Findings and Methodological Lessons from an Evaluation of a Project to Integrate the Arts into Elementary School Reading and Mathematics Instruction

Paul R. Brandon and Brian Lawton
Curriculum Research & Development Group
University of Hawai'i at Mānoa
brandon@hawaii.edu

Annual Meeting of the American Evaluation Association
Baltimore, MD
November 2007

1

Purpose of the Paper

- To report a self-meta-evaluation of a quasi-experimental study of a small federally funded project
- How strong can we expect the conclusions of small quasi-experimental evaluations of federally funded programs to be?

2

Primary Meta-Evaluation Foci

- *"Project-maturity" issues:* Are most small federally funded projects ready experimental designs?
- *Instrumentation issues:* Collecting data on young students
- *Design issues:* Contextual differences among groups; sampling units and appropriate analyses; generalizability of results
- *Implications for social policy:* Funding small federal programs

3

The Evaluated Project

- The evaluation was of the ARTS FIRST Windward Research Project
- Funded by the Arts in Education Model Development and Dissemination (AEMDD) Grant Program.
- Evaluated by doctoral-level generalist evaluators

4

AEMDD Purpose

- *From the RFP for AEMDD:* To support "the enhancement, expansion, documentation, evaluation, and dissemination of innovative, cohesive models that are based on research and have demonstrated that they effectively:
  - (1)Integrate standards-based arts education into the core elementary and middle school curricula;
  - (2) strengthen standards-based arts instruction in these grades; and
  - (3) improve students' academic performance, including their skills in creating, performing, and responding to the arts."

5

Preferred AEMDD Evaluation Designs

- "Evaluation methods using an experimental design are best for determining project effectiveness. . . . If random assignment is not feasible, the project may use a quasi-experimental design with carefully matched comparison conditions. This alternative design attempts to approximate a randomly assigned control group by matching participants--e.g., students, teachers, classrooms, or schools--with non-participants having similar pre-program characteristics. . . ."

6

(Preferred Designs, cont'd.):

- "In cases where random assignment is not possible and participation in the intervention is determined by a specified cutting point on a quantified continuum of scores, regression discontinuity designs may be employed. . . ."
- "Proposed evaluation strategies that use neither experimental designs with random assignment nor quasi-experimental designs using a matched comparison group nor regression discontinuity designs will not be considered responsive to the priority when sufficient numbers of participants are available to support these designs."

7

Required Evaluation Rigor

- Evaluation requirements seem to reflect Institute for Education Science specifications for efficacy and effectiveness studies (focus on whether programs work, not how)
- No provision for using formative evaluation methods for development studies (IES "Goal 2 studies")
  - Our experience suggests most AEMDD projects were development efforts.

8

"Development" or "Enhancement"?

- Note the discrepancy between the program title and description.
  - Title: Arts in Education Model Development and Dissemination (AEMDD) Grant Program
  - Description: ". . . to support "the enhancement, expansion, documentation, evaluation, and dissemination"
  - "Development" is not the same as "enhancement."

9

Project Purpose

- The ARTS FIRST Windward Research Project *purpose:* To examine the extent to which elementary teachers' use of arts strategies in basic skills instruction improved students'
  - achievement in basic skills.
  - attitudes toward school.
  - interest in the arts.

10

Project Model

- Strategies taught in four art forms:
  - Drama
  - Dance
  - Music
  - Visual arts
- Six full-day teacher professional development sessions throughout the year
- In-class mentoring by professional artists with teaching experience.

11

Program Scope and Duration

- Three-year program in three Title I schools, with an evaluation each year
- Served Grades 3-5, beginning with Grade 3 and adding a grade each year.
  - An expensive program:
    - The schools did not have art teachers; artist mentors had to be found and funded.
    - The project could not afford to serve Grade 3 in Year 3.
- Our meta-evaluation primarily examines the third year of the study, when Grades 4 and 5 were served.

12

Evaluation Design

- A quasi-experimental study (non-equivalent control group) with three pairs of schools (project and control) matched on
  - reading achievement
  - SES
  - school size
  - ethnicity
- Matched schools randomly assigned within pairs in Year 1

13

Mixed-Method Evaluation

- *Quantitative data (*Good reliability and validity results, as reported previously):
  - **Students' achievement**
  - **Students' attitudes toward school**
  - **Students' interest in the arts**
  - Teachers' attitudes toward the arts
  - Weekly teacher implementation logs
  - Professional development quality (project group only)
  - Ratings of teachers' quality in using arts strategies
- *Qualitative data* (project group only):
  - student focus groups
  - teacher focus groups
  - professional development quality (open-ended responses)

14

Project-Maturity Issues

15

Ready for Prime Time?

- Contrary to what we had been led to believe about the project at the inception of the study, the AFWRP project evolved considerably over the three years.
  - The arts strategies that were taught were narrowed.
  - The PD methods were revised.
- Not unusual for programs of this nature.

16

Instrumentation Issues Complicating the Interpretation of Results

17

Restriction of Range

- Student attitude and interest-in-the-arts instruments showed ceiling effects in Years 1 and 2.
  - Calculating and analyzing IRT scores for best-discriminating items helped distinguish among groups.

18

How Much About Young Children Can Be Measured Quantitatively?

- Quantitative measures are essential for experiments and quasi-experiments
- Are attitude or student-interest instruments appropriate for measuring children's developmental level?
  - Little found in the literature on measuring elementary school children's attitudes.
- *Bias due to primacy?*
  - Students might have responded to instruments based on recent experiences

19

Design Issues Complicating the Interpretation of Results

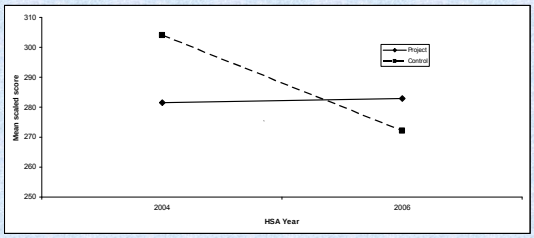
20

Student Quantitative Data Analyses and Results

- Examined project-control differences at the end of Grade 5 (students who had participated all three years), using a fixed-effects statistical model with Grade 3 pretest and propensity scores as covariates.
  - Achievement (results favored project group; post-hoc analysis showed significant difference only between highest-scoring project school and lowest-scoring control school)
  - Attitudes toward school (results favored project)
  - Interest in the arts (dance: results favored project; music: results favored control; others no difference)
  - (Results presented previously [AEA, AERA])

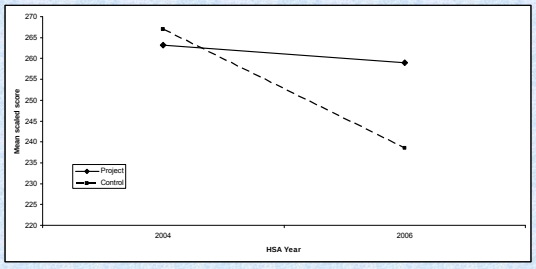
21

Longitudinal Reading Results



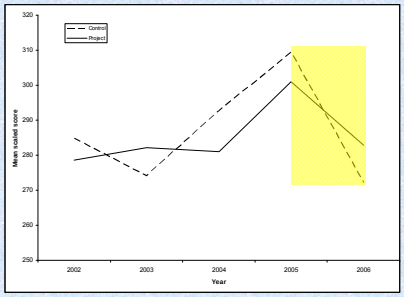
22

Longitudinal Math Results



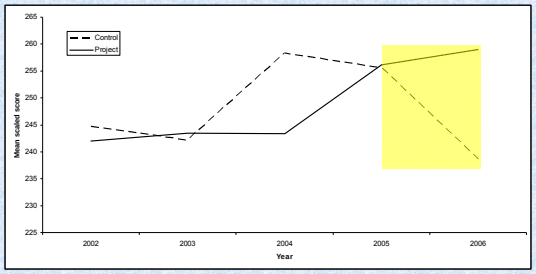
23

Cross-Sectional Validation (Reading)



24

Cross-Sectional Validation (Math)



25

Validity Threats Due to Project History

- *Schools' NCLB Standing*
  - Two control schools were "in good standing, unconditional."
  - Two schools, one project and one control, were in "school improvement Year 2."
  - One project school was in "corrective action."
  - One project school was in "planning for restructuring."
- Reading programs used at the schools (e.g., Success for All used at highest-scoring school)

26

Aspects of the Study That Complicated the Choice of Analysis

- Chose schools as it were a group-randomized (i.e., cluster-sample) study.
  - Six schools was an insufficient size for a valid group-randomized study because of low statistical power.
  - Not surprisingly, the analysis using a random-effects statistical model showed no significant differences among groups.

27

Fixed Effects Model

- We used a fixed-effects statistical model (i.e., schools not considered random effects)
  - This model advocated by IES in the past for Goal 2 (development) grants.
  - The model seems appropriate for us because of the low power of the random-effects model.

28

Problems With Fixed Effects Model

- Cannot generalize results to other sites.
- Some argue strongly against fixed-effects model when groups are the unit of assignment because of the strong possibility of inflating the Type I error rate
  - Murray (1998, p. 108) says that using the fixed effects model is a "seductive trap."

Murray, D. M. (1998). *Design and analysis of group-randomized trials.* New York: Oxford University.

29

Issues About the Use of Propensity Scores

- Propensity scores are recommended as covariates to help equate non-equivalent groups.
  - However, it is often stated that ANCOVA should only be used with randomly assigned sampling units (although the literature is contradictory about this).
- We used propensity scores at the level of the individual in our analysis. Should they have been at the level of the school (the sampling unit)?

30

Social Policy Issues

31

Can AEMDD Do What It Seeks to Accomplish?

- Have enough projects shown prior success (a criterion for eligibility for AEMDD)?
  - Should AEMDD and similar programs emphasize *development*, as the title says, or *enhancement*, as the program description says?
- Evolving projects form moving targets that are inappropriate for experimental study.

32

Funding Levels

- Most AEMDD evaluations were allocated considerably smaller portion of funds than us ($330,000 for 3 years-- about 40-45% of total funding)
- Is funding sufficient for conducting good group-randomized studies?

33

- Low funding a substantial reason for the small number of groups included small studies.
  - "The small number of randomized experiments in education may reflect a simply calculation of how difficult they are to mount." Cook (2001)

34

- Despite low funding, perhaps is good social policy (and politically essential) to support the arts.
- Perhaps it is also good policy to insist evaluators focus more on whether programs work than why they work.

35

Summary

- Many proposals are submitted for funding to develop new projects, not to enhance existing ones.
- These projects are not ready for quasi-experiments.
- Contextual issues may result in validity threats.
- For elementary school programs, quantitative data possibilities are limited.
- The findings of fixed-effects and other quasi-experimental analyses lack strong warrants.
- Funding for small projects may be too little to arrive at the kinds of conclusions that policymakers seek.

36