

# VAMs and Errors of Measurement

Michael Kane

National Conference of Bar Examiners

# Some Basic Presumptions

- VAMs have their origins in economics, but it is not clear that they will work as well in education.
- Where VAMs are used for high-stakes decisions, they need to support causal inferences.
- It seems to be taken for granted that VAMs will have no systemic impact on education.

# A Radical Simplification of Evaluation

- by reducing the outcomes of interest to test scores,
- by reducing the analytic approach to linear (OLS) models,
- and by reducing the input variables of interest to those on which we have data

# Causal Inferences

- The goal is to establish a causal connection between the educational unit (school or teacher) and student learning.
- For this kind of causal claim to be tenable, it is necessary to rule out all other plausible causes for the differences.

# Controlling Systematic Errors

- In experimental settings, we control the extraneous factors, by either fixing the conditions of the facets and/or by random selection or assignment.
- An appealing alternative is to introduce statistical adjustments, but such adjustments are notoriously difficult to implement effectively.

# ANCOVA Revisited

- A fallible covariate (e.g., a previous test score) leads to biased adjustments, and “there is no agreement as to what to do about it” (Pedhazur and Smelkin, 1991, p. 582).
- Linn - even if the models work as intended, the results may be too unreliable for high-stakes contexts.
- Reckase – We have relatively little research on the robustness of the models when assumptions are violated.

# Errors of Measurement

- Systematic errors do not cancel out over repeated measurement, and thus add bias.
  - School context
  - Non-random assignment of students
- Random errors decrease precision but do not add bias.
  - Test scores
  - Students
  - It happens

# Random Errors in “Noisy” Systems

- Random errors cancel out over the long term,
- but, as Keynes said, we live in the short term; in the long term, we are all dead,
- and random rewards/punishments can be pretty effective in driving people crazy.

# Interval Scales

- As Linn pointed out, we can sometimes conclude that a scale (e.g., a percentile scale) is not an interval scale
- It is not so clear how to establish that a scale is an interval scale.
- The tests given at different grade levels are designed to cover different content domains and to differ in difficulty and other characteristics.

# Scales that are Interval Scales

- If we line up adjacent inch marks on a ruler with any adjacent inch marks on any other ruler, they line up exactly.
- Over a range of temperatures and materials, it takes the same energy to raise temperature 1 degree.
- In law, dollars are said to be *fungible*, or interchangeable.

# Statistical Criteria for Interval Scales

- Reckase (2008) suggests two ways of arguing that test scores are on an interval scale:
- by using a score scale that is a linear transformation of an IRT theta scale,
- or by showing that the distribution of observed scores matches some assumed theoretical distribution.

# Concluding Comments

- The state tests that tend to be used in value-added models are quite limited as measures of educational outcomes.
- This complexity and the potential for statistical artifacts in VAMs tend to make them less than fully transparent, and this is not a desirable characteristic for high-stakes testing.