Using Technology to Measure Old and New Domains

Randy Bennett

ランディー・ベネット

ETS
rbennett@ets.org

Presentation at JART, Tokyo, Japan, September 9, 2009

Funded by the National Center for Education Statistics, Institute of Education Sciences, US Department of Education under contract number ED-02-CO-0023

Listening.
Learning.
Leading.

Acknowledgements

- Brent Sandene, Jim Braswell, Bruce Kaplan, Andreas Oranje
- Nancy Horkay, Nancy Allen, Fred Yan
- Hilary Persky, Andy Weiss, Frank Jenkins
- Many others from ETS, Westat, NCES, and NAGB

Overview

- Background
- Using technology to measure existing domains
- Using technology to measure new domains
- Conclusion

What is NAEP?

- National Assessment of Educational Progress
  - The only nationally representative and continuing assessment of what US students know and can do in various subject areas
    - Paper testing program
    - Administered to samples in grades 4, 8, and 12
    - Scores reported for groups but not individuals

Three Research Studies

- Purpose
  - To explore the use of new technology in NAEP to:
    - Measure existing domains
      - Math Online (MOL)
      - Writing Online (WOL)
    - Measure new domains
      - Problem Solving in Technology-Rich Environments (TRE)

Why Should We Care About Comparability Across Delivery Modes?

- If delivery mode affects scores, our ability to draw valid conclusions from test results may be reduced:
  - If results are to be compared over time and the delivery mode has changed from paper to computer
  - If results are to be aggregated across individuals when some individuals have taken the test on paper and others have taken it on computer
  - If groups taking the test on computer are to be compared with one another and computer delivery affects one group more than another

What Do We Mean by "Comparability?"

- Scores can be considered to be comparable when they can be used interchangeably
- *Standards for Educational and Psychological Testing*
  - Highly similar rank-ordering of individuals across conditions
  - Highly similar score distributions across conditions

Existing Literature

- At the K-12 level the literature is limited in that most studies:
  - Are unpublished conference papers
  - Used multiple-choice items only
  - Used convenience samples
  - Considered only level differences

MOL Key Questions

- Do 8th grade students perform differently on a paper vs. computer mathematics test?
- Does delivery mode differentially affect the overall performance of particular NAEP reporting groups?
- Does computer familiarity appear to affect online test performance?

Procedure, Samples, and Instruments

Grade 8	
Online Condition	***Pencil and Paper Condition***
•Paper math pretest (MC)	•Paper math pretest (MC)
•Online tutorial and computer facility measure	
•*Online math test (MC & CR)*	•*Paper version of the "same" math test (MC & CR)*
•Online background questionnaire	•Paper background questionnaire

Computer Facility Measure

- Assessed speed and accuracy in:
  - Pointing, clicking, and scrolling with the mouse
  - Entering numbers
  - Typing and editing text
  - Using the onscreen calculator

Constructed-Response Item Formats

- Figural response
- Numeric entry
- Text entry

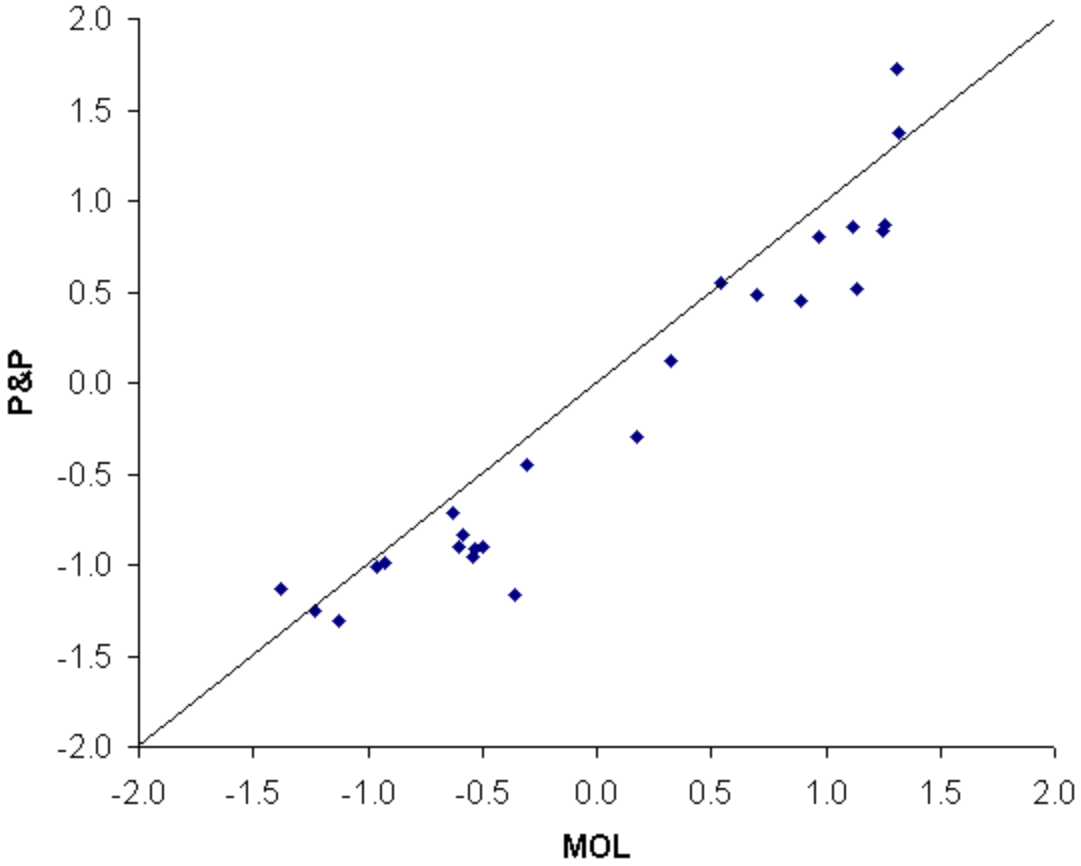
Do Students Perform Differently on Paper vs. Computer?

- 8th grade students scored significantly higher statistically on the paper than online versions of the test
- The difference was:
  - ~ 4 points on a 0-400 scale
  - ~.14 SD units

Students Performed Better on Paper

- Is that effect associated with a few items or is it more pervasive?
- Is the effect associated with constructed-response items more than with multiple-choice items?

Item Difficulties (IRT *b)* for All 8th Grade Items: P&P vs. MOL



Does Delivery Mode Differentially Affect Particular Reporting Groups?

- No significant mode difference for reporting groups categorized by gender, race/ethnicity, region, school location, or school type
- 8th grade students reporting that at least one parent graduated college performed significantly better statistically on paper than on computer
  - ~6 points on a 0-400 scale
  - ~.21 standard deviation units

Does Computer Familiarity Appear to Affect Online Test Performance?

- Computer "familiarity" significantly predicted online test score, after controlling for paper mathematics skill
  - The greater the computer familiarity, the higher the Math Online score
  - Improvement in prediction was ~8 percentage points for 8th grade

What Factors Might Have Caused the Mode Effects Observed in MOL?

- Majority of students (62%) tested via NAEP laptops
  - Many students would have been more familiar with their school computers than with NAEP laptops and could have performed worse on the laptops as a result
- Technology problems interrupted test sessions for 11% of 8th graders
  - Interruptions could have affected student concentration or motivation

What Factors Might Have Caused the Mode Effects Observed in MOL?

- MOL required some degree of computer facility to respond (especially for the more complex CR items) and students varied in their computer facility
- Some items were formatted differently for computer vs. paper presentation and the difference in formatting may have made those items cognitively more difficult on computer

Key Results from MOL (2001)

- The scores from paper and computer tests did not appear comparable in that:
  - Average scores for paper were (marginally) higher than for computer
  - The computer test appeared to measure both math proficiency *and* computer skill
- Technology issues interfered with test delivery and may have contributed to these comparability results

WOL Key Questions

- Do 8th grade students perform differently on a paper vs. a computer writing test?
- Does delivery mode differentially affect the overall performance of particular NAEP reporting groups?
- Does computer familiarity appear to affect online test performance?

Procedure, Samples, & Instruments

Paper Condition	***Online Condition***
2002 Main NAEP (January – March)	
•Paper writing test with two essays	•Paper writing test with two different essays
•Paper background questionnaire	•Paper background questionnaire
2002 WOL (April – May)	
	•Online tutorial and computer facility measure
	•Online writing test with the same two essays as the paper condition
	•Online background questionnaire

Computer Facility Measure

- Typing speed
  - Number of words typed in two minutes from a 78-word passage
- Typing accuracy
  - Sum of errors made in typing the passage
- Text editing
  - Number of tasks completed correctly, including deleting, inserting, modifying, and moving text

Do 8th Grade Students Perform Differently on a Paper vs. a Computer Writing Test?

- No statistically significant mean score differences between modes
- No statistically significant mean word-count differences between modes
- A statistically significantly greater percentage of students responded validly on paper as compared with computer
  - But only for one essay and by only 1 percentage point

Does Delivery Mode Differentially Affect Particular Reporting Groups?

- No statistically significant mean score difference between modes for most reporting groups
- Students from urban-fringe/large-town locations scored significantly higher statistically on paper than on computer
  - ~.2 point on a 0-6 scale
  - ~.15 standard deviation units

Does Computer Familiarity Appear to Affect Online Test Performance?

- "Hands-on computer facility" significantly predicted online writing score, after controlling for paper writing performance
  - The greater the computer facility, the higher the Writing Online score
  - Improvement in prediction was ~11 percentage points

What Factors Might Have Caused the Mode Effects Observed in WOL?

- Very few sessions were interrupted due to technology problems
- Majority of students (65%) tested via NAEP laptops
  - Many students would have been more familiar with their school computers than with NAEP laptops
  - Both a small experiment and a quasi-experimental analysis were conducted

Key Results from WOL (2002)

- The scores from paper and computer tests did not appear comparable
  - Even though mean scores were not measurably different, the computer test appeared to measure both writing proficiency *and* computer skill
- Technological problems did not appear to affect performance

Lesson Learned

- When we deliver a traditional math test on computer, we may be testing a mix of math skills and computer familiarity even though only math skill is the target proficiency
  - Students *had* to use the computer to demonstrate their math proficiency even though the computer was *not* used as a tool for doing mathematics
  - The computer was simply an item-presentation and response-collection mechanism

Implications for Assessing Math

- Render items so that they measure only math proficiency
- Make sure that students have sufficient time to familiarize themselves with the characteristics of the testing system and the item formats
- Deliver the test on familiar hardware
  - Easier to achieve now than in 2001
    - Internet-connected school computers are much more widely available
    - Commercial test delivery software can accommodate a greater range of school technology
    - Students are familiar with a greater variety of computer types

Implications for Assessing Math

- Use the computer-skill requirements of complex CR questions *purposefully*
  - Incorporate computer tools that allow students to demonstrate math skills that couldn't be demonstrated on paper
    - Modeling problem situations mathematically with spreadsheets

Lesson Learned

- When we deliver a writing test on computer, we assess how well students can write using the computer as a tool
  - Writing on computer and writing on paper are not necessarily the same
    - Some students write better on computer than paper
      - Computer may allow greater fluency
      - Computer may allow more revision cycles
    - Other students write better on paper than computer
      - Computer is an impediment because they don't have sufficient text entry and editing skills

Implications for Assessing Writing

- As long as significant numbers of students write better in one or the other mode, we may get different group proficiency estimates from:
  - Testing *all* students on paper
  - Testing *all* students on computer
  - Testing students in the mode in which they typically write

Implications for Assessing Writing

- The approach we take should depend on what we want to know about student writing proficiency
  - How well do students write on paper?
  - How well do students write on computer?
  - How well do students write in their typical mode?

TRE Project Purpose

- *Demonstrate* an assessment that:
  1. Measured important skills not easily tested on paper
  2. Could be delivered successfully on computer by NAEP to 8th graders in a sample of schools throughout the nation
  3. Held together reasonably well psychometrically
  4. Produced credible results

Presentation Overview

- Run through the four intentions, their related outcomes, and some associated issues
- Give my opinion as to what the project did and didn't do effectively
- Suggest how TRE-like measures might be used in large-scale assessments like NAEP
- Offer some closing comments

1. *Measured Important Skills Not Easily Tested on Paper*

- Problem solving with technology
  - Important by virtue of what workers in a knowledge economy, or students in higher education, must know and be able to do
  - By definition, can't be easily measured on paper

Conceptualizing Problem Solving with Technology

	Technology Environment					
Content Domain	Searchable Database	Text Processor	Simulation Tools	Dynamic Displays	Spread-sheet	Comm. Tools
Biology						
Ecology						
Physics						
Balloon Science						
Economics						
History						

What did TRE Attempt to Assess?

- **Scientific-inquiry skill**: being able to *find* information about a given topic, *judge* what information is relevant, *plan and conduct* experiments, *monitor* one's efforts, *organize and interpret* results, and *communicate* a coherent interpretation.
- **Computer skills**: being able to carry out the largely mechanical operations of using a computer to find information, run simulated experiments, get information from dynamic visual displays, construct a table or graph, sort data, and enter text.

TRE What did TRE Attempt to Assess?

- TRE was intended:
  - *Not* as a science assessment on computer
  - As a test of skill in using the computer for problem-solving (in a science-related context)

Two Scenarios

- Each scenario:
  - Attempted to assess a different (small) subset of the elements comprising our conception of problem solving with technology
  - Contained extended tasks offering multiple opportunities to observe student behavior
  - Tried to more faithfully represent than do traditional tests the types of challenges individuals encounter in work and advanced academic settings

TRE Simulation scenario

- Presented the student with a *tool* for asking “what-if” questions
- The student was expected to use this tool for experimentally solving C-R and M-C problems related to the science of gas-balloon flight

Problem 1

How do different payload masses affect the altitude of a helium balloon?

Design Experiment

Choose Values

Make Prediction

Run Experiment

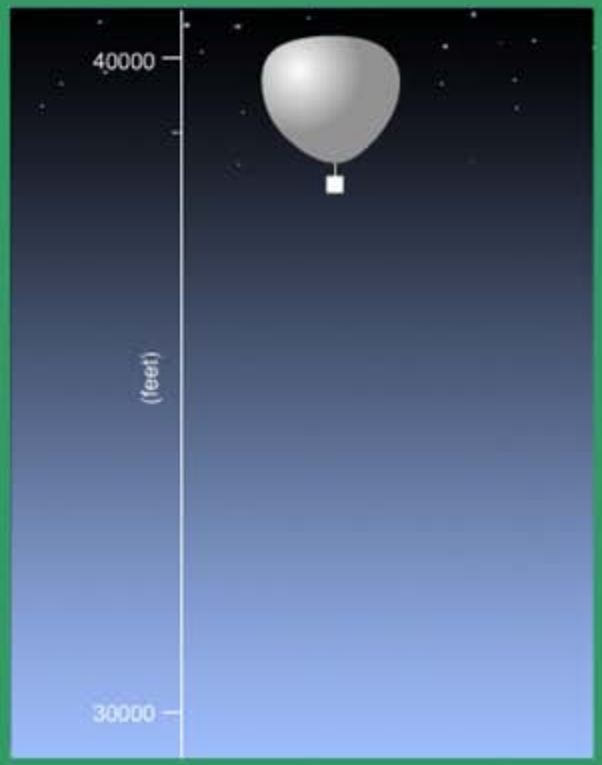
TRY IT!

Interpret results

Make Table

Make Graph

Draw Conclusion



Altitude (feet)	Balloon Volume (cubic feet)	Time to Final Altitude (minutes)
36211	3083	36

Payload Mass (pounds)	Amount of Helium (cubic feet)
10	2275

Glossary

Science Help

Computer Help

Next

TRE Search Scenario

- Presented the student with a *tool* for locating information on a simulated WWW
- The student was expected to employ the tool to answer C-R and M-C questions related to the uses and science of gas-balloon flight

Simulated WWW

- A simulated WWW was chosen to:
  - Increase standardization
  - Prevent visits to inappropriate sites
- The database consisted of ~5,000 pages pulled from WWW, including relevant and irrelevant material
- The information needed to answer the assessment questions was not available on any one page

Judging the Relevance of Web Pages

- All pages were rated by a single judge for pertinence to the motivating problem on a 1-4 scale
- All pages designated as relevant or partly relevant (2, 3, or 4) were independently rated again by two other judges
- Differences were resolved by consensus

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

Go

If you want to narrow your search, try placing the word "and" between your search terms.

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

balloons Go

If you want to narrow your search, try placing the word "and" between your search terms.

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

balloons Go

If you want to narrow your search, try placing the word "and" between your search terms.

Documents 1 to 15 of 300 matching the query "*balloons*".

1. **History of Balloon Delivery & Balloon Decorating Industry 800-424-2323 or 202-785-1290**
1000

Abstract: HISTORY OF BALLOON DELIVERY AND DECORATING INDUSTRY. BALLOONS SINCE 1896 BALLOON BOUQUETS® SINCE 1976. An American success story. CBS Evening News, Charles Kuralt, "On The Road" and Charles Kuralt, On The Road With Charles Kuralt, G.P. Putnam's Sons, 1985. Moves On". . CBS Evening News, Charles Kuralt, "On The Road"

http://naeptba.ets.org/tre/universe/balloon_bouquets/www_balloonbouquets_com/webdoc3.htm
size 5376 bytes - 1/9/2003 4:24:06 PM GMT

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

HISTORY OF BALLOON DELIVERY AND DECORATING INDUSTRY

BALLOONS SINCE 1896
BALLOON BOUQUETS® SINCE 1976

"An American success story...." CBS Evening News, Charles Kuralt, "On The Road", and Charles Kuralt, On The Road With Charles Kuralt, G.P. Putnam's Sons, 1985

It was twenty three years ago in 1976 in Washington, DC that the balloon delivery and arts industry had its beginning. Joe DelVecchio, descendant from three generations of balloon vendors, conceived of the idea of delivering balloons in bunches, and obtained Federal and state trademark registrations for BALLOON BOUQUETS®.

In June, 1976, DelVecchio opened Balloon Bouquets of Washington, DC, and the first delivery of a helium balloon arrangement was made. Balloon Bouquets® franchises opened in major U.S. cities, and an

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

Go

If you want to narrow your search, try placing the word "and" between your search terms.

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

gas balloons Go

If you want to narrow your search, try placing the word "and" between your search terms.

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

NAEP TRE Search Page

Enter your search below:

Tips for searching

gas balloons Go

If you want to narrow your search, try placing the word "and" between your search terms.

Documents 1 to 15 of 300 matching the query *"gas balloons"*.

1. **Gas Balloons - Westward Gallery**
   545

   Abstract: Gas Balloons. From Our Exclusive Collection. Gas Balloons 97 - 1. Gas Balloons 97 - 2. Gas Balloons 97 - 3. Gas Balloons 97 - 4. Gas Balloons 97 - 5. Gas Balloons 95. Copyright © 1995-2000 by Westward Connections Inc. and/or respective content providers. Trade names and trademarks are recognized as properties of the

   http://naeptba.ets.org/tre/universe/westward/www_westward_com/gallery/exhibit7.htm
   size 3464 bytes - 1/9/2003 5:22:39 PM GMT

2. **Difference Between Hot Air and Gas Balloons**
   500

Some scientists study space with large helium gas balloons. These balloons are usually launched from the ground into space but can also be launched from spacecraft near other planets.

Why do scientists use these gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons.

Base your answer on more than one web page or site. Be sure to write your answer in your own words!

You will be scored on:

- how well you search,
- the quality of your bookmarks, and
- how well you answer the questions

Copy = Ctrl + c
Paste = Ctrl + v
Find = Ctrl + f

Ballooning

how balloons work
history
links

Illinois Institute of Technology

How Balloons Work

A balloon generates lifting force when the weight of the gases inside the balloon is less than the weight of the surrounding air that it displaces. The same forces are generated in an air bubble that rises through water - the air in the bubble weighs less than water that it displaces, generating lifting force that causes the bubble to rise. The amount of lifting force is determined by the volume of the balloon and the *density* (weight per unit volume) of the gases. Because hot air and helium are less dense than normal air, they are commonly used to fill balloons.

While in flight a balloon moves in the direction of the prevailing winds, which vary at different altitudes. A balloon pilot can control the altitude of a balloon, but must rely on these winds to carry the balloon in a desired direction. The task of navigating a balloon can therefore be quite challenging.

There are three main types of balloons, as described below:

- Gas Balloons that are inflated with a gas that is less dense than air (e.g., helium)

TRE Scores

- Total
  - Computer Skills
  - Scientific Inquiry

Scoring Student Performance

- Collect evidence of proficiency
  - Used logical analysis, results of previous research, and analysis of pilot-test data to determine which events should be employed as evidence
  - Judged each piece of evidence according to a rubric
  - Aggregated the pieces to form the three scores

Evidence

- Computer Skills
  - Use of advanced search techniques
  - Use of the Back button
  - Number of searches for relevant hits
  - Use of hyperlinks to dig down
  - Use of bookmarking to save pages
  - Use of deletion for unwanted filed pages

Evidence

- Scientific Inquiry
  - Use of relevant search terms
  - Average relevance of hits returned
  - Relevance of pages visited or bookmarked
  - Accuracy and completeness of the answer to the constructed-response question
  - Number right on the four synthesizing multiple-choice questions

A Rubric for Evaluating Bookmarking as Evidence of Computer Skill

- If two or more pages were bookmarked, give full credit.
- If only one page was bookmarked, give partial credit.
- If no pages were bookmarked, give no credit.

Aggregating the Evidence

- Used a series of statistical models to weight and combine the pieces of evidence to create Total, Computer Skills, and Scientific Inquiry scores
  - Item response model
  - Structural equation model
  - Conditioning model
- Reported scores on a scale with a mean of 150 and SD of 35

2. *Delivering on Computer Nationally*

- Student Sample
  - Selected to be nationally representative
  - Participants included 2,134 8th grade students from 222 schools
  - Participation rate was ~80%
    - ~78% for the 2000 paper NAEP grade 8 science assessment
  - Randomly assigned to one of the two scenarios
  - 25 student records (~1%) contained no responses
  - 1,077 students with valid Search responses

Data Collection

- TRE was the third NAEP online study
- All administrations were proctored by NAEP field staff
  - Trained to deal with basic technology-related issues
  - Backed up by telephone tech support
  - Often participants in the two previous NAEP online studies

Data Collection

- Used standard Internet browser software, with common plug-ins and extensions
- Needed to test only 10 students per school
  - Allowed field staff to use NAEP laptops when direct Internet delivery to school computers wasn't feasible

3. *Hold Together Reasonably Well Psychometrically*

- To what degree are Search scores internally consistent?
- Do the three Search scores provide some amount of independent information?
- What search behaviors predict score on the constructed-response question and are these predictions in the expected direction?
- How are Search scores related to reported computer use and are these relations in the expected direction?

Internal Consistency

Scale	Number of Observables	α
Total	11	.74
Scientific Inquiry	5	.65
Computer Skills	6	.73

(Disattenuated) Scale Intercorrelations

Scales	*r*
Scientific Inquiry with Total	.68
Computer Skills with Total	.68
Scientific Inquiry with Computer Skills	.57

(Disattenuated) Correlations of Observables with Scale Score

Observable	Computer Skills	Scientific Inquiry
Relevance of pages visited or bookmarked	.17	***.71***
Accuracy/completeness on CR question	.39	***.70***
Use of relevant search terms	.33	***.51***
Number right on final MC questions	.28	***.44***
Average relevance of hits to motivating problem	.20	***.34***
Use of hyperlinks to dig down	***.69***	.37
Use of Back button	***.65***	.36
Number of searches for relevant hits	***.65***	.33
Use of bookmarking to save pages	***.60***	.45
Use of advanced search techniques	***.46***	.30
Use of deletion for unwanted filed pages	***.24***	.08

N = 672 to 1,077. All values are significantly different from zero at $p < .05$.

The Constructed-Response Question

- Stimulus
  - "…Why do scientists use … gas balloons to explore outer space and the atmosphere instead of using satellites, rockets, or other tools? Be sure to explain at least three advantages of using gas balloons. Base your answer on more than one web page or site. Be sure to write your answer in your own words…"
- Responses rated by human judges once on a 3-point scale for accuracy and completeness
- 25% of responses double-scored independently, with an agreement rate of 90%

A Response Receiving a Top Score

- "One of the advantages of using a balloon is that is has a simple design and can hold a lot of weight. It also costs less to make a balloon rather than making a satelite. You can also launch them in the area you wish to conduct your experiment. It takes little time for it to be constructed as well. This is why it is better to have a balloon rather than a satelite or space shuttle."

Correlations of Observables with the Constructed-Response Raw Score

Observable	*r*
Relevance of pages visited or bookmarked	.55*
Use of bookmarking to save pages	.35*
Use of relevant search terms	.32*
Average relevance of hits to motivating problem	.21*
Use of hyperlinks to dig down	.21*
Use of advanced search techniques	.21*
Number of searches for relevant hits[1]	.20*
Use of back button	.19*
Use of deletion for unwanted filed pages	.03

[1] Fewer searches receives a higher score than more searches.

*$p < .05$.

Relationship Between Search Performance and Reported Computer Use

- On all three TRE Search scales, students who reported:
  - Using a computer daily outside of school scored higher than students who reported using it less frequently
  - Using a computer to find information on the Internet to a large extent scored higher than students who reported using it to find information on the Internet to a small extent
  - Using a word processor, *regardless of extent*, scored higher than students who reported not using a word processor at all

4. *Produced Credible Results*

Mean Scores and Standard Errors for Gender

Group	N	Total	Sci. Inquiry	Comp. Skills
Male	517	**148** (2.4)	**149** (2.7)	**147** (2.5)
Female	560	**151** (2.3)	**150** (2.3)	**152** (1.9)

Mean Scores and Standard Errors for Race/Ethnicity

Group	N	Total	Sci. Inquiry	Comp. Skills
White	643	**161** (1.9)	**160** (1.6)	**158** (1.7)
Black	185	**121** (3.8)	**125** (2.8)	**128** (3.3)
Hispanic	188	**139** (3.4)	**137** (4.8)	**142** (3.4)

Mean Scores and Standard Errors for Parents' Highest Ed. Level

Group	N	Total	Sci. Inquiry	Comp. Skills
Not finish HS	72	**133** (3.7)	**135** (4.3)	**139** (4.5)
Grad HS	214	**142** (4.4)	**143** (2.9)	**145** (3.1)
Post HS	202	**155** (3.0)	**154** (2.7)	**154** (2.6)
Grad College	497	**157** (2.4)	**156** (2.4)	**155** (2.4)

Mean Scores and Standard Errors for Eligibility for School Lunch

Group	N	Total	Sci. Inquiry	Comp. Skills
Not eligible	656	**160** (1.6)	**158** (2.0)	**158** (1.8)
Reduced-price	70	**145** (4.3)	**148** (3.7)	**147** (4.4)
Free lunch	300	**129** (2.5)	**131** (2.6)	**133** (2.5)

What Did TRE Appear to Do (and Not Do) Effectively

- Did
  - Define problem solving with technology based on prior research, break it down into measurable (process and product) components, and create a demonstration performance assessment
- Didn't
  - Ground the assessment in a NAEP content framework
  - Cover problem solving with technology, scientific inquiry, computer skills, or even search skill, very broadly or deeply
  - Use the real Internet

What Did TRE Appear to Do (and Not Do) Effectively

- Did
  - Deliver on computer to a national sample of students with participation rates comparable to paper NAEP and without any significant technology problems
- Didn't
  - Deliver to either a large sample of schools or to a large sample of students

What Did TRE Appear to Do (and Not Do) Effectively

- Did
  - Produce scores that appeared to function in a reasonable way psychometrically
  - Produce population-group results basically consistent with findings from NAEP assessments in associated content areas
- Didn't
  - Provide convincing evidence of validity
  - Produce results that can be taken as estimates of problem solving with technology, scientific inquiry, computer skills, or search skill for the nation's 8th-grade students

How Would Such a Measure Be Used in NAEP?

- As part of a content assessment
  - 2009 NAEP Science Assessment
    - Included extended online tasks administered to a student subsample as a complement to the paper assessment
- As part of an ICT assessment
  - Built of multiple scenarios covering a range of substantive contexts and using a variety of technology tools

Lessons Learned

- We can successfully measure some important domains that can't be assessed through paper-and-pencil, M-C tests
- Going beyond traditional testing is extremely challenging
  - There usually isn't a well-developed research base, nor a widely accepted content framework
  - Designing performance tasks for computer is a relatively new activity
  - Many schools do not yet have the technology infrastructure
  - Students produce *extensive* information when taking these tests

Implications

- The need for measures of new domains is *not* going to go away
- Assessment agencies will have to learn how to assess such domains
- We will *never* learn how to measure these domains effectively if we're not willing to invest the time, effort, and money in trying

Conclusion

- In trying to measure:
  - An existing domain on computer like math, we may be testing a mix of domain and computer skills when domain skill is the intended target proficiency
    - Either remove computer skills from the test or redefine the construct to include computer skills
  - An existing domain on computer like writing, however, we are bordering upon measuring a new domain
    - The computer is a tool for doing writing
    - But it is not the *only* tool for doing writing
- Some domains can only be measured on computer because the computer is central to the domain definition

Conclusion

- What we choose to measure on computer, and how we choose to measure it, should depend on construct definition and on whether we wish that definition to be:
  - As it was for paper testing
  - Intentionally changed by computer delivery
  - Defined in significant part by computer delivery itself

Using Technology to Measure Old and New Domains

Randy Bennett

ランディー・ベネット

ETS
rbennett@ets.org

Presentation at JART, Tokyo, Japan, September 9, 2009

Funded by the National Center for Education Statistics, Institute of Education Sciences, US Department of Education under contract number ED-02-CO-0023

Listening.
Learning.
Leading.