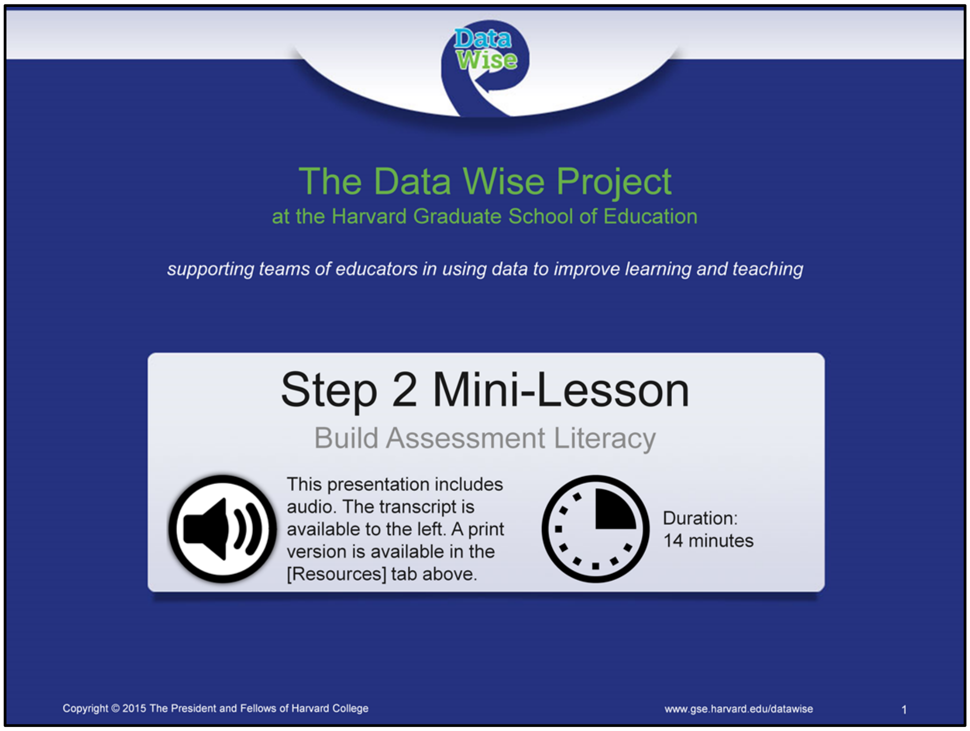


The Data Wise Project

at the Harvard Graduate School of Education

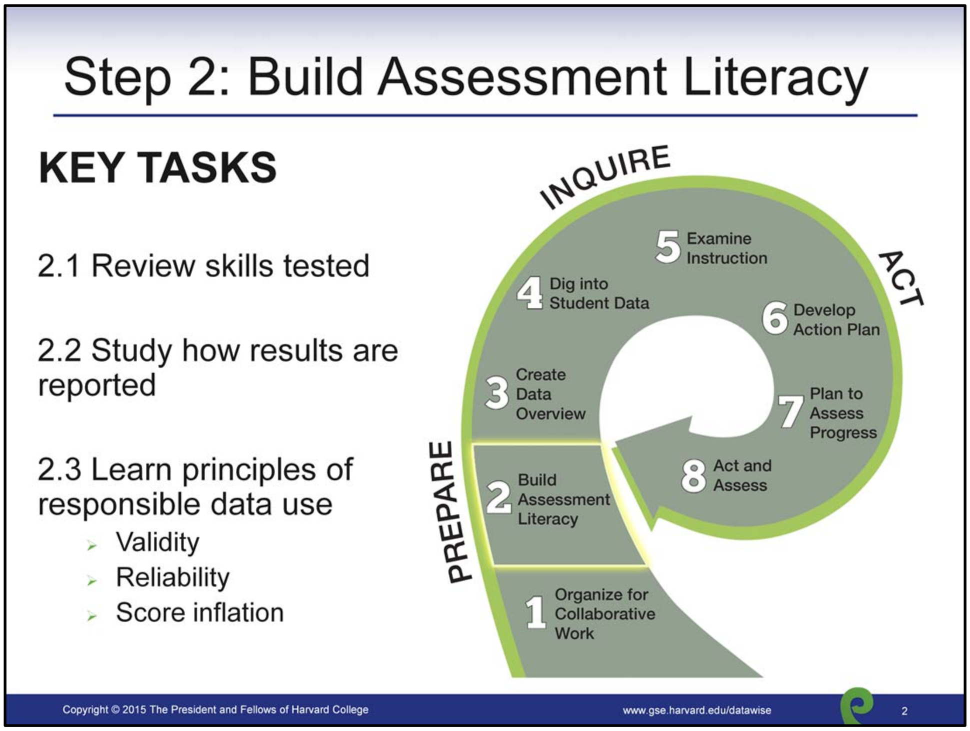
supporting teams of educators in using data to improve learning and teaching

Step 2 Mini-Lesson

Build Assessment Literacy

This presentation includes audio. The transcript is available to the left. A print version is available in the [Resources] tab above.

Duration:
14 minutes



Step 2: Build Assessment Literacy

KEY TASKS

2.1 Review skills tested

2.2 Study how results are reported

2.3 Learn principles of responsible data use

- Validity
- Reliability
- Score inflation

PREPARE

1 Organize for Collaborative Work

2 Build Assessment Literacy

3 Create Data Overview

INQUIRE

4 Dig into Student Data

5 Examine Instruction

ACT

6 Develop Action Plan

7 Plan to Assess Progress

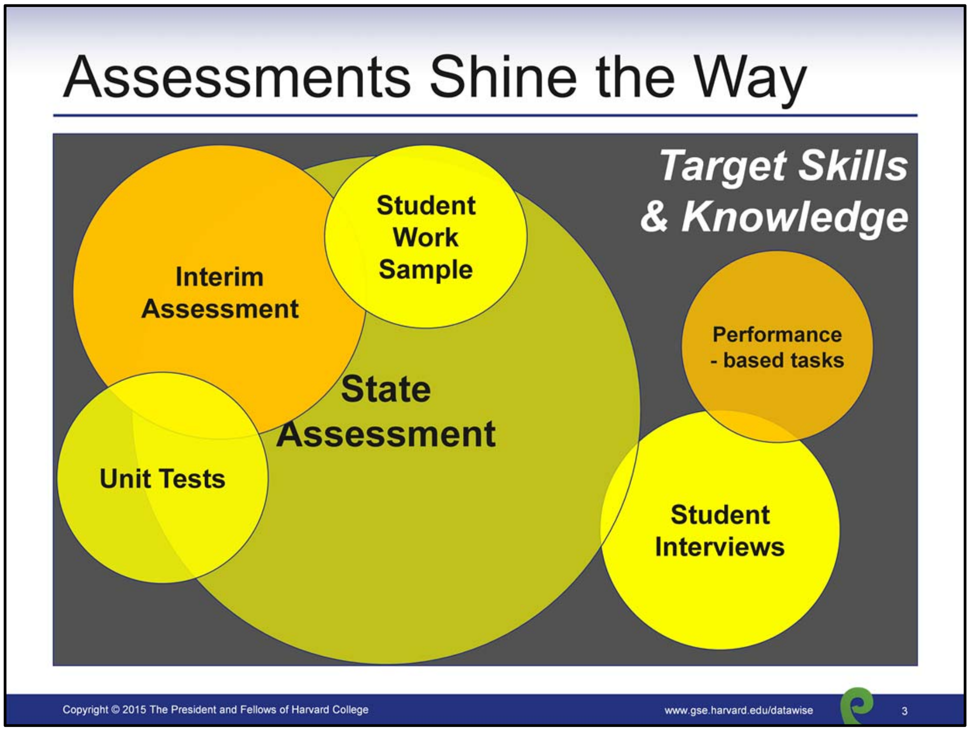
8 Act and Assess

www.gse.harvard.edu/datawise 2

Welcome to Step 2 of the Data Wise Improvement Process; building assessment literacy Like Step 1: Organizing for Collaborative Work, Step 2 is the foundation on which your cycle of inquiry is built. There are three key tasks in Step 2: 1) Reviewing Skills Tested; 2) Study How Results Are Reported and 3) Learn Principles of Responsible Data Use. Together, these are essential skills for all faculty in order to use data to improve instruction.

In this lesson, we will focus on learning principles of responsible data use. Your course team meetings will allow you the opportunity to work through Key Tasks 2.1 and 2.2 with your own school data.

In this presentation we will identify common “intuitive” misconceptions about testing data by exploring three concepts: 1) validity of inferences; 2) reliability of test scores and 3) score inflation.



Assessments Shine the Way

Target Skills & Knowledge

Interim Assessment

Student Work Sample

State Assessment

Unit Tests

Performance - based tasks

Student Interviews

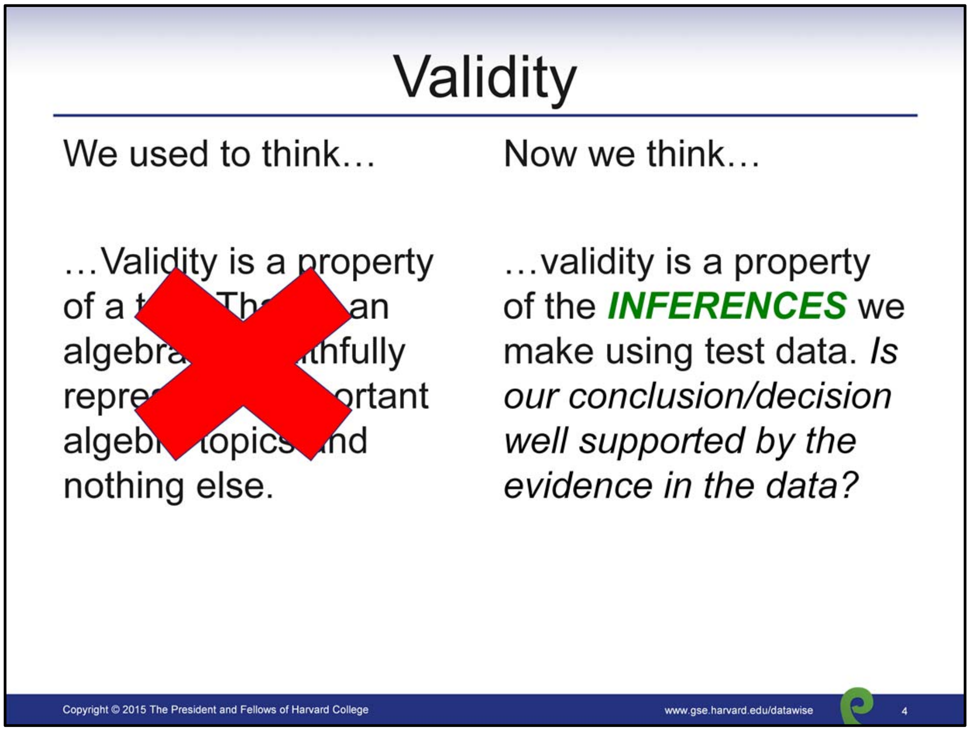
www.gse.harvard.edu/datawise 3

This gray box represents the skills and knowledge we want students to gain. Schools and teachers are inundated with testing data. From state assessment, district interim assessment, unit tests, performance-based tasks, student work in class, sometimes it can feel like we are under a student data pile and don't know where to begin!

At Data Wise, we see each one of these assessments as giving us additional information to help understand and clarify what are students are struggling to learn. We use high level assessment results, like state tests or PSAT data, to ask questions about student learning. Why is it our students are scoring below the state average in math? High level data provokes questions, but rarely provides answers. And so that's when we seek out MORE DATA. What does our interim assessment data tell us about student learner? What does a close examination of actual student work, and not just percentages correct, reveal?

By triangulated across multiple data sources, we begin to narrow in on a student learning challenge that we are confident is real and worthy of our focused attention.

In what follows, we will apply the concepts of validity and reliability to testing data to help you become a more critical consumer of the information available in these reports.



Validity

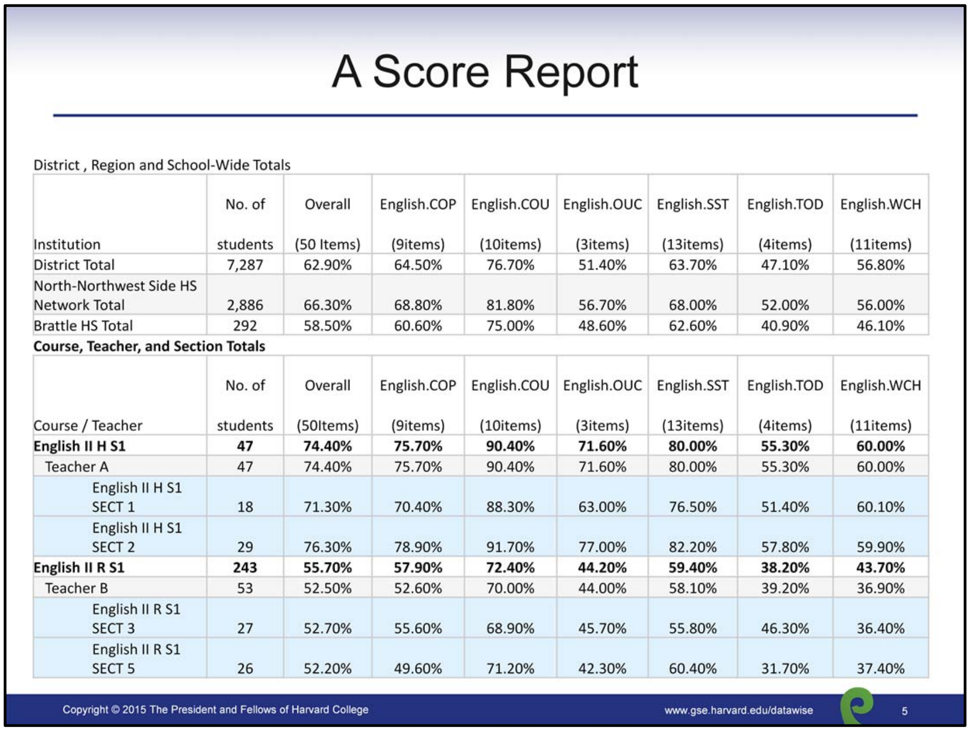
We used to think…

…Validity is a property of a t Th an algebra thfully repres ortant algebr topics nd nothing else.

Now we think…

…validity is a property of the ***INFERENCES*** we make using test data. *Is our conclusion/decision well supported by the evidence in the data?*

We used to think validity was a property of a test. That is, if an algebra test faithfully represented important algebraic topics, we would call it valid. Now, however, we think about validity as a property of the INFERENCES we make using the test data. We should be asking ourselves: is our conclusion/decision well supported by the evidence in the data? Is there other data we need to consider? Other information that may bear on our interpretation? Let’s take a look at how validity applies to reading score reports.



A Score Report

District , Region and School-Wide Totals

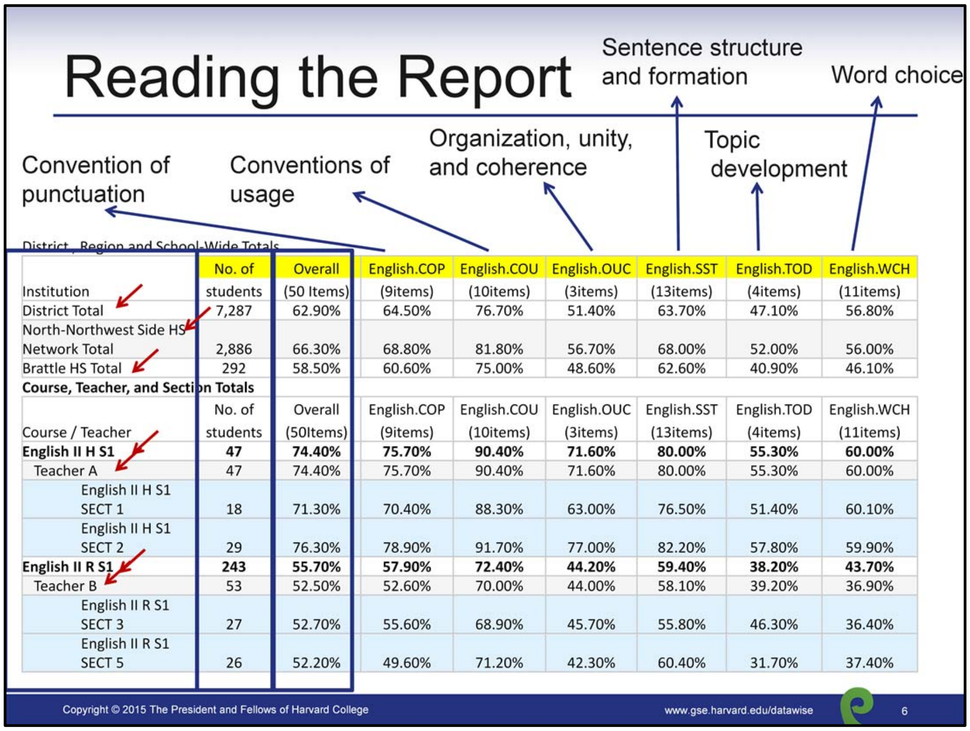
	No. of	Overall	English.COP	English.COU	English.OUC	English.SST	English.TOD	English.WCH
Institution	students	(50 Items)	(9items)	(10items)	(3items)	(13items)	(4items)	(11items)
District Total	7,287	62.90%	64.50%	76.70%	51.40%	63.70%	47.10%	56.80%
North-Northwest Side HS Network Total	2,886	66.30%	68.80%	81.80%	56.70%	68.00%	52.00%	56.00%
Brattle HS Total	292	58.50%	60.60%	75.00%	48.60%	62.60%	40.90%	46.10%

Course, Teacher, and Section Totals

	No. of	Overall	English.COP	English.COU	English.OUC	English.SST	English.TOD	English.WCH
Course / Teacher	students	(50Items)	(9items)	(10items)	(3items)	(13items)	(4items)	(11items)
English II H S1	**47**	**74.40%**	**75.70%**	**90.40%**	**71.60%**	**80.00%**	**55.30%**	**60.00%**
Teacher A	47	74.40%	75.70%	90.40%	71.60%	80.00%	55.30%	60.00%
English II H S1 SECT 1	18	71.30%	70.40%	88.30%	63.00%	76.50%	51.40%	60.10%
English II H S1 SECT 2	29	76.30%	78.90%	91.70%	77.00%	82.20%	57.80%	59.90%
English II R S1	**243**	**55.70%**	**57.90%**	**72.40%**	**44.20%**	**59.40%**	**38.20%**	**43.70%**
Teacher B	53	52.50%	52.60%	70.00%	44.00%	58.10%	39.20%	36.90%
English II R S1 SECT 3	27	52.70%	55.60%	68.90%	45.70%	55.80%	46.30%	36.40%
English II R S1 SECT 5	26	52.20%	49.60%	71.20%	42.30%	60.40%	31.70%	37.40%

Here is the state testing score report from public high school in Chicago. Data reports such as these can be overwhelming at first, which is why Key Task 2.2: Study how results are reported is critical to helping teachers engage with data.

Let's walk through some key attributes of this report.



Reading the Report

Sentence structure and formation

Word choice

Organization, unity, and coherence

Topic development

Convention of punctuation

Conventions of usage

District , Region and School-Wide Totals

	No. of	Overall	English.COP	English.COU	English.OUC	English.SST	English.TOD	English.WCH
Institution	students	(50 Items)	(9items)	(10items)	(3items)	(13items)	(4items)	(11items)
District Total	7,287	62.90%	64.50%	76.70%	51.40%	63.70%	47.10%	56.80%
North-Northwest Side HS Network Total	2,886	66.30%	68.80%	81.80%	56.70%	68.00%	52.00%	56.00%
Brattle HS Total	292	58.50%	60.60%	75.00%	48.60%	62.60%	40.90%	46.10%

Course, Teacher, and Section Totals

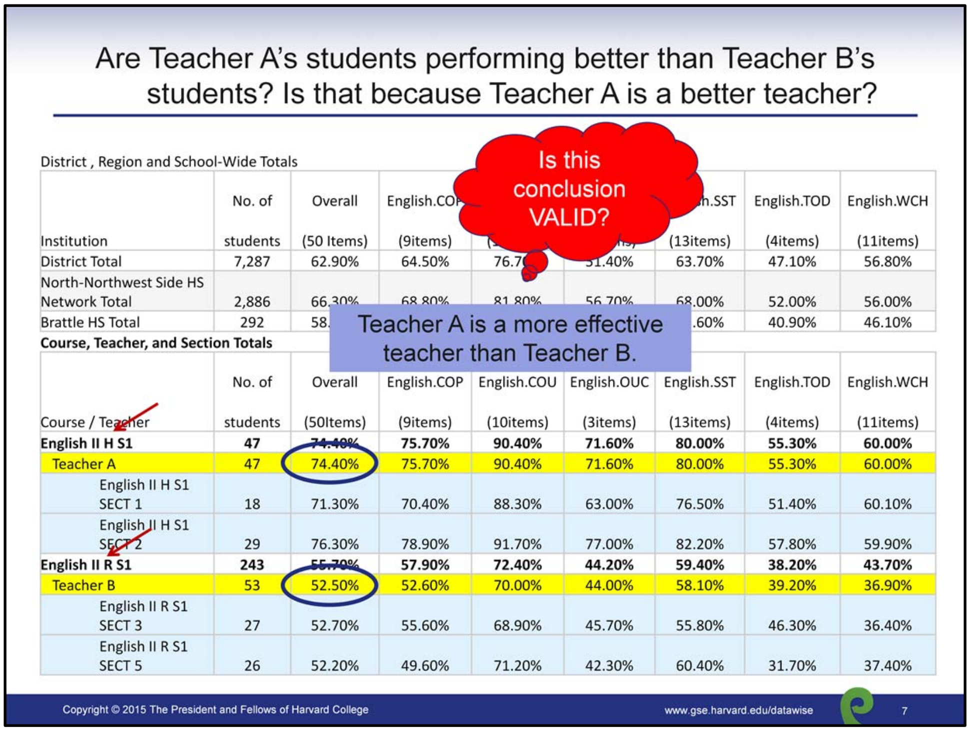
	No. of	Overall	English.COP	English.COU	English.OUC	English.SST	English.TOD	English.WCH
Course / Teacher	students	(50Items)	(9items)	(10items)	(3items)	(13items)	(4items)	(11items)
English II H S1	**47**	**74.40%**	**75.70%**	**90.40%**	**71.60%**	**80.00%**	**55.30%**	**60.00%**
Teacher A	47	74.40%	75.70%	90.40%	71.60%	80.00%	55.30%	60.00%
English II H S1 SECT 1	18	71.30%	70.40%	88.30%	63.00%	76.50%	51.40%	60.10%
English II H S1 SECT 2	29	76.30%	78.90%	91.70%	77.00%	82.20%	57.80%	59.90%
English II R S1	**243**	**55.70%**	**57.90%**	**72.40%**	**44.20%**	**59.40%**	**38.20%**	**43.70%**
Teacher B	53	52.50%	52.60%	70.00%	44.00%	58.10%	39.20%	36.90%
English II R S1 SECT 3	27	52.70%	55.60%	68.90%	45.70%	55.80%	46.30%	36.40%
English II R S1 SECT 5	26	52.20%	49.60%	71.20%	42.30%	60.40%	31.70%	37.40%

www.gse.harvard.edu/datawise 6

First notice the headings on the top row. This is an English language arts report and the the headings indicate the domains tested, such as conventions of punctuation and conventions of usage. In addition, the first column list particular subgroups of students, the second column represents the number of students tested and the third column represents the overall average score.

Now let’s examine the first column. We can see scores are being reported for the district, a region and our school, Brattle HS. Going further down, we see that scores are further disaggregated. First by course sections and then by teacher.

Now let’s turn our attention to the rows titled Teacher A and Teacher B.



Are Teacher A’s students performing better than Teacher B’s students? Is that because Teacher A is a better teacher?

Is this conclusion VALID?

Teacher A is a more effective teacher than Teacher B.

District , Region and School-Wide Totals

Institution	No. of students	Overall (50 Items)	English.COP (9items)	[illegible]	[illegible]	...h.SST (13items)	English.TOD (4items)	English.WCH (11items)
District Total	7,287	62.90%	64.50%	76.7	51.40%	63.70%	47.10%	56.80%
North-Northwest Side HS Network Total	2,886	66.30%	68.80%	81.80%	56.70%	68.00%	52.00%	56.00%
Brattle HS Total	292	58.	[illegible]	[illegible]	[illegible]	.60%	40.90%	46.10%

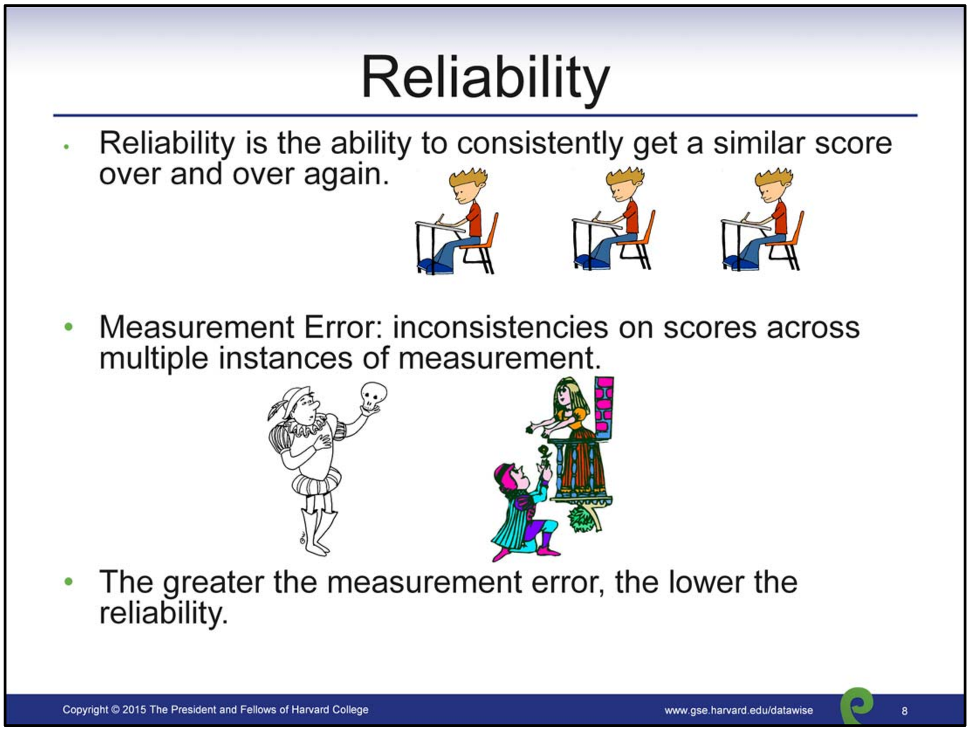
Course, Teacher, and Section Totals

Course / Teacher	No. of students	Overall (50Items)	English.COP (9items)	English.COU (10items)	English.OUC (3items)	English.SST (13items)	English.TOD (4items)	English.WCH (11items)
English II H S1	**47**	**74.40%**	**75.70%**	**90.40%**	**71.60%**	**80.00%**	**55.30%**	**60.00%**
Teacher A	47	74.40%	75.70%	90.40%	71.60%	80.00%	55.30%	60.00%
English II H S1 SECT 1	18	71.30%	70.40%	88.30%	63.00%	76.50%	51.40%	60.10%
English II H S1 SECT 2	29	76.30%	78.90%	91.70%	77.00%	82.20%	57.80%	59.90%
English II R S1	**243**	**55.70%**	**57.90%**	**72.40%**	**44.20%**	**59.40%**	**38.20%**	**43.70%**
Teacher B	53	52.50%	52.60%	70.00%	44.00%	58.10%	39.20%	36.90%
English II R S1 SECT 3	27	52.70%	55.60%	68.90%	45.70%	55.80%	46.30%	36.40%
English II R S1 SECT 5	26	52.20%	49.60%	71.20%	42.30%	60.40%	31.70%	37.40%

www.gse.harvard.edu/datawise 7

Upon first inspection I notice that the students in Teacher A’s class on average are scoring higher than the students in Teacher B’s class, as the average score in Teacher A’s class is 74.40% and the average score in Teacher B’s class is 52.5%. Notice that I am staying low on the ladder of inference. However, I might start to climb the ladder and draw a conclusion by saying Teacher A is a more effective teacher than Teacher B. But is this conclusion VALID? Are the average percent correct enough evidence to make this claim?

If we look a little closer, we might also notice that Teacher A’s courses are labeled with an H while Teacher B’s courses are labeled with an R. In fact, the H stands for “honors” section and the R stands for “regular” section. With this information, we should begin to question the claim that Teacher A is a more effective teacher than Teacher B as we know that Teacher A is teaching an honor’s class, which has *preselected* high scoring students. Indeed, it is not a valid inference based on how students are grouped. ***We’ll need more evidence than this score report provides*** to decide the extent to which higher scores are due to Teacher A rather than the sample of students in Teacher A’s class.



Reliability

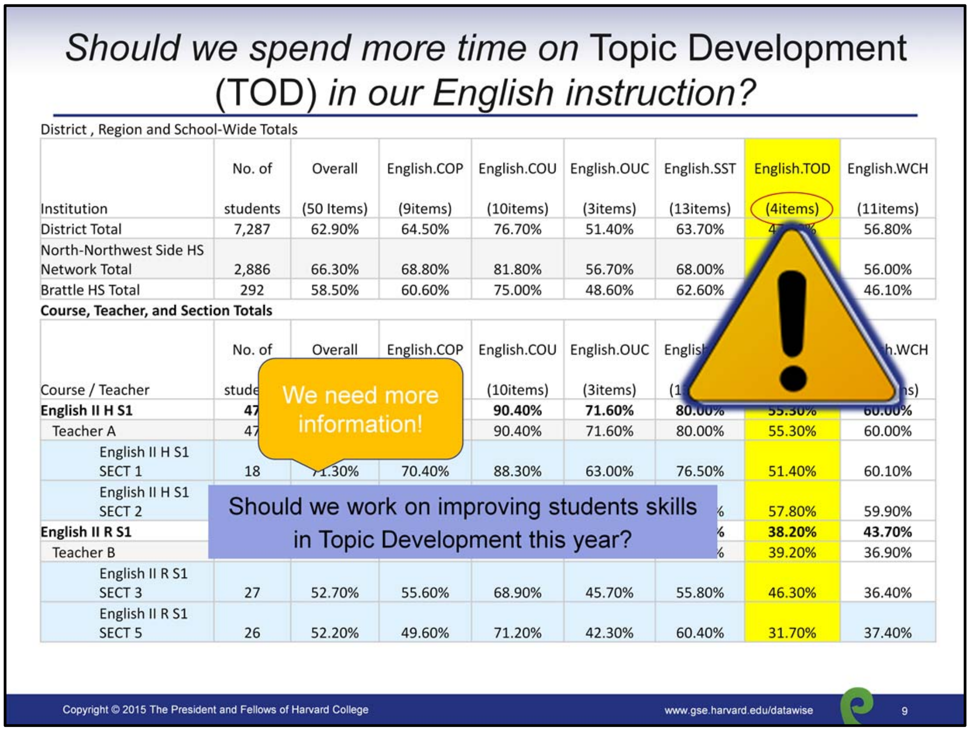
- Reliability is the ability to consistently get a similar score over and over again.
- Measurement Error: inconsistencies on scores across multiple instances of measurement.
- The greater the measurement error, the lower the reliability.

www.gse.harvard.edu/datawise 8

Now let's turn our attention to the concept of reliability. Reliability describes how ***consistent*** we expect a student's or group of student's score to be on a test. We ask ourselves two questions: 1) If a student took an assessment again, how likely is it they would get a similar score? 2) If a group of students took an assessment again, how likely is it they would get a similar average score? For example, if a student took the SAT three weekends in a row, would we expect the scores on each take to be very close to one another or very far? If the scores are very close to one another, we consider this a very reliable test.

Another example: let's say the 10th grade at Adams high school takes a state English Language arts exam every year. Do we expect the average score from year to year to be very similar to each other or very different? The answer to this question relies on multiple factors. For example, in a small school, average scores tend to vary a lot from year to year as a result of the particular group of students taking the test. For example, perhaps one year the school just happens to have a few more high achieving students than the previous year and this causes the average score to “jump up”. In larger schools, however, such changes in the make up of the class get washed out on average. Thus the average test score in a large school is more stable than a small school and hence more reliable and less likely to change from year to year.

Measurement error also effects the reliability of scores. Sampling of items, personal student factors and environmental factors all effect measurement error. For example: sometimes a particular test just happens to have questions on content very familiar or unfamiliar to students. Perhaps an English Language Arts tests asks a questions about a Hamlet passage that students had recently analyzed in class. We would expect these students to do particularly well on these prompts compared to students who had read Romeo and Juliet as opposed to Hamlet. But does this mean that those students are really more proficient in English Language Arts skills more broadly? The more measurement error in a test score, the lower the reliability.



Should we spend more time on Topic Development (TOD) *in our English instruction?*

District , Region and School-Wide Totals

Institution	No. of students	Overall (50 Items)	English.COP (9items)	English.COU (10items)	English.OUC (3items)	English.SST (13items)	English.TOD (4items)	English.WCH (11items)
District Total	7,287	62.90%	64.50%	76.70%	51.40%	63.70%	4	56.80%
North-Northwest Side HS Network Total	2,886	66.30%	68.80%	81.80%	56.70%	68.00%		56.00%
Brattle HS Total	292	58.50%	60.60%	75.00%	48.60%	62.60%		46.10%

Course, Teacher, and Section Totals

Course / Teacher	No. of stude	Overall	English.COP	English.COU (10items)	English.OUC (3items)	Englis (1		h.WCH ns)
English II H S1	**47**			**90.40%**	**71.60%**	**80.00%**	**55.30%**	**60.00%**
Teacher A	47			90.40%	71.60%	80.00%	55.30%	60.00%
English II H S1 SECT 1	18	1.30%	70.40%	88.30%	63.00%	76.50%	51.40%	60.10%
English II H S1 SECT 2						%	57.80%	59.90%
English II R S1						%	**38.20%**	**43.70%**
Teacher B						%	39.20%	36.90%
English II R S1 SECT 3	27	52.70%	55.60%	68.90%	45.70%	55.80%	46.30%	36.40%
English II R S1 SECT 5	26	52.20%	49.60%	71.20%	42.30%	60.40%	31.70%	37.40%

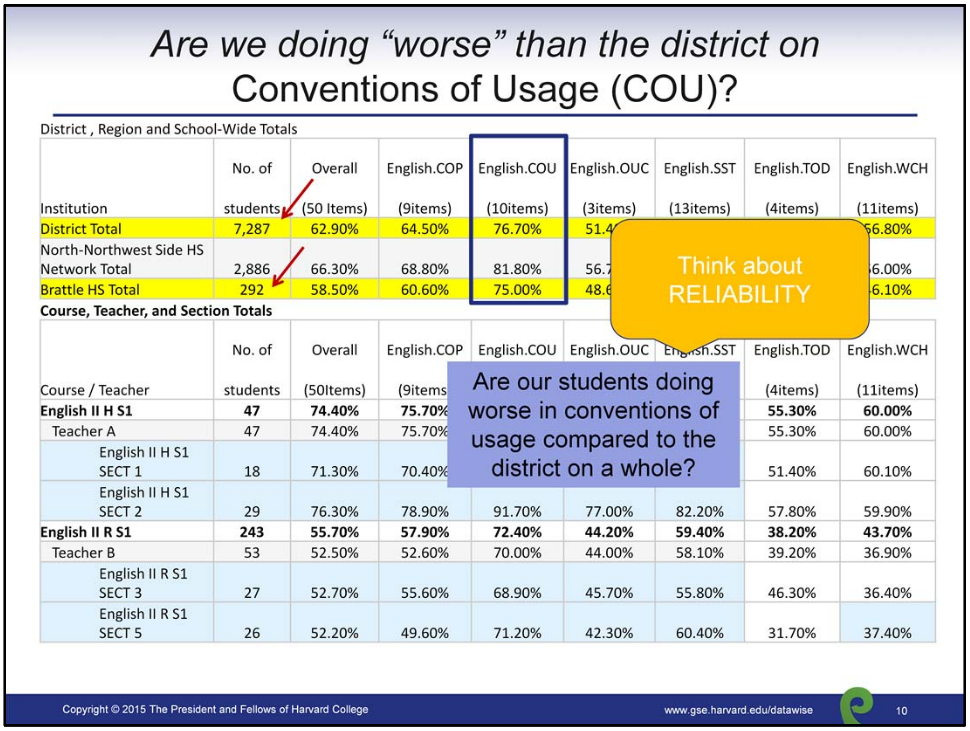
Should we work on improving students skills in Topic Development this year?

www.gse.harvard.edu/datawise 9

Let's apply the concept of reliability when interpreting test results. Teachers may look at this test report and be drawn to the Topic Development column highlighted here. Indeed, students at Brattle High School got less than 50% of the questions correct on average. It is Brattle's lowest scoring strand. The English Department may ask itself: Should we work on improving students skills in Topic Development this year?

Before deciding, however, a savy english department will consider more information. First, notice that there are only 4 items in Topic Development, less than many other strands. Thus one unusually difficult or poorly phrased item could be driving the low average. Second, topic development is not just the lowest strand for Brattle, but it is the lowest strand for the entire state. Before taking on topic development as a focus, the english department may consider more information. If available, they may examine the actual items on the test. Doing so will allow them to consider how important these skills are for their students future in comparison to other strands. It may be that other strands represent more important skills to work on with their students.

The point is this: Beware of drawing quick conclusions about strands with small sample of questions. There are only four items in that domain, and these items could be designed to be difficult. Comparing percentages of correct items across domains may not be sufficient to decide where we should spend our time.



Are we doing "worse" than the district on Conventions of Usage (COU)?

District , Region and School-Wide Totals

Institution	No. of students	Overall (50 Items)	English.COP (9items)	English.COU (10items)	English.OUC (3items)	English.SST (13items)	English.TOD (4items)	English.WCH (11items)
District Total	7,287	62.90%	64.50%	76.70%	51.4	[illegible]	[illegible]	56.80%
North-Northwest Side HS Network Total	2,886	66.30%	68.80%	81.80%	56.7	[illegible]	[illegible]	6.00%
Brattle HS Total	292	58.50%	60.60%	75.00%	48.6	[illegible]	[illegible]	6.10%

Course, Teacher, and Section Totals

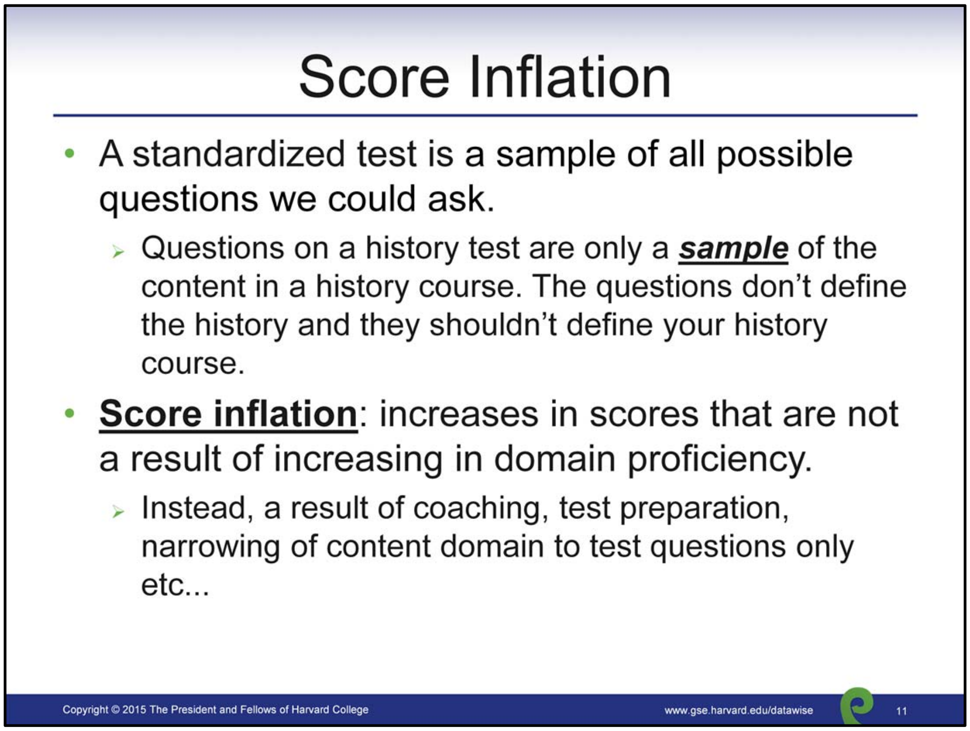
Course / Teacher	No. of students	Overall (50Items)	English.COP (9items	English.COU	English.OUC	English.SST	English.TOD (4items)	English.WCH (11items)
English II H S1	**47**	**74.40%**	**75.70%**	[illegible]	[illegible]	[illegible]	**55.30%**	**60.00%**
Teacher A	47	74.40%	75.70%	[illegible]	[illegible]	[illegible]	55.30%	60.00%
English II H S1 SECT 1	18	71.30%	70.40%	[illegible]	[illegible]	[illegible]	51.40%	60.10%
English II H S1 SECT 2	29	76.30%	78.90%	91.70%	77.00%	82.20%	57.80%	59.90%
English II R S1	**243**	**55.70%**	**57.90%**	**72.40%**	**44.20%**	**59.40%**	**38.20%**	**43.70%**
Teacher B	53	52.50%	52.60%	70.00%	44.00%	58.10%	39.20%	36.90%
English II R S1 SECT 3	27	52.70%	55.60%	68.90%	45.70%	55.80%	46.30%	36.40%
English II R S1 SECT 5	26	52.20%	49.60%	71.20%	42.30%	60.40%	31.70%	37.40%

Think about RELIABILITY

Are our students doing worse in conventions of usage compared to the district on a whole?

www.gse.harvard.edu/datawise 10

Administrators and teachers often look at assessment reports and compare how their students are doing to others. For example, we might look at this data report and wonder: Are our students doing worse in conventions of usage compared to the district on a whole? This is another opportunity to consider reliability. The district has many more students, and thus their average is much more stable. Brattle can expect their average to bounce around more year to year because they have fewer students. These percentages are so close, it's easy to imagine the percentage changing as a result of which particular group students took the test. Especially be cautious with sub group data that is based on a small number of students as these averages are often unreliable



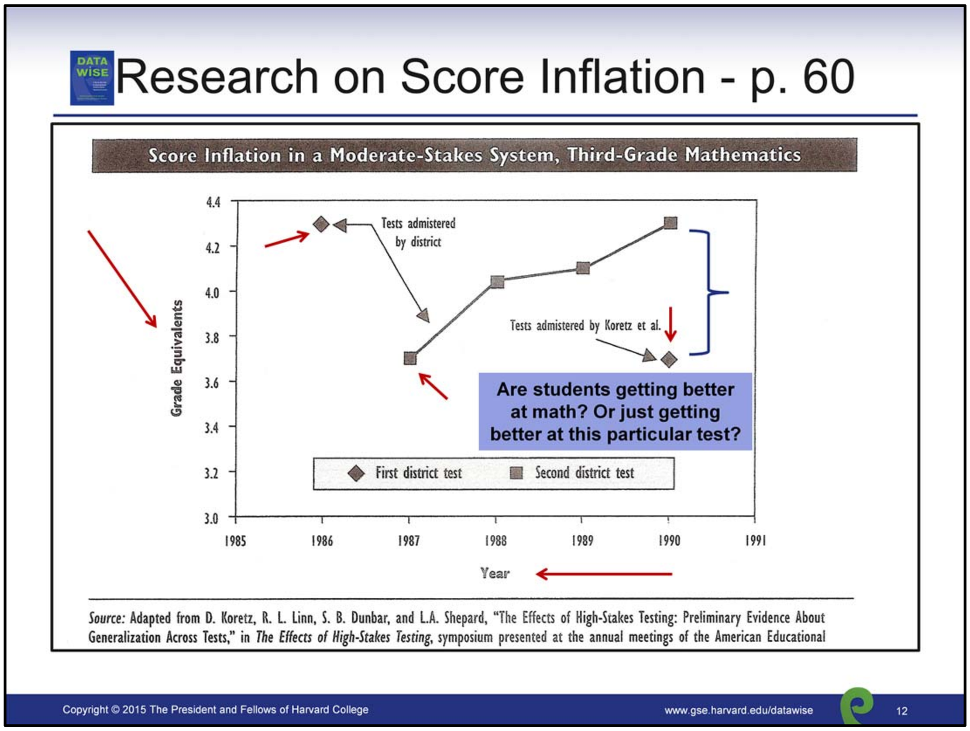
Score Inflation

- A standardized test is a sample of all possible questions we could ask.
  - Questions on a history test are only a ***sample*** of the content in a history course. The questions don’t define the history and they shouldn’t define your history course.
- **Score inflation**: increases in scores that are not a result of increasing in domain proficiency.
  - Instead, a result of coaching, test preparation, narrowing of content domain to test questions only etc...

www.gse.harvard.edu/datawise 11

Finally, we want you to consider score inflation. Any standardized tests is a sample of all possible questions we could ask. For example, questions on a history test are only a sample of the content in a history course. The questions don't define history and they shouldn't define your course.

Score inflation refers to any increase in scores that are NOT a result of increases in domain proficiency. For example, here in Massachusetts our 8th grade state exam asks many questions about the slope of a line. As a teacher, there are a few questions I can predict occur every year on the exam. I might spend class time making sure students practice that particular item type over and over again until they know exactly what procedure to apply to get it right. Their scores are likely going to go up as a result. But have they really learned more about slope? Is their understanding of slope truly improved? Or have they just memorized how to answer one narrow type of question? The latter is score inflation.



Research on Score Inflation - p. 60

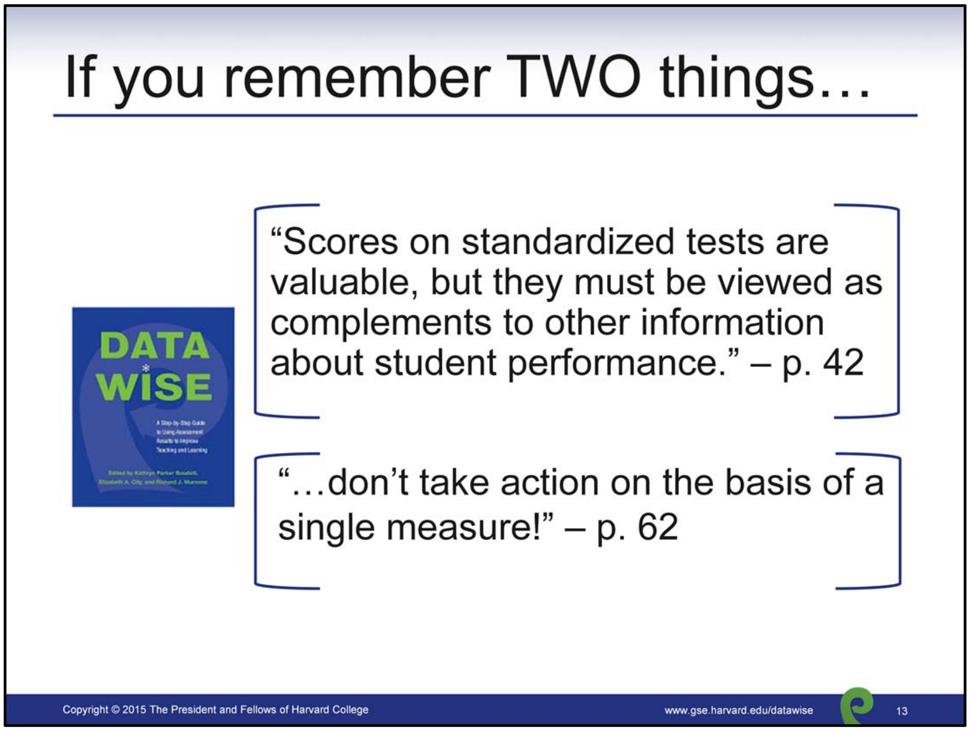
Score Inflation in a Moderate-Stakes System, Third-Grade Mathematics

Are students getting better at math? Or just getting better at this particular test?

Source: Adapted from D. Koretz, R. L. Linn, S. B. Dunbar, and L.A. Shepard, "The Effects of High-Stakes Testing: Preliminary Evidence About Generalization Across Tests," in *The Effects of High-Stakes Testing*, symposium presented at the annual meetings of the American Educational

www.gse.harvard.edu/datawise 12

Dan Koretz at Harvard University has studied score inflation. In this graph we present some of his research. This graph display students performance on a third grade mathematics test. On the x-axis are years and on the y-axis is students' average performance on a state test, measured in grade-level equivalents. The first diamond represents the average score on the state test in 1986. In 1987 the state changed the format and items on the test. The scores plummeted, as you can see from the square. Does this mean students suddenly got worse at math? Or did they just get worse at the skills of taking the test? Over the next 3 years score steadily increased. So are students getting better at math? Or just getting better at this particular test? In 1990 Dan Koretz's team also administered students the "old version" of the test. We see that represented by the diamond on the right. Notice the big gap between the new test performance and old test performance in 1990. If students were really getting better at math, we would expect these two averages to be similar. Indeed, this gap is evidence that the increases in score between 1987 and 1990 are likely about students getting better at a test, not students getting better at math.



DATA WISE
A Step-by-Step Guide to Using Assessment Results to Improve Teaching and Learning
Edited by Kathryn Parker Boudett, Elizabeth A. City, and Richard J. Murnane

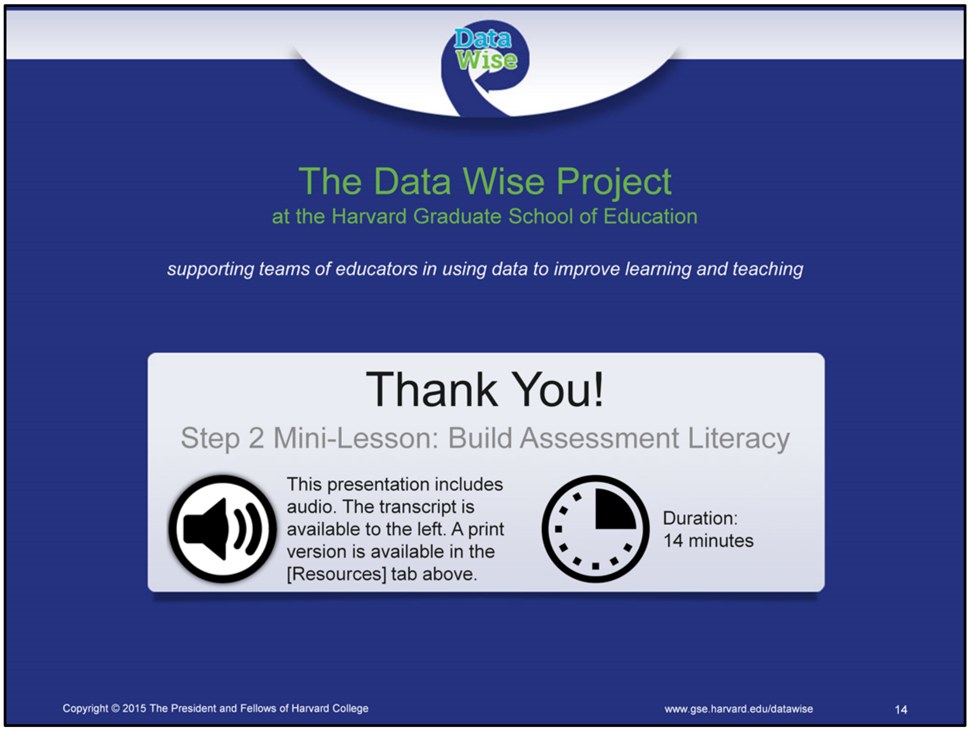
If you remember TWO things…

“Scores on standardized tests are valuable, but they must be viewed as complements to other information about student performance.” – p. 42

“…don’t take action on the basis of a single measure!” – p. 62

Copyright © 2015 The President and Fellows of Harvard College www.gse.harvard.edu/datawise 13

So how can you use data wisely to improve student learning? In the end you want to put in a place a process that encourage you to triangulate across data sources before taking action. We leave you with two powerful quotations from Data Wise.



The Data Wise Project

at the Harvard Graduate School of Education

supporting teams of educators in using data to improve learning and teaching

Thank You!

Step 2 Mini-Lesson: Build Assessment Literacy

This presentation includes audio. The transcript is available to the left. A print version is available in the [Resources] tab above.

Duration:
14 minutes