Using State Tests for Evaluation Purposes

CCSSO National Conference on Student Assessment

June 21, 2010

CPRE mdrc MATHEMATICA Policy Research, Inc.

Whether and How to Use State Tests in Education Experiments

Henry May
CPRE, University of Pennsylvania

Irma Perez-Johnson, Josh Haimson
Samina Sattar, Phil Gleason
Mathematica Policy Research

CCSSO National Conference on Student Assessment
June 21, 2010

CPRE

MATHEMATICA
Policy Research, Inc.

Overview

I. Deciding Whether to Use State Tests

II. Key Issues when Using State Tests

III. Recommendations for Best Practices

Key Issues in Deciding Whether to Use State Tests

- Validity Issues
  - Relevance to questions & intervention is key
    - Narrow outcomes (e.g., reading fluency)
    - Broad outcomes (e.g., proficiency)
  - A lack of comparability across grades and/or states can be problematic
  - Recognizing conflicting opinions about combining results from different tests is important when presenting results

Key Issues in Deciding Whether to Use State Tests

- Reliability Issues to Consider
  - Conditional measurement error
  - Ceiling and floor effects
  - Implications for statistical power
- Feasibility Issues to Consider
  - Consent / Privacy (FERPA)
  - Following mobile students

Key Issues when Using State Tests

- Use of baseline or historical data
  - Pre-intervention verification of equivalent groups
  - Improving statistical power though covariance analysis
- Impact analyses using scale scores is preferred
- Analysis and interpretation of proficiency level scores (e.g., Below Basic, Basic, Proficient, Advanced) and proficiency rates can be very messy
    - Proficiency rates vary widely across states
    - Proficiency scores are not interval-scaled

Key Issues when Using State Tests

- Complications in Studies with Multiple Grades/States
  - It may be difficult to interpret results when tests measure different skills/knowledge
  - Why combine results across grades or states
    - Similarity across tests and study samples (unlikely)
    - Modest sample sizes from each grade or state
    - Desire for broad-based impact estimates

Key Issues when Using State Tests

- It is important to establish a consistent reference population in multi-grade and multi-state studies
  - Standardized impact estimates can be thrown off by shifts in the total variability of the study sample in each grade or state.
  - Standardization relative to the statewide population can account for differences in study samples across states and grades.

Recommendations: RCT Design

- In order to use a state test in an RCT, the assessment should…
  - exhibit adequate alignment with the research questions and/or the intervention theory of action.
  - have adequate reliability for the target population.
  - have baseline and post-intervention data available.

Recommendations: RCT Design

- In order to produce combined impact estimates across multiple grades and/or states, the individual state tests should also…
  - exhibit similar alignment with the research questions and/or the intervention theory of action.
  - have similar reliability (i.e., no ceiling/floor effects) for the target population.
  - have similar participation rates for the target population.

Recommendations: RCT Design

Calculating multi-grade/state impact estimates

- If the tests are all on a common vertically-equated scale from a single state, analyses should utilize the vertical scale scores.
- If the tests are not vertically equated, or are from multiple states, then the test scores must be rescaled to a common metric before estimating combined impacts.
  - If the target population is similarly represented in each grade and state, then test scores can be rescaled using sample means and SDs
  - Otherwise, the test scores should be rescaled using statewide means and SDs

Recommendations: RCT Design
Calculating multi-grade/state impact estimates

- For large RCTs, meta-analytic techniques are best for combining impact estimates across multiple grades and/or states because meta-analytic methods…
  - explicitly test for variation in effects across grades/states.
  - provide a mechanism to explain variation in effects.
  - allow impact estimates to be pooled (or not) as the results warrant.

Recommendations: RCT Design
Calculating multi-grade/state impact estimates

- Fixed effects meta-analyses…
  - can be implemented by pooling impact estimates, or by pooling individual-level data.
  - may use grade*TRT or state*TRT interactions to test for variation in treatment effects.
  - are most appropriate when the number of grades or states are small (e.g., <10), and results will not be generalized beyond those grades and states.

Recommendations: RCT Design
Calculating multi-grade/state impact estimates

- Random effects meta-analyses…
  - can also be implemented by pooling impact estimates, or by pooling individual-level data in an HLM model.
  - allow TRT effects to vary randomly by grade and/or state.
  - are most appropriate when the number of grades or states are not small (e.g., ≥10), and results will be generalized (e.g., nationwide).

Recommendations: RCT Design
Calculating multi-grade/state impact estimates

- Impacts should be combined only when variation across grades/states can be…
  - explained/predicted through moderator analyses

    OR
  - attributed to random sampling variation

    OR
  - deemed ignorable based on the desire for an impact estimate that is pooled across different sets of state standards

Summary

- Data from state tests can be an efficient and relevant means for evaluating program impacts
- Researchers should consider first the nature of the outcomes posed by the research questions
- Studies involving multiple states and/or grades must address numerous complicated issues in analysis and interpretation
- Assumptions implied by analytical choices must be acknowledged and evaluated