Large Scale Assessments & Accountability Systems

Michal Beller

mbeller.rama@education.gov.il

National Authority for Measurement and Evaluation in Education
“That which cannot be measured, cannot be improved”
Accountability Processes

- Goals
- Inputs
- Processes
- Measurement
- Feedback
- Change

Source: Eva Baker, 2005
Large Scale Assessments in Israel

- Matriculation Exams – Bagrut (Math et al.)
- Meytzav (Reading, Math, Science, English)
- Mashov Artzi (Reading, Math, English)

International Assessments:
- TIMSS - Math and Science Study
- PIRLS - Reading Assessment
- SITES - Information Technology
  - (1997, 2006)
- PISA - Reading, Mathematical & Scientific Literacy
Matrix Sampling

- Sampling plan in which different **samples** of respondents take different **samples** of items

20) How does NAEP use matrix sampling? What is focused BIB spiraling, and what are its advantages for NAEP?

Typically, several hundred questions are needed to reliably test the many specifications of the complex frameworks that guide the NAEP assessments. Administering the entire collection of cognitive questions to each student would be far too time-consuming to be practical.

Therefore, a number of test booklets are printed for a particular subject, with each booklet containing a different selection of cognitive blocks. This design, called matrix sampling, allows NAEP to assess the entire subject area within a reasonable amount of testing time. In matrix sampling, different portions from the entire pool of cognitive questions are printed in separate booklets and administered to different but equivalent samples of students. This design minimizes the assessment time required per student while allowing complete coverage of the subject being assessed.
הישגיהם הלימודים של תלמידי תיכונים בחינוך והבעלות מבית הוראה והקשר התרבותי בין החינוך של oranot בישראל

ممצעי המחבר
הבינלאומי של
רות זודובסקי
המרץ לחינוך מדעי וטכנולוגי
בית הספר לחינוך
אוניברסיטת תל-אביב

TIMSS 2003

IEA's TIMSS 2003 International Report on Achievement in the Mathematics Cognitive Domains
Findings from a Developmental Project

In a V.S. Mullis
Michael O. Martin
Pierre Foy

רומז לחריטה
למאדורה
בעיתון
© 2006 # 6
TIMSS 2007 is the fourth in a cycle of internationally comparative assessments dedicated to improving teaching and learning in mathematics and science for students around the world. Carried out every four years at the fourth and eighth grades, TIMSS provides data about trends in mathematics and science achievement over time.

To inform educational policy in the participating countries, this world-wide assessment and research project also routinely collects extensive background information that addresses concerns about the quantity, quality, and content of instruction. For example, TIMSS 2007 will continue collecting detailed information about mathematics and science curriculum coverage and implementation, as well as teacher preparation, resource availability, and the use of technology.
TIMSS 2007 Assessment Frameworks

Developing the TIMSS 2007 Assessment Frameworks represents an extensive collaborative effort involving individuals and expert groups from more than 60 countries around the world. The document contains three frameworks for implementing TIMSS 2007—the Mathematics Framework, the Science Framework, and the Contextual Framework for questionnaires. It also provides an overview of the assessment design, including general parameters for item development.

The TIMSS content frameworks for 2007 rely heavily on the extensive efforts expended to update the frameworks for 2003, when specific assessment objectives were developed for Grades 4 and 8. For 2007, there was a further effort to consolidate the major content domains and present them separately for the two grades, as follows:

Mathematics
- Grade 4: Number, Geometric Shapes and Measures, Data Display
- Grade 8: Number, Algebra, Geometry, Data and Chance

Science
- Grade 4: Life Science, Physical Science, Earth Science
- Grade 8: Biology, Chemistry, Physics, Earth Science

Both the mathematics and science assessment frameworks also have a cognitive dimension—Knowing, Applying, and Reasoning. To enable reporting by cognitive domains, these also were revised for TIMSS 2007 to sharpen the distinction among categories.

IEA is extremely grateful for the support of the National Center for Education Statistics of the US Department of Education, the US National Science Foundation, the World Bank, the United Nations Development Programme, and the participating countries.

Download Files

Publications are formatted in Portable Document Format (PDF) and intended for use with Acrobat Reader 7.

Get the full report (3MB)

If you need to download a free copy of Adobe Acrobat Reader, which will enable you to read and print the reports, press the icon below.

The reports are also available in printed and bound volumes which can be ordered for a fee. Shipping and handling included for orders within the United States. International orders require US$10 per book for shipping. Print a fax order form or phone order call +1(617)355-1600.
What PISA Assesses

PISA assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society. In all cycles, the domains of reading, mathematical and scientific literacy are covered not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life.

In the PISA 2003 cycle, an additional domain of problem solving was introduced to continue the examination of cross-curriculum competencies.

Method of assessment
Designing Accountability Systems
Accountability in education is not new. The mechanisms may have changed, but the questions remain the same: Who is responsible for what and to whom?

Although the questions are straightforward, the answers are not, and accountability systems are complex.

Source: Accountability in Education in Canada
Accountability in Education

- Accountability is a tool for policy makers to employ in order to achieve their goals
- It is not a goal in itself
- Accountability is most valuable if it not only documents the current status of the system but also generates information that can support continuous improvement
- School effectiveness with respect to increasing student test scores can be an important source of such information

Henry Braun, ETS (Personal Communication)
Linn (2000) identified five “waves” of reforms that occurred during the last 50 years in the U.S.:

- 1950s: tracking and selection
- 1960s: program accountability
- 1970s: minimum competency testing
- 1980s: school and district accountability
- 1990s: standards based accountability systems
STANDARDS-BASED ACCOUNTABILITY

TEN SUGGESTIONS

Robert L. Linn

1. Tests are relatively inexpensive.
Compared to changes that involve increases in instructional time, reduced class size, training, and attracting better teachers, assessment is very low-cost.

2. Testing changes can be implemented relatively quickly.
Other school reforms may take years to implement, and it may take even longer to know if they have improved schooling.

3. Test results are visible and draw media attention.
Poor results in the first year of a new testing program are usually followed by increased pressure in subsequent years, giving the appearance that schools are improving.

4. Testing can create other changes that would be difficult to legislate.
Research has shown that state- or district-level testing and assessment requirements motivate changes in curriculum and teaching at the school and classroom levels. It is much more difficult to directly legislate changes in the classroom.

Unfortunately, when tests are used to make major decisions about schools and students, these attractive features frequently result in unexpected problems. Test results may be incomplete or misleading, resulting in poor policy decisions. Nevertheless, the policy need for rapid information about student progress and school quality ensures a continued high interest in educational testing.

Robert L. Linn is co-director of the National Center for Research on Evaluation, Standards, and Student Testing and Distinguished Professor of Education at the University of Colorado at Boulder. He is the current chairperson of the National Research Council's Board on Testing and Assessment.

UCLA Graduate School of Education & Information Studies
1

Set standards that are high, but attainable
Non-Realistic Goals

NAEP Mathematics Percent Proficient or Above Trends and Projections (Grades 4 and 8)
Develop standards, *then* assessments
טכנית הלימודים

סטנדרטים של תכנון

סטנדרטים של בישראל

مبחרי נש (::)

שאלונים אחרים
Alignment of Tests to Content Standards

In a project coordinated by the Council of Chief State School Officers and led by Norman Webb, researchers have developed a systematic approach to facilitate alignment and check to see if it exists.

Four criteria for alignment are involved.

- **Categorical Concurrence**: the extent to which both standards and the test incorporate the same content.
- **Depth of Knowledge Consistency**: the extent to which what is elicited from the students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.
- **Range of Knowledge Correspondence**: the extent to which a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, what students need to correctly answer the test questions.
- **Balance of Representation**: the degree to which one objective is given more emphasis than another.

What is included and excluded is systematic: the most challenging objectives are the ones that are under-sampled or omitted entirely. Thus, many of the tests in use by a state cannot be judged to be aligned to the states’ standards — even though most of the items map to some standard or objective.

Include all students in testing programs except those with the most severe disabilities.
Include All Students

- Individuals with Disabilities Education Improvement Act of 2004 (IDEA)
- Universal Design of Assessments

### Creating Accessible Assessments Through UDL

<table>
<thead>
<tr>
<th>Universal Design Principle</th>
<th>Possible Classroom Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple Means of Presentation</strong></td>
<td>Provide student with multiple representations of content and concepts.</td>
</tr>
<tr>
<td>Present alternatives to reduce barriers</td>
<td>E.g. graphics, digital text, printed text, text-to-speech, audio, multimedia, animations, simulations.</td>
</tr>
<tr>
<td>and adjust to the different ways students recognize and process information.</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Means of Expression</strong></td>
<td>Provide student choice in forms of expression.</td>
</tr>
<tr>
<td>Accommodate students’ varied strategic</td>
<td>E.g. hand writing, keyboarding, illustrating, graphic organizers, voice recording, speech-to-text.</td>
</tr>
<tr>
<td>and motor systems.</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Means of Engagement</strong></td>
<td>Provide student with content and interaction options.</td>
</tr>
<tr>
<td>Allow students opportunities for</td>
<td>E.g. choice of material, simulation environments, representation and expression options</td>
</tr>
<tr>
<td>establishing interest in tasks.</td>
<td></td>
</tr>
</tbody>
</table>

http://www.ihdi.uky.edu/uda/
Useful high-stakes accountability requires new high-quality assessments each year that are comparable to those of previous years.
Figure 2
Results of Changing to a New Test Form*

*Based on Linn, Graue, & Sanders, 1991.
High-Quality Assessments

- Validity and Reliability
- ECD – Evidence Centered Design of Assessments (Mislevy et. al)
- Technology
Framework Content Areas That Might Be Measured Better With Computer
Don’t put all of the weight on a single test when making important decisions about students and schools (i.e., retention, promotion, probation, rewards)
In analyses of the role of national educational assessment, insufficient attention has been paid to the central place of the classroom. Rather than encouraging a two-way flow of information, today's "standards-based" frameworks tend to direct the flow of accountability from the outside into the classroom.

The authors of this volume emphasize that assessment, as it exists in schools today, consists mainly of the measurements that teachers themselves design, evaluate, and act upon every day. Improving the usefulness of assessment in schools primarily requires assisting and harnessing this flood of assessment information, both as a means of learning within the classroom and as the source of crucial information flowing out of classrooms.

This volume aims to encourage debate and reflection among educational researchers, professionals, and policymakers. Five source chapters describe successful classroom assessment models developed in partnership with teachers, while additional commentaries give a range of perspectives on the issues of classroom assessment, standardized testing, and accountability.
6

Place more emphasis on comparisons of performance from year to year than from school to school
Figure 1 shows that white math scores were consistently higher than black or Hispanic students' scores for the entire period.
"In this monograph, we clarify the primary questions raised by the use of Value-added modeling (VAM) for measuring teacher effects, review the most important recent applications of VAM, and discuss a variety of the most important statistical and measurement issues that might affect the validity of VAM inferences. Although the document focuses on measures of teacher effectiveness, many of the points discussed here also apply to measures of school effects. The monograph should be of interest to policymakers who are considering the use of VAM for teacher evaluations or accountability. It will also be of interest to researchers who are looking to use VAM to understand teachers or looking for ways to improve VAM models."
Conferences and Events
Linking and Aligning Scores and Scales (ETS-sponsored)

Overview

Date: June 24, 2005 - June 25, 2005  
Location: Princeton, New Jersey  
Event: Linking and Aligning Scores and Scales

Date: June 24, 2005 - June 25, 2005  
Location: Princeton, New Jersey

Linking scores and aligning scales of different tests are fundamental issues for assessment. The history of linking has been evolutionary. There are a variety of data collection designs and statistical procedures for linking. There are different types of linkage.

The Linking and Aligning Scores and Scales conference brought together leading experts in the field to address both the theory and practice of linking scores and aligning scales. Conference sessions addressed the history, data collection, procedures, and the theory and practice associated with different types of links.

This conference was held in honor of Ledyard R Tucker’s theoretical and practical approach to linking educational test scores and scales. Tucker engineered custom solutions to testing problems that made use of available computational capacity. In particular, he made seminal contributions to linking scores and aligning score scales.
Different State Performance Standards

Percent Proficient or Above

<table>
<thead>
<tr>
<th></th>
<th>NAEP</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td>34</td>
<td>67</td>
</tr>
<tr>
<td>Missouri</td>
<td>28</td>
<td>21.3</td>
</tr>
</tbody>
</table>
Set both long- and short-term school goals for all schools to reach.
Uncertainty should be reported in all test results
Uncertainty should be reported

Cautions in Interpretations

Users of this website are cautioned against interpreting NAEP results as implying causal relations. Inferences related to subgroup performance or to the effectiveness of public and nonpublic schools, for example, should take into consideration the many socioeconomic and educational factors that may also impact performance.

Beginning in 2002, the NAEP national sample was obtained by aggregating the samples from each state, rather than by obtaining an independently selected national sample. As a consequence the national sample size increased, and smaller differences between years or between types of student were found to be statistically significant than would have been detected in previous assessments.

The NAEP mathematics scale makes it possible to examine relationships between students' performance and various background factors measured by NAEP. However, a relationship that exists between achievement and another variable does not reveal its underlying cause, which may be influenced by a number of other variables. Similarly, the assessments do not reflect the influence of unmeasured variables. The results are most useful when they are considered in combination with other knowledge about the student population and the educational system, such as trends in instruction, changes in the school-age population, and societal demands and expectations.

**standard error.** A measure of sampling variability and measurement error for a statistic. Because of NAEP's complex sample design, sampling standard errors are estimated by jackknifing the samples from first-stage sample estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using plausible values.
Evaluate not only the hoped-for positive effects of standards-based assessments, but the unintended negative effects of the testing system.
Two Schools vs. Fixed Target (Proficient or Above)

Year 1

Year 2

Percent

School 1

School 2

Fixed Target
“…Attempting to meet the provisions of NCLB could well lead to extraordinary attention being given to students who are just below the basic/proficient cut point (sometimes called “bubble” students), with perhaps considerably less attention given to other students. From these perspectives, I would argue that the current reauthorization of ESEA might be more aptly titled “Most Children Left Behind.” (Brennan, 2004)
Between-subjects reallocation: WA, grade 4

SOURCE: Stecher et al., 2000a, 21.
Teaching to the Test
הכנה למתיז'ב

ששתון מתכון למתיז'ב

ב-13 בדצמבר יא getch מ"ב ליתון ח".решת אורת מפאשר לכל תלמידי ישראל

בעברית לעברית - להתקין למכרה במעדות אמות של אבני בני-

האריך מילוי 200 שאלונות על פי ענפים המבחנים מערערים טונשלון מדריך אקדמיה, מדריך, משלים מילוי

רבייה והפתיחה.

ברוך ברוך ברכה מברכה מברכה מברכה מברכה

בכרך כל תערוכת מבחר 미"ב זמרו קצריםلاح אחד פיתחה סינתלמודים, מゝ שמעון בינד

מברית המורמים לעברין את התלמודים זכאי לדינה. הם הקים הזה לא נمسلم למתו. הוא

ממידי شكלאון רבו להכין ע⾝ים, והם מתו ממידי – בעברה ובעברת.

אתה ההכנה למתיז'ב

השאלה מוצגת בדיעבד אקייאת ב-4 אופי הירוצים המסוגלים למתלים משלים מעשיים: לברך

עם שמעון, המטאפר קובלת מספר פילוליلاح אחד של שאלת שלטונות פטרון: הירוצים עם ציוו

משלס יראם ל-10 שאלות של שאלת מתנחת לפורט: הירוצים עם מ להתמדה, המטרפים קראים יידה

טקטיסים מעניינים – באומנות שאילו פיתחת שאולות רב-ברידר; פיתח钵 ממידי "ב

קדום, שב⽤ים המתחד את ינו הערכות הידר של ₪ כל ברך אתنشי המבחנים, ואחר

כיתור אָרְצוּ יְשָׁרָאָל

ב-13 בדצמבר יא getch מ"ב ליתון ח".решת אורת מפאשר לכל תלמידי ישראל

בעברית לעברית - להתקין למכרה במעדות אמות של אבני בני-

האריך מילוי 200 שאלונות על פי ענפים המבחנים מערערים טונשלון מדריך אקדמיה, מדריך, משלים מילוי

רבייה והפתיחה.

ברוך ברוך ברכה מברכה מברכה מברכה מברכה

בכרך כל תערוכת מבחר 미"ב זמרו קצריםلاح אחד פיתחה סינתלמודים, מゝ שמעון בינד

מברית המורמים לעברין את התלמודים זכאי לדינה. הם הקים זה לא נمسلgovernment למתו. הוא

ממידי شكלאון רבו להכין ע⾝ים, והם מתו ממידי – בעברה ובעברת.

אתה ההכנה למתיז'ב

השאלה מוצגת בדיעבד אקייאת ב-4 אופי הירוצים המסוגלים למתלים משלים מעשיים: לברך

עם שמעון, המטאפר קובלת מספר פילוליلاح אחד של שאלת שלטונות פטרון: הירוצים עם ציוו

משלס יראם ל-10 שאלות של שאלת מתנחת לפורט: הירחים עם מitatד, המטרפים קראים יידה

טקטיסים מעניינים – באומנות שאילו פיתוח שאולות רב-ברידר; פיתח钵 ממידי "ב

קדום, שב⽤ים המתחד את ינו הערכות הידר של ₪ כל ברך אתنشי המבחנים, ואחר

כיתור אָרְצוּ יְשָׁרָאָל
What PISA Assesses

PISA assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society. In all cycles, the domains of reading, mathematical and scientific literacy are covered not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life.

In the PISA 2003 cycle, an additional domain of problem solving was introduced to continue the examination of cross-curriculum competencies.

Method of assessment

"הכנה לבתניה" או שינו
נדרש של תכנית הלימודים?
לאה קי המלונות בבחרה יホーム למבוקש
בגרלים?  

במוקים לאדם את כל המלונים. התפקידים מתאימים פיצח,คอม.
שהוקע את המלונות, היא אוצרת את המלונים
שהוקע את המלונות.getMethod: "ה Jazeera"  
ולש進め ב Lansing."יח" האונין המריץ: "וורוד ובעניין
הצבור".

פגך עזר:  

משמע ההתקפה על האונין המריץ. לה סטרים המקצוע ולחゲームים
לשם ב Merrion. "יח" האונין המריץ: "וורוד ובעניין
הצבור".

מכה את המלונות, היא אוצרת את המלונים על קרזים המקצוע ולח
 الوزراء של המלונים. méthod: "ה Jazeera"  
ולש進め ב Lansing."יח" האונין המריץ: "וורוד ובעניין
הצבור".

The Programme for International Student) (רה"מ פי'ה)  

ה컫ת מתן ה组织领导 מלבניJuan.  

בגרלים: "ה Jazeera"  
ולשenville ב Lansing."יח" האונין המריץ: "וורוד ובעניין
הצבור".

לשתתף ב Merrion. "יח" האונין המריץ: "וורוד ובעניין
הצבור".
Performance on Coached and Uncoached Tests

![Graph showing the comparison of Grade Equivalents for Test C and Test B from 1985 to 1991](image)

SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)
Narrowing the achievement gap means that we must provide all children with the teachers and resources they need in order to reach our high expectations.
Welcome to the National Grid for Learning, the gateway to quality assured educational resources on the Internet. The NGfL provides a network of selected links to websites that offer high quality content and information. Whether you are learning, supporting, teaching or managing, there are resources on the NGfL for you.

- Vision for 21st-century learning
- New tools to create learning materials
- City council tops broadband take-up list
- All news stories...

Email Alerts
Set up weekly email alerts from the NGfL

Online Event
Life Onboard HMS Victory

NGfL search snippet
Add the NGfL search to your website

The Teaching Zone
The Baccalleau: Crossroads for Cultures
Special Operations Executive in the Far East
מגנזר פריטי מבוקשים
אטגריה
مؤינה
תורשה
מים
רבייה
בתי נופים
חצרות
חומרים
ה.DriverManager
ועד הסקר
ענני
מרז
© 2006 # 48
1. עבדה שלג' לפי הודי סדר המשולח
הטلمיד בכל את 7 ו- 5 בمقפוק לכל את 7-0 וואה
מכף התחת בכל בדרך
במקינר ריבי השניה נושאת בכפל 7 בפורת השראו
ולא נושאת בכפל 7 בפורת המאות. (minate שפורת ה-0
בפרת ורה ליגה)

בｂיווע שגיאות דומינון התכבותלה התשואות:

| 208 x 7 = 3556 |
| 504 x 7 = 3648 |
| 504 x 7 = 4248 |

2. טעות שמקדשה בכל בפג
הטلمיד פותר: 7x0=7

3. טעות שמקדשה "שכחת" או חוסר הבנת המבנה

| 504 x 7 = | 3598 |
| 504 x 7 = | 3508 |
رام"ה
מדידה והערכה
מדידה בשיחות הלמידה
תליכים
משוב
הtrzymה ושינוי
משות ועדיים
תודה!

mbeller.rama@education.gov.il

שאלות? הענות?