Validity Issues for International Large Scale Assessments: 
*Truth and Consequences*

MICHAEL J. FEUER
TEACHERS COLLEGE/ETS
MARCH 2012

THE GEORGE WASHINGTON UNIVERSITY
GRADUATE SCHOOL OF EDUCATION
AND HUMAN DEVELOPMENT
The Leaning Tower...
The news hits

2003—

- We are not just slipping, we are in serious trouble
- Major reforms are needed!
- Our standing in the world is at risk!!
PISA shock!

“… PISA results … fundamentally undermined the implicit self-assessment of the country's education system, an assessment that had been characterized by self-confidence, belief in its efficiency and the system's important role in Germany’s economic achievements…” (Ertl)
Block that metaphor!

“Observers have compared the consequences of PISA on education in Germany with the Sputnik shock (Ostermann, 2005) or even with the French Revolution (Herrmann, 2004) and have argued that PISA has had a more far-reaching impact on German education than A Nation at Risk had on education in the USA (Gruber, 2006).”

(Ertl, 2006)
Sound familiar?

“The end is near for American education” (Turner; Huffington Post, October 2011);
“Your child left behind” (Ripley; The Atlantic, December 2010);
“The failure of American schools” (Klein; The Atlantic, June 2011);
“Why can’t American students compete?” (Hanushek & Peterson; Newsweek, August 2011);
”'Wake up call': US students trail global leaders” (Armario; MSNBC, December 2010)

“if an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war…” Nation at Risk, 1983
“In striking contrast to the international acclaim during the 1990s for Japanese schools’ instructional excellence and solid curricular contents, the Japanese media, scholars, politicians, and the public continued to perceive their country’s schooling as steeped in a dire crisis … [and the public debate] framed education reform as having direct bearing on the nation’s rise and fall, evoking a sense of urgency for immediate intervention …” (Takayama, 2007).
Why the angst?

PISA Math, Selected Countries, 2006 and 2009

2006 2009

Int'l avg 494 496 504 513 442 447 523 529 547 546 474 487
Germany 504 513 504 513 504 513 504 513 504 513 504 513
Israel 442 447 447 447 447 447 447 447 447 447 447 447
Japan 523 529 529 529 529 529 529 529 529 529 529 529
Korea 547 546 546 546 546 546 546 546 546 546 546 546
USA 474 487 487 487 487 487 487 487 487 487 487 487
The three r’s: revulsion, reevaluation, reform

“as early as July 2003 blueprints of national standards for German, Mathematics and the first foreign language (English or French) for the end of Realschule education …were introduced. A revised version of the standards was passed …in December 2003. All Lander agreed to introduce the national standards at the beginning of the 2004/05 school year as a basis for the development of new curricula and for initial and continuous teacher training…” (Ertl, 2006).
“PISA” and “TIMSS” in English language Publications, 1995-2008

Via Google

* [http://books.google.com/ngrams/info](http://books.google.com/ngrams/info)

With thanks to Jim Williams and Laura Engle
“PISA” in books published in German, 1995-2008
Logic Model

US economy at risk from global rivals

As shown by slips in productivity growth and other indicators

American workers won’t be able to compete, especially at reasonable wages

Economic decline parallels poor performance on ILSA

Lousy schools lead to ill-prepared workforce

American workers won’t be able to compete, especially at reasonable wages

Economic decline parallels poor performance on ILSA

Lousy schools lead to ill-prepared workforce

US economy at risk from global rivals
Or in abridged form

How we do as a society…

Depends on our international economic standing…

Which depends on our international educational standing
So it pays to check the numbers

For example:

Or these

Total Productivity Growth Rates, Selected Countries, 1986-2004

Source: Lach, Shil, Trajtenberg, 2008
Or these

Average annual Productivity Growth, Selected Countries, 1870-1979

<table>
<thead>
<tr>
<th>Country</th>
<th>1870-80</th>
<th>1900-13</th>
<th>1929-38</th>
<th>1950-60</th>
<th>1970-79</th>
<th>% change from previous decade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>2.2</td>
<td>2.7</td>
<td>0</td>
<td>3.1</td>
<td>1.8</td>
<td>-41.9</td>
</tr>
<tr>
<td>Finland</td>
<td>1.3</td>
<td>2.4</td>
<td>1.9</td>
<td>4</td>
<td>2.6</td>
<td>-45.2</td>
</tr>
<tr>
<td>France</td>
<td>2.3</td>
<td>1.8</td>
<td>2.8</td>
<td>4.4</td>
<td>4.1</td>
<td>-9.7</td>
</tr>
<tr>
<td>Germany</td>
<td>1.5</td>
<td>1.4</td>
<td>2.3</td>
<td>6.6</td>
<td>4.5</td>
<td>-67.7</td>
</tr>
<tr>
<td>UK</td>
<td>1.6</td>
<td>0.9</td>
<td>0.9</td>
<td>2.2</td>
<td>2.8</td>
<td>19.4</td>
</tr>
<tr>
<td>USA</td>
<td>2.3</td>
<td>2</td>
<td>0.7</td>
<td>2.4</td>
<td>1.9</td>
<td>-16.1</td>
</tr>
</tbody>
</table>

Source: Baumol, et al, 1989
So much for the golden age...

**NAEP trends, mathematics, 1973-2008**

<table>
<thead>
<tr>
<th>Age group</th>
<th>Whites</th>
<th>Blacks</th>
<th>Hispanics</th>
<th>Average change</th>
</tr>
</thead>
<tbody>
<tr>
<td>9-year olds</td>
<td>25 points</td>
<td>34</td>
<td>32</td>
<td>+30</td>
</tr>
<tr>
<td>13-year olds</td>
<td>16 points</td>
<td>34</td>
<td>29</td>
<td>+26</td>
</tr>
<tr>
<td>17-year olds</td>
<td>4 points</td>
<td>17</td>
<td>16</td>
<td>+12</td>
</tr>
</tbody>
</table>

With thanks to Hal Salzman
And back to **P I S A**

![Bar chart showing average math scores for different countries and years](chart.png)

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>527</td>
<td>527</td>
</tr>
<tr>
<td>Finland</td>
<td>548</td>
<td>541</td>
</tr>
<tr>
<td>France</td>
<td>496</td>
<td>497</td>
</tr>
<tr>
<td>Germany</td>
<td>504</td>
<td>513</td>
</tr>
<tr>
<td>Israel</td>
<td>442</td>
<td>447</td>
</tr>
<tr>
<td>Korea</td>
<td>547</td>
<td>546</td>
</tr>
<tr>
<td>UK</td>
<td>495</td>
<td>492</td>
</tr>
<tr>
<td>US</td>
<td>474</td>
<td>487</td>
</tr>
<tr>
<td>Shanghai</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>Selected avg</td>
<td>504</td>
<td>517</td>
</tr>
<tr>
<td>Int'l avg</td>
<td>494</td>
<td>496</td>
</tr>
</tbody>
</table>
Which, by the way, brought on a new strain of PPSS*

*post PISA stress syndrome
Comparing our “best” to the world

• In science at grade 4, Singapore was the only education system with a higher average score or with a higher percentage of students reaching the Advanced benchmark than Massachusetts (Singapore was higher on both measures).

• In science at grade 8, no countries had measurably higher average scores than Massachusetts. Singapore, Chinese Taipei, Japan, and the Republic of Korea had average scores not measurably different from Massachusetts.

From TIMSS 2007,
Achievement and attainment

Enrollment in US Degree-Granting Postsecondary Institutions, Selected Years, by Race/Ethnicity, in 1000's

source:
NCES: Digest of Education Statistics, 2010: Table 235
Lessons from 100+ years of testing

- Validity of inferences (not tests)
- Measures vs. estimates
- Accountability matters
- Constructs, content, consequences

Sam Messick (1931 – 1998) on validity (excerpts):

evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use…

it is important to accrue evidence of …
positive consequences as well as evidence that adverse consequences are minimal…
What else we have learned

Political economy = measurement of externalities

Political economy of assessment = externalities of measurement

Some references:
OTA 1992; NAS, 2010; Feuer, various
An assertion:

Given the special rhetorical power of international comparisons:
- readers of daily newspapers and blogs; members of Congress and their staffs; federal, state, and local education officials; parents and other concerned citizens are owed extra assurance that policy declarations are supported by sound empirical evidence.
A framework for policy and research (i)

<table>
<thead>
<tr>
<th>Principle</th>
<th>Guiding questions</th>
<th>Data requirements and the politics of validation: to be continued...</th>
</tr>
</thead>
</table>
| Articulation of intended rationales of using comparative data to guide policy. | • Does the use of the proposed assessment, e.g., PISA or TIMSS, presuppose agreement on the core values embedded in the assessment items?  
• Or, is use of the assessment intended as a tool to foster public deliberation about core values of education? |                                                                 |
| Transparency of embedded assumptions and logic of relationship of ILSA results to policy choices. | • To what extent do inferences derived from ILSA rest on defensible logic and empirical evidence? |                                                                 |
## A framework for policy and research (ii)

<table>
<thead>
<tr>
<th>Principle</th>
<th>Guiding questions</th>
<th>Data requirements and the politics of validation: to be continued...</th>
</tr>
</thead>
</table>
| **Estimation of potential benefits and risks of using assessment results to guide educational practice and policy.** | • Do international rankings provide an adequate basis for development of education reform policies?  
• To what extent are downside risks of making reform decisions based on comparative scores taken into account by policy makers and educators?  
• Do flaws in the comparative score data and their interpretation distort policy judgments and, if so, how are the effects of such errors distributed across economically and socially diverse schools and school systems? | |

Continued...
A framework for policy and research (iii)

<table>
<thead>
<tr>
<th>Principle</th>
<th>Guiding questions</th>
<th>Data requirements and the politics of validation: to be continued...</th>
</tr>
</thead>
</table>
| Acknowledgement of the compelling nature of international comparative rhetoric. | • To what extent does reliance on ILSA contribute to erosion of morale about the quality and prospects of genuine school reform and improvement of teaching and learning for all students?  
  • Is there a macro-level downside risk associated with exaggerated claims of decline and stagnation in educational performance, especially as it is implicitly or explicitly linked to long run economic performance? |                                                                                                                                               |
| Benefits and costs of participation in ILSA.                              | What criteria should guide decisions by policy makers to invest in continued improvement of assessment programs and continued participation?                                                                       |                                                                                                                                               |
| Validation as “procedural rationality” (see Feuer, 2006).                 | How can a comprehensive approach to validity of ILSA promote and facilitate continuous improvement in teaching, learning, and education policy?                                                                   |                                                                                                                                               |
Thanks to...

Madhabi
Eugene
Alina
Judy
Nancy

And many other friends at ets and tc who have endured my ruminations on international comparisons lo these many years...

Contact: mjfeuer@gwu.edu