of effectiveness*

*Internal to the project, working with designers*

## Program Evaluation

Reframed as **Program Evaluation**

*Study of implementation and results of the **R&D** activities*

*External to the project, third-party perspective*

# Innovations vs. Programs

## Research & Development

6. Scale-up
5. Effectiveness
4. Efficacy
3. Design & Development
2. Early-Stage/Exploratory
1. Foundational

## Program Evaluation

- Implementation-Results
- Process-Product
- Monitoring
- Performance Reporting
- Formative Feedback

*Examines both **research** & **development** activities!*

# Today's Conversation

## **Evaluating STEM Teaching-and-Learning Innovations at Multiple Scales**

- Challenges of Multilevel, Multisite Evaluation
- Research & Development Orientation
- Studying Innovations versus Programs
- Implications for the NSF INCLUDES Program

# Implications for INCLUDES

## From the INCLUDES Design and Development Launch Pilots Solicitation (NSF 17-522)

- **Broader Impacts** – Anticipated societal goals
- **Goal** – Ultimate aim; *“to achieve impact at the national level”*
- **Objectives** – Typically expected to be measurable; may be activities or outcomes
- **Outcomes** – Typically framed as lasting results; persistent changes for groups of people

# Implications for INCLUDES

## From the INCLUDES Design and Development Launch Pilots Solicitation (NSF 17-522)

- **Strategies** – Higher-level statements of “what the money is being spent on”
- **Activities** – Specific actions intended to generate **outputs** and result in **outcomes**
- **Outputs** – The completion of, or deliverables generated by, activities

# Implications for INCLUDES

## From the INCLUDES Solicitation (NSF 17-522)

- **Social Innovation Framework** – Conceptual model for a novel solution (effective, sustainable) to a social problem; value accrues primarily to society (NSF *Broader Impacts*)
  - Associating, Questioning, Observing, Networking, and Experimenting
  - Burning, Sensing, Questioning, Idea Networking, Associating, Experimenting, and Impacting

# Implications for INCLUDES

## From the INCLUDES Solicitation (17-522)

- **Collective Impact** – Specific social innovation framework
  - Common agenda
  - Mutually reinforcing activities
  - Continuous communication
  - Shared measures
  - Backbone organization

*Note: NSF is prescribing processes toward a single outcome*

# From INCLUDES Solicitations

Meaningful assessment and evaluation of NSF funded projects should be based on appropriate metrics, keeping in mind the likely correlation between the effect of broader impacts and the resources provided to implement projects. (16-544 & 17-522)

- *“Appropriate metrics” translates into “common measures” in a collective impact model*



# From INCLUDES Solicitations

[DDLDP projects] demonstrate how extant teams and organizations can be reconfigured and joined together to **form new alliances** with common goals and purposes and collective impact-style approaches, with a strategy for how the **effective practices** of the Alliance are likely to be **deployed at scale** (16-544)

- *Common measures may examine how teams form alliances, how “effective” practices are, or how efforts to “deploy them at scale” work*

# From INCLUDES Solicitations

If the size of the activity is limited, evaluation of that activity in isolation is not likely to be meaningful. ...may best be done at a higher, more aggregated, level than the individual project.

(16-544 & 17-522)

- Use of common measures allows evaluation across instances of an activities, with or without disaggregation at the “site” level

# From INCLUDES Solicitations

Describe agreed-upon ways to measure and report success, including the selection of an external evaluator (16-544 & 17-522)

- An external evaluator may be best positioned to provide external evaluation of R&D activities and outputs, from a “critical friend” outsider perspective

# From INCLUDES Solicitations

Include a description of an evaluation plan that uses benchmarks, indicators, logic models, road maps, or other evaluative methods to document progress toward goals, objectives, and outcomes defined in the proposal. (17-522)

- The logic model should illustrate theoretical relationships among factors that translate NSF \$\$ into “active participation of under-represented groups in STEM”

# Implications: What is studied?

## **Research** *& Development*

- “Bold, innovative ways for solving a broadening participation challenge in STEM”
- Models or prototypes
- Effective practices
- Strategies for seeking & developing STEM talent

## *Program* **Evaluation**

- Synthesis activities
- Building of a research base
- Spreading/adapting effective practices
- Support for partnerships to develop those strategies

# Implications: What is studied?

## Research & Development

- Innovations developed by a **Collective Impact** approach

*Considering an R&D orientation starting with **Design & Development Research...***

## Program Evaluation

- **Collective Impact R&D work**
- Common agenda; mutually reinforcing activities
- Shared measures
- Nature and quality of communication
- Processes and quality of collaboration (e.g., backbone organization)

# Implications for INCLUDES

## Practical Considerations

- INCLUDES may be an example of the most complex kind of multisite, multilevel evaluation
- Development of common measures requires shared definitions of outcomes; harder than it might seem
- Implementation quality and fidelity will be hard to understand given variability and multisite, multilevel nature of the program

# Implications for INCLUDES

## Practical Considerations

- Evaluation of collaboration, communication, and consensus building (common agenda and shared measures) must examine **quality** of processes and outputs; not easy or cheap
- Consensus re: measures of quality are even harder to develop than other common measures



# Implications for INCLUDES

## Potential for Scaling

- Guidance from NSF is not explicit re: scaling the innovation vs. scaling the R&D activities (good and bad news of ambiguity)
- Scale-up may be addressed in terms of either – I think!

# Implications for INCLUDES

## Theory-based Evaluation Design

- INCLUDES theories of action are very long (e.g., like US ED institution-level grants)
- Actual outcome (broadened participation) is probably years in the future; clear logic is crucial!
- Opportunities for “mutation” are greater—fidelity issues, transformation of intentions, or growth beyond the R&D model of scale (Dede/Microsoft Scaling Framework)

# Implications for INCLUDES

## Research & Development

- |                            |   |                |
|----------------------------|---|----------------|
| 6. Scale-up                | → | Evolution      |
| 5. Effectiveness           | → | Shift          |
| 4. Efficacy                | → | Spread         |
| 3. Design & Development    |   | Sustainability |
| 2. Early-Stage/Exploratory |   | Depth          |
| 1. Foundational            |   |                |

**Questions? Discussion?**

**Kirk Knestis PhD**

CEO

Hezel Associates, LLC

731 James Street #410

Syracuse, NY 13203

kirk@hezel.com