Made to be broken:
The paradox of student growth prediction

Andrew Ho, Harvard Graduate School of Education
Frank B. Womer Invitational Keynote
Michigan School Testing Conference
Ann Arbor, Michigan. February 14, 2018
The paradox of student growth prediction

Student growth predictions should be:

a) Accurate.
b) Ultimately, incorrect.
c) Both a) and b).
Supporting Flexibility for States and Districts and Promoting High Expectations for All Students:
Title I, Part A
General Statutory Requirements
- ESSA requires that states establish college-and career-ready standards and maintain high expectations when assessing all students against those standards. These regulations support innovation and flexibility while maintaining a high bar for quality of the tests states use to assess all students against state-developed college- and career-ready expectations.

ACT and SAT College Benchmarks

What do we mean by “a high probability of success”?
Students who meet a Benchmark on the ACT or ACT Compass have approximately a 50 percent chance of earning a B or better and approximately a 75 percent chance of earning a C or better in the corresponding college course or courses. Students who meet a Benchmark on ACT Explore or ACT Plan are likely to have approximately this same chance of earning such a grade in the corresponding college course(s) by the time they graduate high school.
**Predictive accuracy** is not a goal but something to sacrifice toward other goals: transparency, usability, and positive outcomes.

Remember: These models are only valid if they are wrong.

FIGURE 1. Tiers of assessment.
“Although **predictive purposes** are important in high-stakes testing situations, we suspect that **there are few assessment systems where the sole purpose for the system is prediction. Rather, most users want additional information to help them improve the performance of students for whom failure is predicted**” (p. 8).

“... if the test predicts that a student is on track to perform at the basic level, and then appropriate interventions are used to bring the student to proficient, **the statistical analysis of the test’s predictive validity should underpredict student performance over time**” (p. 8).
• Let’s clarify: The model *should* make accurate predictions about what would happen, on average, *had there been business as usual*.

• This is often represented by using data from past cohorts.

\[
.2X_{October} + .4X_{December} + .4X_{February} = \hat{X}_{April}
\]

\[
.2(1640) + .4(1650) + .4(1660) = 1652
\]

• The prediction for any student in April, 2018, is made assuming the same background, conditions, and context as the previous cohort in 2016-2017.

• These older, “reference cohorts” serve as a business-as-usual baseline from which we hope we make improvement.
Here is some clearer counterfactual reasoning:

- If I predict Ed will score a 1652 on his spring test, he should score near 1652, in the absence of the prediction and prediction-based feedback.
- If I predict Ed will score a 1652 on his spring test, he should score significantly higher than 1652, with the prediction and feedback.

This motivates experimental or quasiexperimental designs:

- For example, randomly assign students to prediction-based feedback.
- Few testing organizations developing prediction-based feedback systems are undertaking such rigorous evaluation efforts.

They are unlikely to be successful unless, for every ounce invested in prediction, a pound is invested in the quality of feedback.
What question should we ask prediction-based feedback systems?

• Do you have any evidence that predictions will lead to positive outcomes?

• Your default assumption should be: predictions from interim tests provide negligible additional information beyond the past year’s test scores.

• Lorrie Shepard: “A school board member says we need an early warning system to know which students won't meet the proficiency standard at the end of the year so they can receive extra help.
  – “(This assumes, first and foremost, that a significant majority of teachers does not already know which students are at risk... 
  – “This is an assumption that should be directly tested as part of any validity investigation.” (Shepard, 2009, pp. 35-36)
Possible Outcomes for Prediction-Based Feedback Systems

Ideal (Presumed) Scenario

- Fall Score (Predictor)
- Spring Score (Outcome)

No Predictions

With Predictions

Pass

Fail
Possible Outcomes for Prediction-Based Feedback Systems

Triage Scenario

- No Predictions
- With Predictions

Spring Score (Outcome)

- Pass
- Fail

Fall Score (Predictor)
Possible Outcomes for Prediction-Based Feedback Systems

Inflation Scenario

- **No Predictions**
- **With Predictions**
- **With Predictions, Generalizable Performance**

Spring Score (Outcome) vs. Fall Score (Predictor)
The unfortunate marriage of *prediction* and *usually reasonable psychometric pursuits*.

**Act 1:** Growth reporting on interim assessments

**Act 2:** Standard setting

**Act 3:** Growth-to-standard models for accountability
Act 1) Growth Reporting as Will-o’-the-wisp
Act 1) How should we report & predict growth from interim assessments?

**Linear Fit (Regression on Time)**

**Straight Average (Surely a straw man?)**
Act 1) Finding: Linear fit dramatically underperforms an average score

Magnitude of typical errors in prediction (RMSE)

- Within-grade vertical scales work for narrow purposes, usually at high cost (learning progressions).
- Reporting growth curves on within-grade vertical scales and using them for spring test predictions is folly.

Waldman, 2014
• Coefficient Adjustment
  – Using last year’s data, estimate slopes and intercepts for a) data “to date” and b) all data through the year.
  – Predict coefficients from b) using coefficients from a).
  – Take coefficients from this year “to date” and adjust them through the fitted model.

• Ensemble Model (Hastie et al., 2009)
  – Using last year’s data, find the best combination of a) average scores and b) linear fitted predictions, that predicts last year’s outcome.
  – Use these weights to adjust this year’s coefficients.
  – Think of this method as using the best of the average and the linear fitted predictions.
Act 1) Finding: Alternative models improve prediction

Magnitude of typical errors in prediction (RMSE)

- A simple average of scores to date
- Coefficient adjustment
- Ensemble model

Including all available scores to this date

Waldman, 2014
Act 1) Let predictive models predict

• Clear recommendations for predictive modeling.
  – Don’t use predictions from simple linear extrapolations.
  – Do use prior-year information to inform predictions.
  – Do explore modern analytic methods for prediction.

• But the central question remains:
  – How do we sacrifice the prediction? How do we exceed it?
The unfortunate marriage of *prediction* and *usually reasonable psychometric pursuits*.

**Act 1:** Growth reporting on interim assessments

**Act 2:** Standard setting

**Act 3:** Growth-to-standard models for accountability
Students scoring at this benchmark have “approximately a 50% chance of earning a B or better and approximately a 75% chance or better of earning a C or better in the corresponding college course or courses” (ACT, 2007, p. 24, emphasis added).

Or, the benchmark is “the SAT score associated with a 65% probability of earning a first-year GPA of 2.67 (B-) or higher” (Wyatt, Kobrin, Wiley, Camara, & Proestler, 2011, p. 5, emphasis added).

Problem: A student’s predicted score is a function of a) her actual achievement and b) the predictive utility of the test.

Question: Have you ever seen the predictive utility of tests?
Tests don’t predict college outcomes very well

Freshman GPA vs. Grade 8 Math Scores

n=16K; r=.34
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.

\[ r = 0.9 \]
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.

\[ r = .8 \]
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.
Act 2) What cut score predicts a future B-?

• Correlation between current and future scores (obviously) affects the standard.

$r = .6$

Empirical Standard (81% above)
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.

$r = .2$

![Scatter plot with correlation](image)
Act 2) What cut score predicts a future B-?

- Correlation between current and future scores (obviously) affects the standard.
Act 2) What cut score predicts a future B−?

- Correlation between current and future scores (obviously) affects the standard.
Comparing Standards Across Grades

• Let’s set empirical cut scores for Grades 10, 8, 6, and 4, predicting a college grade of a B-. Say that a B- is the 40th percentile.
• Say correlations are .6, .5, .4, and .2, respectively.
• The percentages of students expected to be on track will be 66%, 69%, 74%, and 90% on track, respectively.
• Lower grades have higher proficiency percentages because the tests aren’t predicting as well, not because students are performing better.

Comparing Standards Across Subjects

• Say we wish to predict a college grade of a B- (30th percentile)
• But the mathematics correlation is .6 whereas the reading correlation is .4.
• Then the percentage of students on track will be 66% for math but 74% for reading.
• Reading has higher proficiency percentages because the test isn’t predicting as well, not because students are performing better.
• Standard setting can feel capricious.
• Predictive standard setting appears to solve the problem.
• But, at best, predictive standard setting imparts a false sense of empirical certitude.
• At worst, it leads to uneven stringency across grades and subjects.

• Instead, 1) set standards as usual, 2) articulate them across grades (smooth patterns of proficiency percentages across grades), then...
• 3) express the implications of cut scores empirically, in terms of $X\%$ chance of achieving $Y$ future score.
• Don’t let the prediction drive the standard setting.
• Afterwards, ask, again: How do we sacrifice the prediction? How do we exceed it?
The unfortunate marriage of *prediction* and *usually reasonable psychometric pursuits*.

**Act 1:** Growth reporting on interim assessments

**Act 2:** Standard setting

**Act 3:** Growth-to-standard models for accountability
Act 3) A surprising consensus

Why hasn’t there been more attention paid to principles around which there is considerable consensus?

1) Multiple measures
2) Achievable targets
3) Moderate stakes, that are...
4) Commensurate with supports
Act 3) How predictive accuracy can create unachievable targets

Three students with equal predictions from a linear trajectory model.

The same three students’ predictions from a regression-based prediction model.

Three students with equal predictions from a regression-based prediction model.
Act 3) Regression models dominate in accuracy

From Ho (2014): The error of the trajectory model is 1.5 to 2 times that of the regression model, for $k = 2 \ldots 5$ available years of data for prediction, and common intergrade correlations shown.
Three students with equal predictions from a regression-based prediction model.

- Requiring students to empirically overcome initial low scores leads to unrealistically high targets.
- Also leads to absurdly low targets for high scorers.
- Trajectory models are poor predictors but are a reasonable basis for setting aspirational, achievable targets.
The unfortunate marriage of *prediction* and *usually reasonable psychometric pursuits*.

Act 1: Growth reporting on interim assessments

Act 2: Standard setting

Act 3: Growth-to-standard models for accountability
The paradox of student growth prediction

*Predictive accuracy* is not a goal but something to sacrifice toward other goals: transparency, usability, and positive outcomes.

...and more on the theories, practices, tools, policies, and incentives that will lengthen *this* arrow.

Remember: *These models are only valid if they are wrong.*