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
My goals today…

- Express my appreciation
- Try to say something that would resonate 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
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

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).
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.

<table>
<thead>
<tr>
<th>Perspective</th>
<th>Theoretical</th>
<th>Practical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>Latent Process</td>
<td>Content Validity &amp; Reliability</td>
</tr>
<tr>
<td>External</td>
<td>Nomological Networks</td>
<td>Utility &amp; Impact</td>
</tr>
</tbody>
</table>

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

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

![Validity Framework for Growth Models in Teacher Accountability](image)

Actionable Framework...

• Framework
• Articulates Claims
• Suggests Evidence.

### Example from Framework

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Claim</th>
<th>Evidence</th>
</tr>
</thead>
</table>
| 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 |

Another Example from Framework

<table>
<thead>
<tr>
<th>Proposition</th>
<th>Claim</th>
<th>Evidence</th>
</tr>
</thead>
</table>
| 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 |
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
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:

<table>
<thead>
<tr>
<th>Purpose of the Growth Model</th>
<th>No.</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information on School and Student Achievement</td>
<td>37</td>
<td>25%</td>
</tr>
<tr>
<td>Accountability</td>
<td>27</td>
<td>18%</td>
</tr>
<tr>
<td>Identifying Successful School Improvement Strategies</td>
<td>20</td>
<td>13%</td>
</tr>
<tr>
<td>Instructional Support</td>
<td>18</td>
<td>12%</td>
</tr>
<tr>
<td>Program Evaluation</td>
<td>17</td>
<td>11%</td>
</tr>
<tr>
<td>Recognition of Schools</td>
<td>14</td>
<td>9%</td>
</tr>
<tr>
<td>Teacher Effectiveness (link to students)</td>
<td>13</td>
<td>9%</td>
</tr>
<tr>
<td>Financial Incentives</td>
<td>4</td>
<td>3%</td>
</tr>
</tbody>
</table>

Total Responses from 43 States: 150

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

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

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

Overview of a Validity Agenda

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

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Type of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Purpose</td>
<td>Specification of the purpose of the growth model and the types of claims that will be made.</td>
<td>Evaluate and document whether goals are understood via surveys, focus groups, and/or interviews.</td>
</tr>
<tr>
<td>2. Audience</td>
<td>Clear indication of who the audience and type of information that they will receive</td>
<td>Document review. Interviews.</td>
</tr>
<tr>
<td>3. Alignment</td>
<td>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</td>
<td>Alignment studies across tests and to standards. Performance Level Descriptors and process for developing them. Learning/skill progressions across tests.</td>
</tr>
</tbody>
</table>
### Overview of a Validity Agenda (cont’d)

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
<th>Type of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Scale Development</td>
<td>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.</td>
<td>Variety of psychometric methods. Review of linking design, methodology, and results.</td>
</tr>
<tr>
<td>5. Time Frame</td>
<td>Clear indication of when testing will occur and the sequence of tests.</td>
<td>Documentation</td>
</tr>
<tr>
<td>6. Longitudinal Data</td>
<td>Capturing data on students over time with links to other data as needed.</td>
<td>Statistical analyses including examination of missing data. Extent to which recommendations of DQC are met.</td>
</tr>
<tr>
<td>Component</td>
<td>Description</td>
<td>Type of Evidence</td>
</tr>
<tr>
<td>----------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>7. Model</td>
<td>Selection of model that matches claims to be made.</td>
<td>Variety of research studies. Evaluation of standard setting procedures (if applicable)</td>
</tr>
<tr>
<td>8. Validity</td>
<td>Evaluation of the claims being made.</td>
<td>Documentation of each component and evidence indicated for each. Studies after implementation to evaluate whether claims can be supported</td>
</tr>
<tr>
<td>9. Use, utility, impact</td>
<td>Examination how information from growth models are being used and impacting the target audience.</td>
<td>Utilize program evaluation methods to gather evidence. Use national databases. Implement surveys, interviews, and/or focus groups</td>
</tr>
</tbody>
</table>
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...

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.
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?
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


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


References (cont’d)


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!!