A Validity Agenda for Growth Models: One size doesn't fit all!

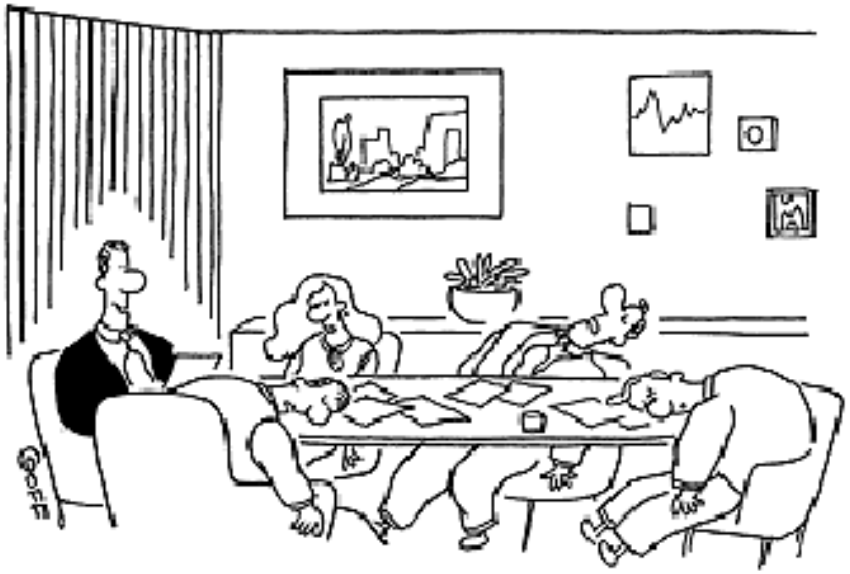
Thanos Patelis
The College Board

Presentation at the Joint Business Meeting of the
AERA Special Interest Groups on
Test Validity Research and Evaluation,
Large Scale Assessment,
Advanced Study of National Databases, and
Computer and Internet Applications in Education
April 14, 2012
Vancouver, British Columbia, Canada

CollegeBoard
inspiring minds™

My goals today…

- Express my appreciation
- Try to say something that would resonant with each SIG represented here.
- Keep it short
- Try to pull together some excellent work and thinking done by many others (including many of you all)
- Offer a practitioner's perspective → validation practices
- Try to make it entertaining
- Try to convince you that there isn't one growth model, there is no silver bullet, and you need a research plan.
- Call to arms
- Bribe you with drinks if you clap and say nice things
- Hope you don't throw me out and revoke my membership



"At last we've reached a consensus! This meeting is boring!"

My takeaway statements

1. Assessments, National Databases, Validity Work, and Technology are components that cannot be separated and must exist to do this right.
2. There are a number of components that must be in place or done when introducing growth models in a large scale setting (e.g., states).
   - Growth models should be built by design, but we all are faced with the reality of needing to retrofit.
3. It is our professional responsibility to make sure that there is evidence to support the claims being made.
4. A short- and long-term validity agenda is needed to permit the collection of evidence across all components of the growth model.
   - State claims to be made
   - Evaluate whether the claims can be supported
5. Because there exist a variety of goals and diverse contexts, there is no one growth model.
6. Because there exist a variety of goals, growth models embedded within diverse contexts, the design of the validity agenda will be personalized and end up being similarly varied.
7. Don't forget the report! In fact start with the report as an expression of the claims and validate.
8. There's a lot of work that needs to be done and it's hard work. The targets of the results of growth models (children, teachers, administrators) need your thinking and your efforts.

The Logic of a Joint SIG Business Meeting

Assessments

Data & Databases

Validity

Technology

Standards for Educational and Psychological Testing

Standard 1.1: A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.

--(AERA, APA, & NCME, 1999, p. 17)

Standards for Educational and Psychological Testing

Standard 1.4: If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.

-- (AERA, APA, & NCME, 1999, p. 18)

Components in Introducing Growth Models

Build it by design or retrofit carefully

1. Specify purpose→ Claims
2. Audience
3. Examine or Build Alignment
4. Scale development
5. Time frame
6. Longitudinal data
7. The model
8. Validity evidence
9. Examination of the use (intended and unintended) utility, and impact of the information

Recommend *Implementer's Guide to Growth Models* (2008) published by CCSSO

Auty et al., 2008; Gong, 2010; O'Malley et al., 2011; Patelis et al., 2012.

Examples of validity studies

- There's been activity to generate validate evidence around growth models.
- Even RFP's are asking for evidence in support of them..
- Some examples of work done:

1. Examination of validity of the growth model in North Carolina's accountability system (Brown, 2008).
2. Examination of the validity of the *Insight* growth model developed by the Pioneer Regional Educational Service Agency in 13 school systems in Georgia by gathering evidence of the utility and impact of the information provided through interviews and document reviews (Crane, 2011).
3. Examination of the validity of a growth model using the California Standards Test in a large urban school district. (Horner, 2009).
4. Examination of the validity of college readiness and steps towards college readiness (Camara, 2011).

A Validity Model for Tests

- Creating some heated debate among scholars, a model was proposed for talking about and examining validity.
- Regardless of the position you take, putting aside all the valid arguments in favor and against the terminology (e.g., Sireci, 2007; Gorin, 2007), and suspending any judgments on what type of evidence is more or less important, this model offered a practical framework that could be applied to growth models.

	Focus	
Perspective	Theoretical	Practical
Internal	Latent Process	Content Validity & Reliability
External	Nomological Networks	Utility & Impact

Source: Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity in education. *Educational Researcher, 36,*(8) 437-448.

Another Validity Model for Tests

- Another model has been proposed that deepened the framework and specified internal and external sources of evidence.

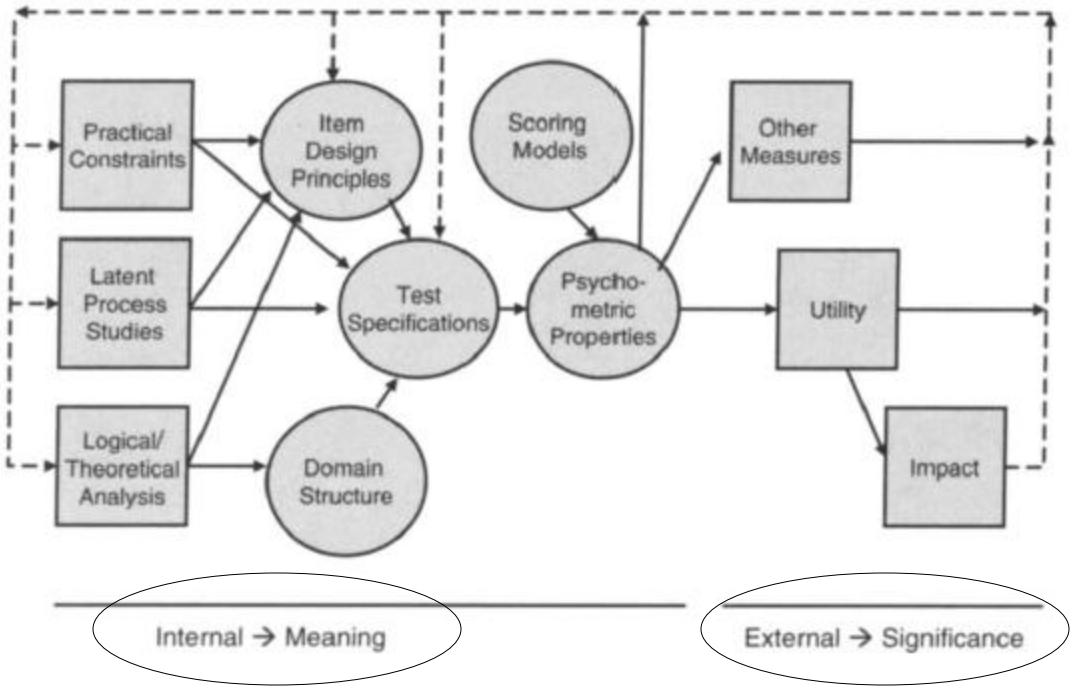


FIGURE 2. *A universal system for validity.*

Source: Embretson, S. (2007). Construct validity: A universal validity system or just another test evaluation procedure. *Educational Researcher, 36*, 449-455.

Suggestions from Michael Kane…

- In a recent lecture in Cambridge, Kane said:

The argument-based framework is quite simple and involves two steps. First, specify the proposed interpretations and uses of the scores in some detail. Second, evaluate the overall plausibility of the proposed interpretations and uses.

The argument-based framework is quite flexible in the sense that it does not specify any particular kind of interpretation or use for assessment scores, and invites assessment developers and users to specify their proposed interpretations and uses. Any kind of interpretation or use can be proposed, but the claims being made should be justified, and more ambitious interpretations and uses impose more demands for justification.

---- Kane (2011, p. 4)

Validity Framework for Growth Models in Teacher Accountability

- A set of propositions and associated claims have been proposed that must be supported by evidence for using growth models for teacher accountability.

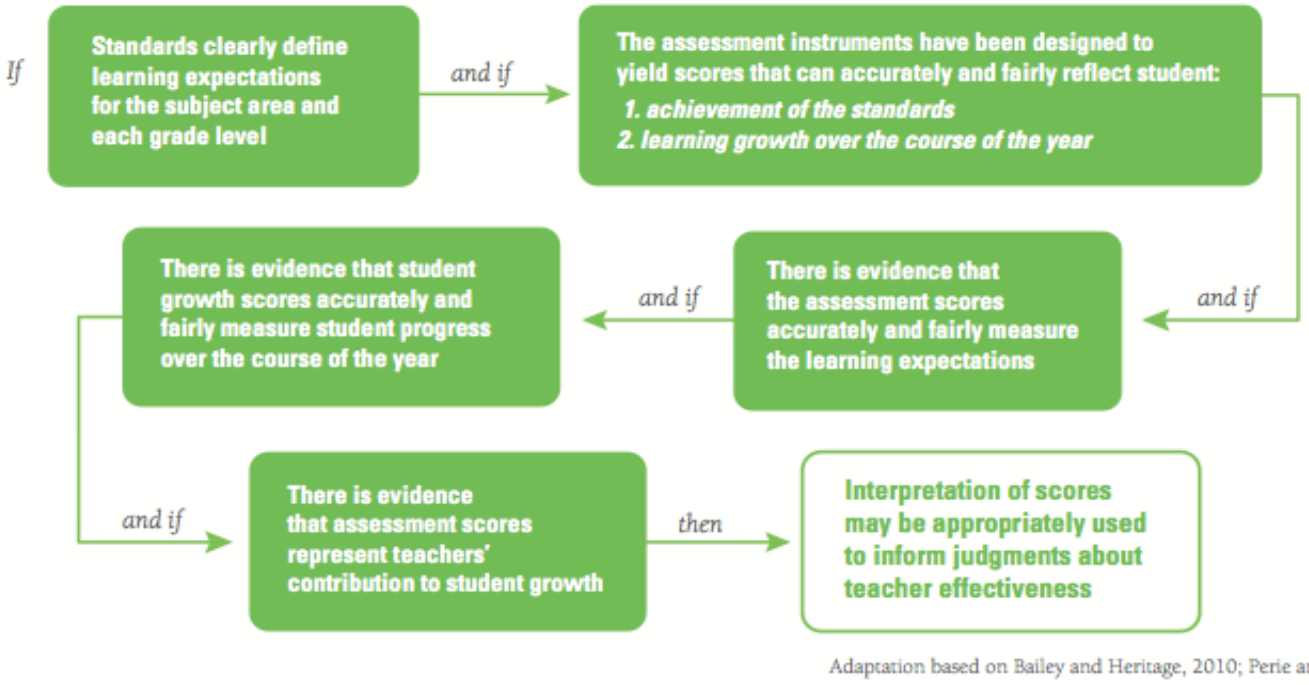


Figure 1. Propositions that justify the use of these measures for evaluating teacher effectiveness.

Source: Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Developing and Selecting Assessment of Student Growth for Use in Teacher Evaluation Systems. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Actionable Framework…

- Framework
- Articulates Claims
- Suggests Evidence.

Table 1. Propositions and Claims Critical to the Validity Evaluation.

Proposition 1 - Standards clearly define learning expectations for the subject area and each grade level

Design Claims:	Evidence
• Learning expectations are clear • Learning expectations are realistic • Learning expectations reflect a progression (at minimum for the span of a grade level)	• Expert reviews

Proposition 2a - The assessment instruments have been designed to yield scores that can accurately and fairly reflect student achievement of the standards

Design Claims:	Evidence
• Specifications/blueprint for assessment reflect the breadth and depth of learning expectations • Assessment items and tasks are consistent with the specifications and comprehensively reflect learning expectations • Assessment design, administration, and scoring procedures are likely to produce reliable results • Assessment tasks and items are designed to be accessible and fair for all students	• Expert reviews of alignment • Measurement review of administration and scoring procedures • Sensitivity reviews

Proposition 2b - The assessment instruments have been designed to yield scores that can accurately and fairly reflect student learning growth over the course of the year

Design Claims:	Evidence
• Assessments are designed to accurately measure the growth of individual students from the start to the end of the school year • Cut scores for defining proficiency levels and adequate progress, if relevant, are justifiable • Assessments are designed to be sensitive to instruction	• Expert reviews • Research studies

Proposition 3 - There is evidence that the assessment scores accurately and fairly measure the learning expectations

Psychometric Claims:	Evidence
• Psychometric analyses are consistent with/confirm the assessment's learning specifications/blueprint • Scores are sufficiently precise and reliable • Scores are fair/unbiased	• Psychometric analyses • Content analysis

Proposition 4 - There is evidence that student growth scores accurately and fairly measure student progress over the course of the year

Psychometric Claims:	Evidence
• Score scale reflects the full distribution of where students may start and end the year • Growth scores are sufficiently precise and reliable for all students • Growth scores are fair/relatively free of bias • Cut points for adequate student progress are justified	• Psychometric modeling and fit statistics • Sensitivity/bias analyses

Proposition 5 - There is evidence that scores represent individual teachers' contribution to student growth

Psychometric Claims:	Evidence
• Scores are instructionally sensitive • Scores representing teacher contribution are sufficiently precise and reliable • Scores representing teachers contributions are relatively free of bias	• Research studies on instructional sensitivity • Precision and stability metrics • Advanced statistical tests of modeling alternatives and tenability of assumptions

Based on Herman & Choi, 2010

Source: Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Developing and Selecting Assessment of Student Growth for Use in Teacher Evaluation Systems. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Example from Framework

Proposition	Claim	Evidence
Proposition 2b - The assessment instruments have been designed to yield scores that can accurately and fairly reflect student learning growth over the course of the year	• Assessments are designed to accurately measure the growth of individual students from the start to the end of the school year • Cut scores for defining proficiency levels and adequate progress, if relevant, are justifiable • Assessments are designed to be sensitive to instruction	• Expert reviews • Research studies

Source: Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Developing and Selecting Assessment of Student Growth for Use in Teacher Evaluation Systems. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Another Example from Framework

Proposition	Claim	Evidence
Proposition 4 - There is evidence that student growth scores accurately and fairly measure student progress over the course of the year	• Score scale reflects the full distribution of where students may start and end the year • Growth scores are sufficiently precise and reliable for all students • Growth scores are fair/relatively free of bias • Cut points for adequate student progress are justified	• Psychometric modeling and fit statistics • Sensitivity/bias analyses

Source: Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Developing and Selecting Assessment of Student Growth for Use in Teacher Evaluation Systems. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Components in Introducing Growth Models

Build it by design or retrofit carefully

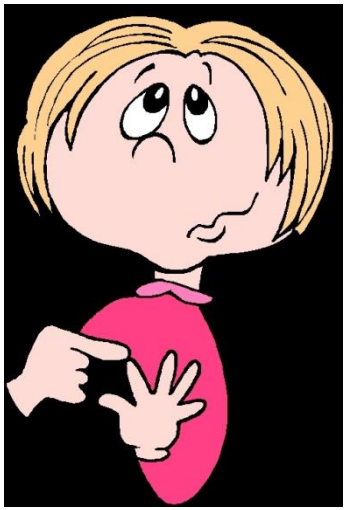
1. Specify purpose→ Claims
2. Audience
3. Examine or Build Alignment
4. Scale development
5. Time frame
6. Longitudinal data
7. The model
8. Validity evidence
9. Examination of the use (intended and unintended) utility, and impact of the information

CollegeBoard
inspiring minds™

The State of Affairs with Growth Models

- In a survey of states in early part of 2010, 95% of the 43 that responded indicated that they have implemented, planning to, or considering growth models.
- The reported goals of the growth models were as follows:

Purpose of the Growth Model	No.	%
Information on School and Student Achievement	37	25%
Accountability	27	18%
Identifying Successful School Improvement Strategies	20	13%
Instructional Support	18	12%
Program Evaluation	17	11%
Recognition of Schools	14	9%
Teacher Effectiveness (link to students)	13	9%
Financial Incentives	4	3%
Total Responses from 43 States:	150	



Source: Blank, R. K. (2010). *State Growth Models for School Accountability: Progress on Developing and Reporting Measures of Student Growth*. Washington, DC: Council of Chief State School Officers.

The State of Affairs with Growth Models (cont'd)

- The CCSSO reviewed and analyzed state web-based reporting on growth models for 22 states.

Component	Summary
1. Purposes	Many (previous slide)
2. Audience	Administrators, Teachers, Parents, Public
3. Alignment	--
4. Scale Dev.	--
5. Time Frame	Grades 3-8 or 4-8
6. Long. Data	2-3 years of data (unclear if longitudinal)
7. Model	Many (VAMs, Transition, Projection, SGPs, etc)
8. Validity	--
9. Use, Utility, Impact	--

Source: Blank, R. K. (2010). *State Growth Models for School Accountability: Progress on Developing and Reporting Measures of Student Growth*. Washington, DC: Council of Chief State School Officers.

Overview of a Validity Agenda

The word "valid" is derived from the Latin validus, meaning strong

Component	Description	Type of Evidence
1. Purpose	Specification of the purpose of the growth model and the types of claims that will be made.	■ Evaluate and document whether goals are understood via surveys, focus groups, and /or interviews.
2. Audience	Clear indication of who the audience and type of information that they will receive	■ Document review. ■ Interviews.
3. Alignment	Since claims will be made over time across assessments, the content alignment must be examined across assessments and to standards associated with claims to be made	■ Alignment studies across tests and to standards. ■ Performance Level Descriptors and process for developing them. ■ Learning/skill progressions across tests.

Overview of a Validity Agenda (cont'd)

Component	Description	Type of Evidence
4. Scale Development	Scale metric must provide reliable and valid information at each testing time and across testing times. Type of linkage across tests must be articulated and evaluated.	Variety of psychometric methods. Review of linking design, methodology, and results.
5. Time Frame	Clear indication of when testing will occur and the sequence of tests.	Documentation
6. Longitudinal Data	Capturing data on students over time with links to other data as needed.	Statistical analyses including examination of missing data. Extent to which recommendations of DQC are met.

Overview of a Validity Agenda (cont'd)

Component	Description	Type of Evidence
7. Model	Selection of model that matches claims to be made.	Variety of research studies. Evaluation of standard setting procedures (if applicable)
8. Validity	Evaluation of the claims being made.	Documentation of each component and evidence indicated for each. Studies after implementation to evaluate whether claims can be supported
9. Use, utility, impact	Examination how information from growth models are being using and impacting the target audience.	Utilize program evaluation methods to gather evidence. Use national databases. Implement surveys, interviews, and/or focus groups

Some specific thoughts….

- Camara (2011) has offered the following comments about gathering evidence to support college readiness statements (a type of growth model):
  - Task is to make inferences about high school students readiness for postsecondary education (college, workplace training).
  - Difficulty is that we often are making these inferences 2, 3, or 4 years in advance.
  - Predicting future academic behaviors
  - Suggestion: Back map or sequence postsecondary proficiencies (KSAs) to establish a trajectory of skill acquisition.
    - Caution: Individual differences, contextual differences.
- **A validity argument depends on more than one proposition.** Strong evidence in support of one does not diminish the need for evidence to support other propositions.
- A few lines of very solid evidence regarding a proposition are better than numerous lines of evidence of questionable quality.
- Interpretation of results should be based on multiple sources of convergent and collateral data (and understanding of normative, empirical and theoretical foundations).

Two words on data…

#1 DATAQUALITY CAMPAIGN
Using Data To Improve Student Achievement
www.DataQualityCampaign.org

	DQC: 10 Essential Elements	2010	2011
K-12 only	(1) a unique statewide student identifier that connects student data across key databases across years	52 states	52 states
K-12 only	(2) student-level enrollment, demographic & program participation information	52 states	52 states
K-12 only	(8) student-level graduation & dropout data	52 states	52 states
K-12 only	(9) the ability to match student records between the p-12 & higher education systems	41 states	49 states
K-12 only	(10) a state data audit system assessing data quality, validity & reliability	52 states	52 states
K-12 only	(3) the ability to match individual students' test records from year to year to measure academic growth	52 states	52 states
K-12 only	(4) information on untested students & the reasons they were not tested	49 states	51 states
K-12 only	(5) a teacher identifier system with the ability to match teachers to students	35 states	44 states
K-12 only	(6) student-level transcript information, including information on courses completed & grades earned	37 states	41 states
K-12 only	(7) student-level college readiness test scores	46 states	50 states

	COMPETES Act: Required Elements	2010
P-12 & postsecondary education	(1) a unique statewide student identifier that does not permit a student to be individually identified by users of the system	43 states
P-12 & postsecondary education	(2) student-level enrollment, demographic, & program participation information	45 states
P-12 & postsecondary education	(3) student-level information about the points at which students exit, transfer in, transfer out, drop out, or complete P-16 education programs	36 states
P-12 & postsecondary education	(4) the capacity to communicate with higher education data systems	33 states
P-12 & postsecondary education	(5) a State data audit system assessing data quality, validity, & reliability	48 states
P-12 only	(6) yearly test records of individual students with respect to ESEA assessments	49 states
P-12 only	(7) information on students not tested by grade & subject	49 states
P-12 only	(8) a teacher identifier system with the ability to match teachers to students	30 states
P-12 only	(9) student-level transcript information, including information on courses completed & grades earned	28 states
P-12 only	(10) student-level college readiness test scores	40 states
Postsecondary only	(11) information regarding the extent to which students transition successfully from secondary school to postsecondary education, including whether students enroll in remedial coursework	28 states
Postsecondary only	(12) other information determined necessary to address alignment & adequate preparation for success in postsecondary education	29 states

http://dataqualitycampaign.org/

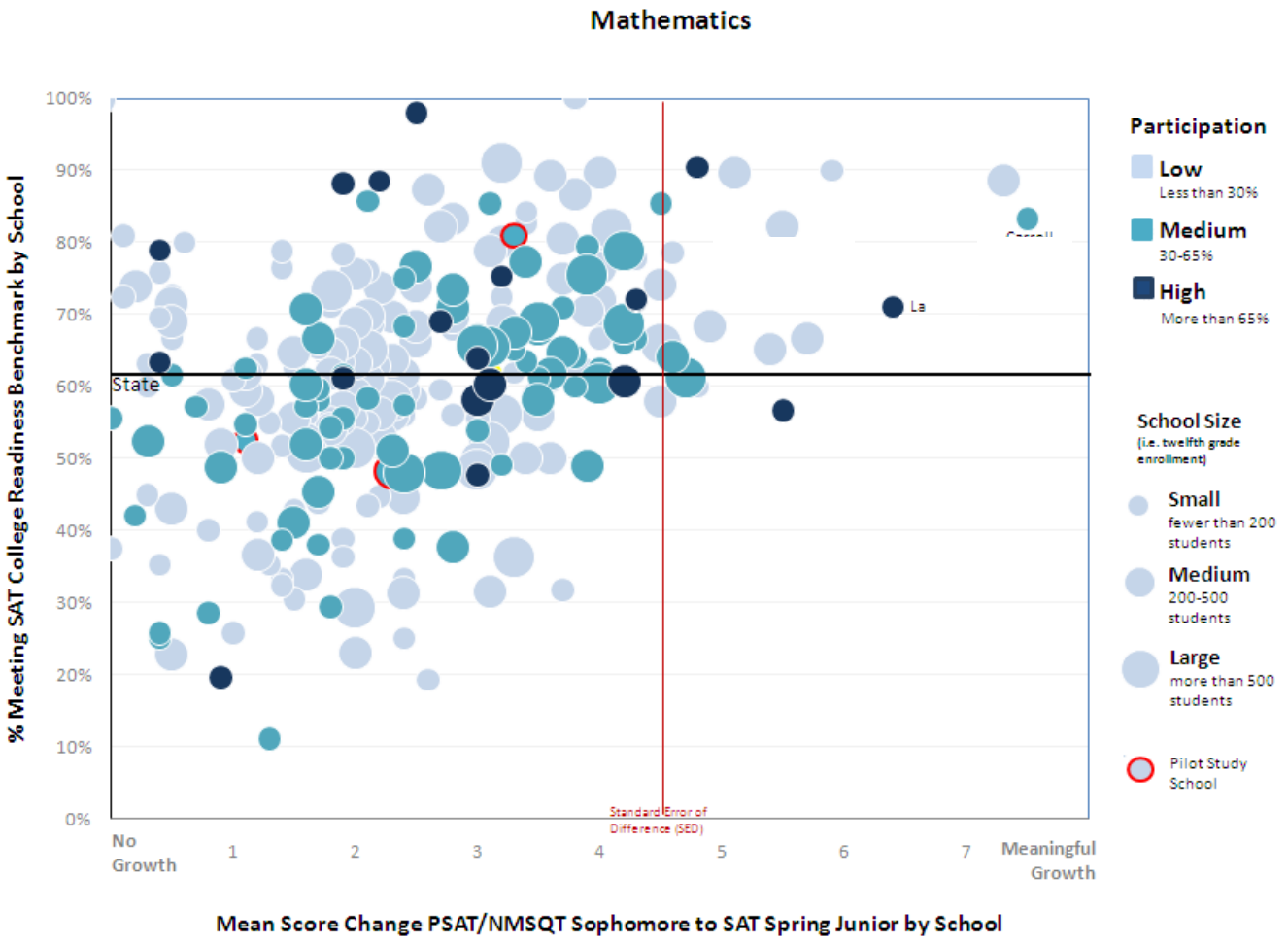
#2 Can look to national data sources!
Work to influence of state-wide and national databases.
Ensure access is available for research.

CollegeBoard inspiring minds™

Comments on Reporting…

- At the end of the day, the targeted audience only has the reports as the results of any testing or growth modeling.
- This is the deliverable and not the growth model or the alignment studies.
- This is where technology and the use of powerful visualizations of the information can make a difference.
- CCSSO has offered some recommendations on this, see Auty et al (2008).
- Some experts are working on the science of score reporting (Ron Hambleton, John Hattie, Sandip Sinharay, Krista Breithaupt, Joe Ryan, Patelis & Matos-Elefonte) and its role in making the validity agenda concrete and focused.
- States are turning to web applications to display not only student-level results but also aggregate results. The web and good technological solutions offer a means to drill down offering more information making reports actionable.
- Comment: The report is the expression of the claims.
  - Whatever is on the report should be the object of validation.
  - Start with the report, represent the claims that you want to make, and gather evidence to support the claims on there.

Reporting – Whatever is on here should be validated!



Research Questions:

- What does the score change mean?
- How big of a change is meaningful?
- Is there evidence to support the benchmark?
- Is this characteristic of the school or only a subset of students?
- Will school values bounce around?
- What does a school with little change and small percentage of students at the benchmark mean?
- Is participation rate an important feature?
- Is school size an important feature?
- Will the target audience make correct inferences?
- What inferences can be made that were not intended?
- How absolute are those cut-points represented by the reference lines?

CollegeBoard
inspiring minds™

My takeaway statements

1. Assessments, National Databases, Validity Work, and Technology are components that cannot be separated and must exist to do this right.
2. There are a number of components that must be in place or done when introducing growth models in a large scale setting (e.g., states).
   - Growth models should be built by design, but we all are faced with the reality of needing to retrofit.
3. It is our professional responsibility to make sure that there is evidence to support the claims being made.
4. A short- and long-term validity agenda is needed to permit the collection of evidence across all components of the growth model.
   - State claims to be made
   - Evaluate whether the claims can be supported
5. Because there exist a variety of goals and diverse contexts, there is no one growth model.
6. Because there exist a variety of goals, growth models embedded within diverse contexts, the design of the validity agenda will be personalized and end up being similarly varied.
7. Don't forget the report! In fact start with the report as an expression of the claims and validate.
8. There's a lot of work that needs to be done and it's hard work. The targets of the results of growth models (children, teachers, administrators) need your thinking and your efforts.

References

Auty, W., et al. (2008). *Implementer's Guide to Growth Models*. Washington, DC: Council of Chief State School Officers. http://www.ccsso.org/Documents/2008/Implementers_Guide_to_Growth_2008.pdf

Blank, R. K. (2010). State Growth Models for School Accountability: Progress on Developing and Reporting Measures of Student Growth. Washington, DC: Council of Chief State School Officers. http://www.ccsso.org/Documents/2010/State_Growth_Models_for%20School_2010.pdf

Brown, K. T. (2008). Testing the Testing: Validity of a State Growth Model. *International Journal of Education Policy and Leadership, 3(*6). Retrieved November 2011 from http://www.ijepl.org.

Camara, W. J. (June, 2011). *College & Career Readiness: An Initial Validation Argument.* Presentation at CCSSO's National Conference on Student Assessment, Orlando, FL.

Crane, E. W. (2011). Validity study of Pioneer RESA's Insight growth model: Final Report. http://www.pioneerresa.org/Insight%20WestEd%20Report%20Final%2005-20-11.pdf

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher*, *36*(8), 449-455.

Gong, B. (June, 2010). *Using Growth Data to Improve Learning, Teaching, and School Functioning*. Presentation at the CCSSO National Conference on Student Assessment, Detroit, MI. http://www.nciea.org/publications/CCSSO_BG2010.pdf

Gorin, J. S. (2007). Reconsidering Issues in Validity Theory. *Educational Researcher*, *36*(8), 456-462.

References (cont'd)

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). *Developing and Selecting Assessment of Student Growth for Use in Teacher Evaluation Systems.* Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Horner, M. (2009). *Quantifying student growth analysis of the validity of applying growth modeling to the California Standards Test.* Dissertation at the University of Southern California. http://digitallibrary.usc.edu/assetserver/controller/item/etd-Horner-2920.pdf

Kane, M. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1) 3–17.

O'Malley, K. J., McClarty, K. L, Murphy, D., & McBride, Y. (2011). *Overview of Student Growth Models. Pearson.* http://www.pearsonassessments.com/hai/Images/tmrs/Student_Growth_WP_083111_FINAL.pdf

Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity in education. *Educational Researcher, 36*(8), 437-448.

Patelis, T., Barry, C., Bausmith, J., & Matos-Elefonte, H. (2012). *Considerations in Growth Models and Reporting*. White Paper. New York, NY: The College Board.

Sireci, S.G. (2007). On Validity Theory and Test Validation. *Educational Researcher*, *36*(8),477-481.

Questions, Comments, Suggestions

- Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board presentations do not necessarily represent official College Board position or policy.
- Please forward any questions, comments, and suggestions to: Thanos Patelis tpatelis@collegeboard.org or 212-649-8435
- Please go the College Board's web site for much more information and this presentation: www.collegeboard.org/research.

Thank you!!