Evidence of Student Growth from Common District Assessments: A Closer Look at Effect Size

Jack B. Monpas-Huber, Ph.D.
Director of Assessment & Student Information

SHORELINE PUBLIC SCHOOLS

Observations

1. Achieving a standard is no longer enough. Now we must also show evidence of student growth for teacher evaluations.
2. We lack a consistent definition of “low, medium, and high” growth.
3. We’re beleaguered by a different metrics of growth for different assessments:
    - DIBELS (between-window increases correct sounds/words per minute toward established benchmarks)
    - easyCBM (between-window increases in total raw scores, percentiles)
    - SRI (increases in Lexile scale scores toward established benchmarks)
    - STAR Math (increases in scale scores)
4. In the spirit of PLC, we’d like to make comparisons of growth among sites using the same assessment
5. We want to be able to say we’re “having an effect”

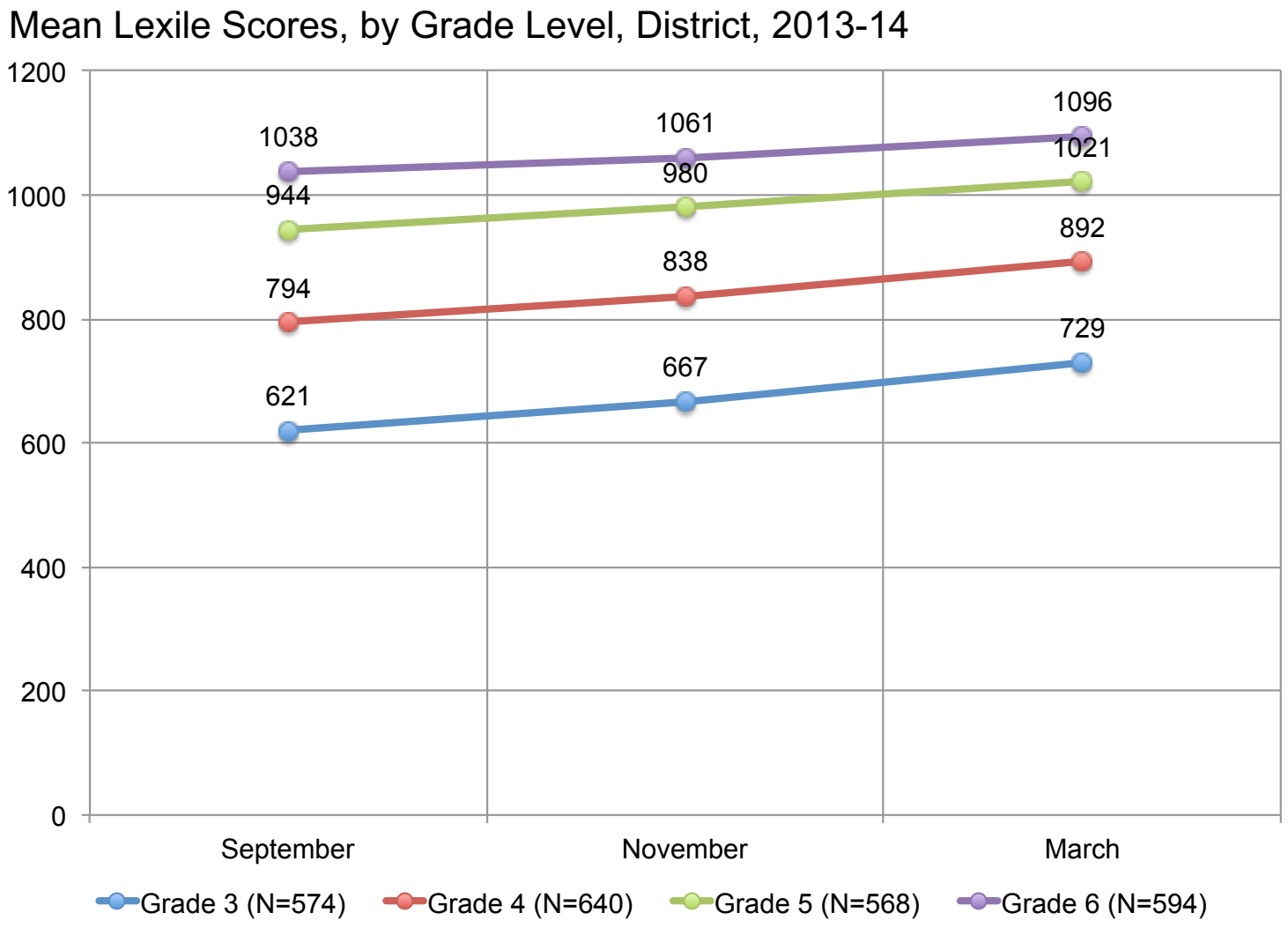
SHORELINE PUBLIC SCHOOLS

Purpose of this presentation

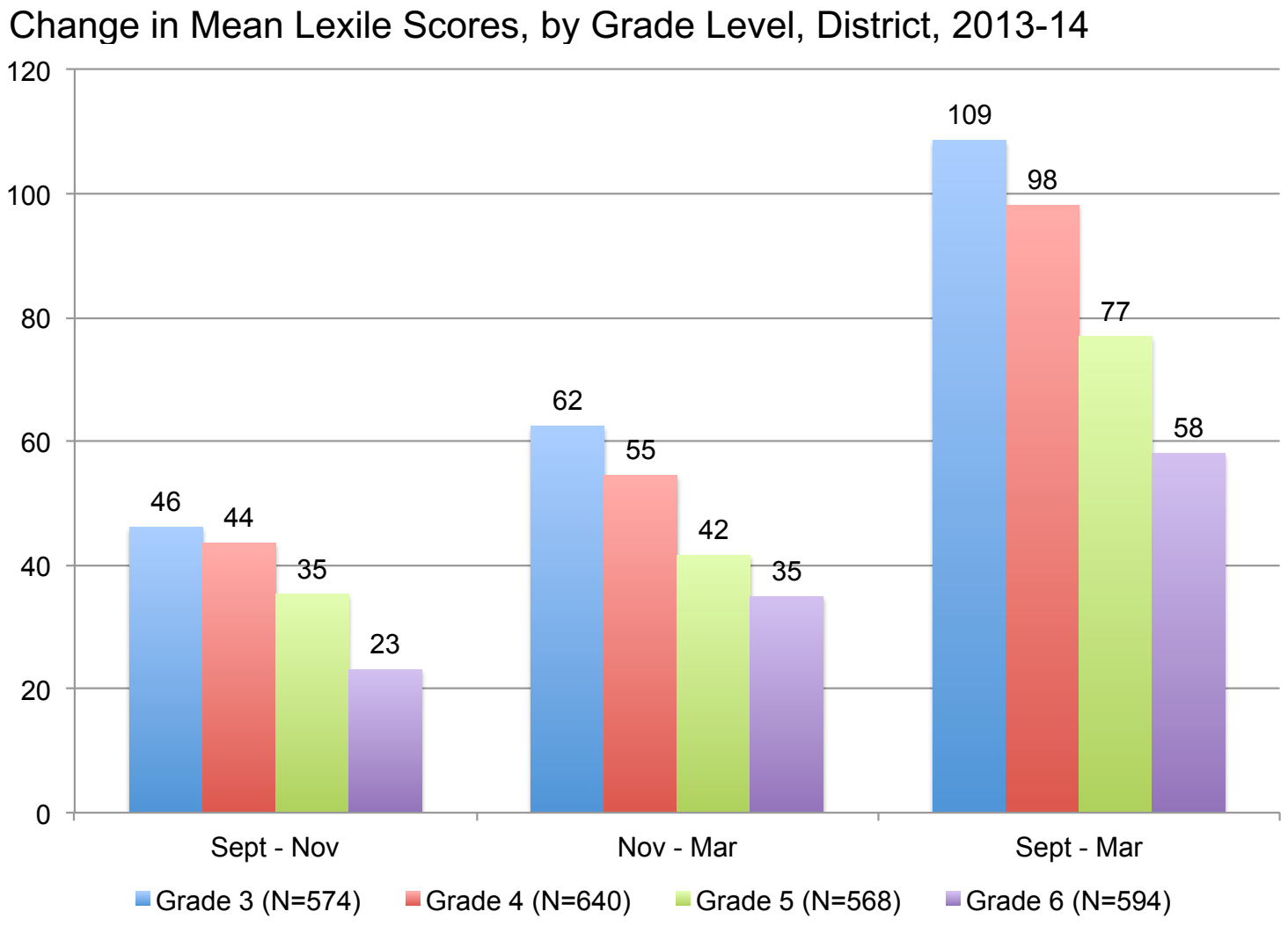
1. Board Priority #1, 4d. Use common assessments (summative and formative) to inform instruction: *"Structure principal meetings to discuss PLC use of common assessment data to inform instruction and interventions . . ."*
2. Provide principals with a sense of "normal growth" on primary District common assessments
3. Introduce effect size as a a way to think and talk about "low, medium, high" growth, school effects, teacher effects, and group gaps

SHORELINE PUBLIC SCHOOLS

What's "normal growth" in Lexile comprehension from September to March?

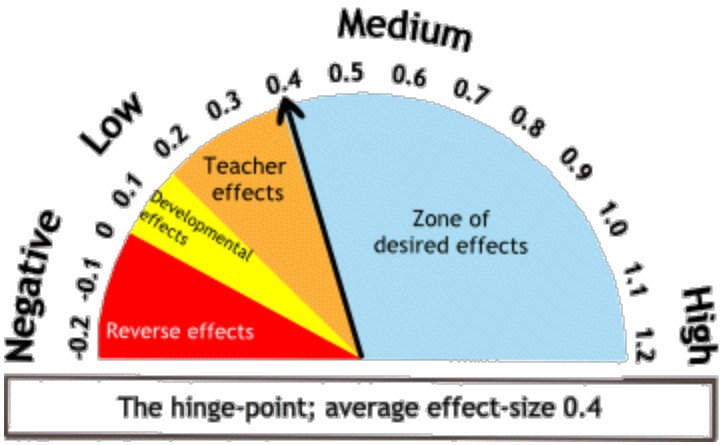


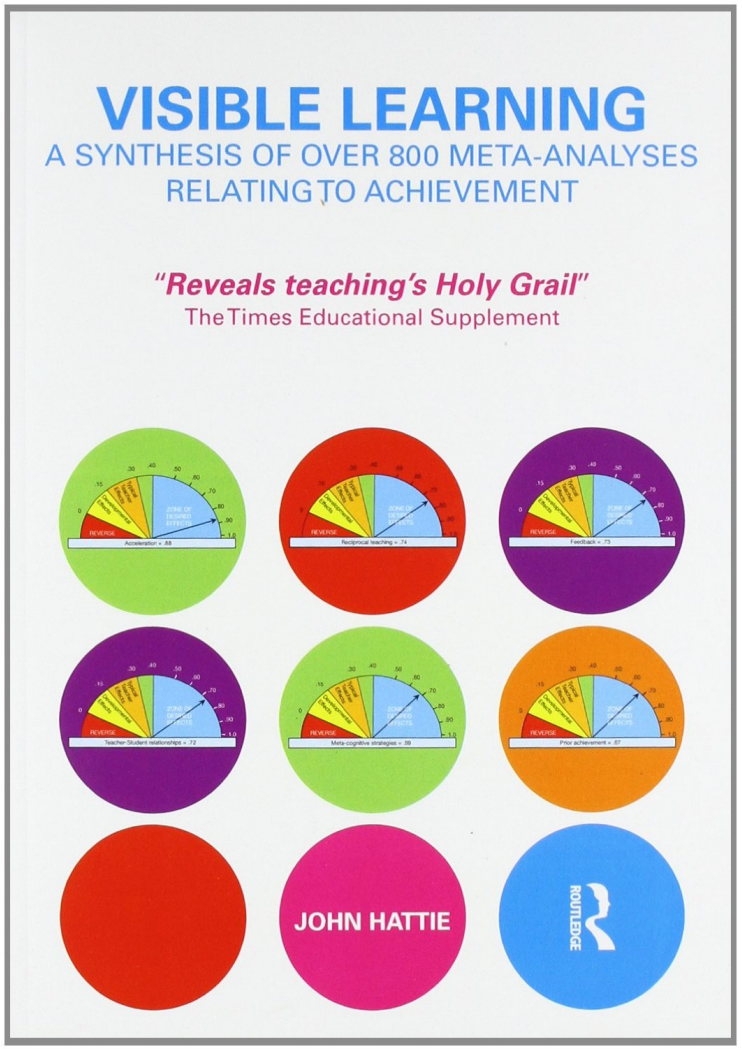
What's "normal growth" in Lexile comprehension from September to March?



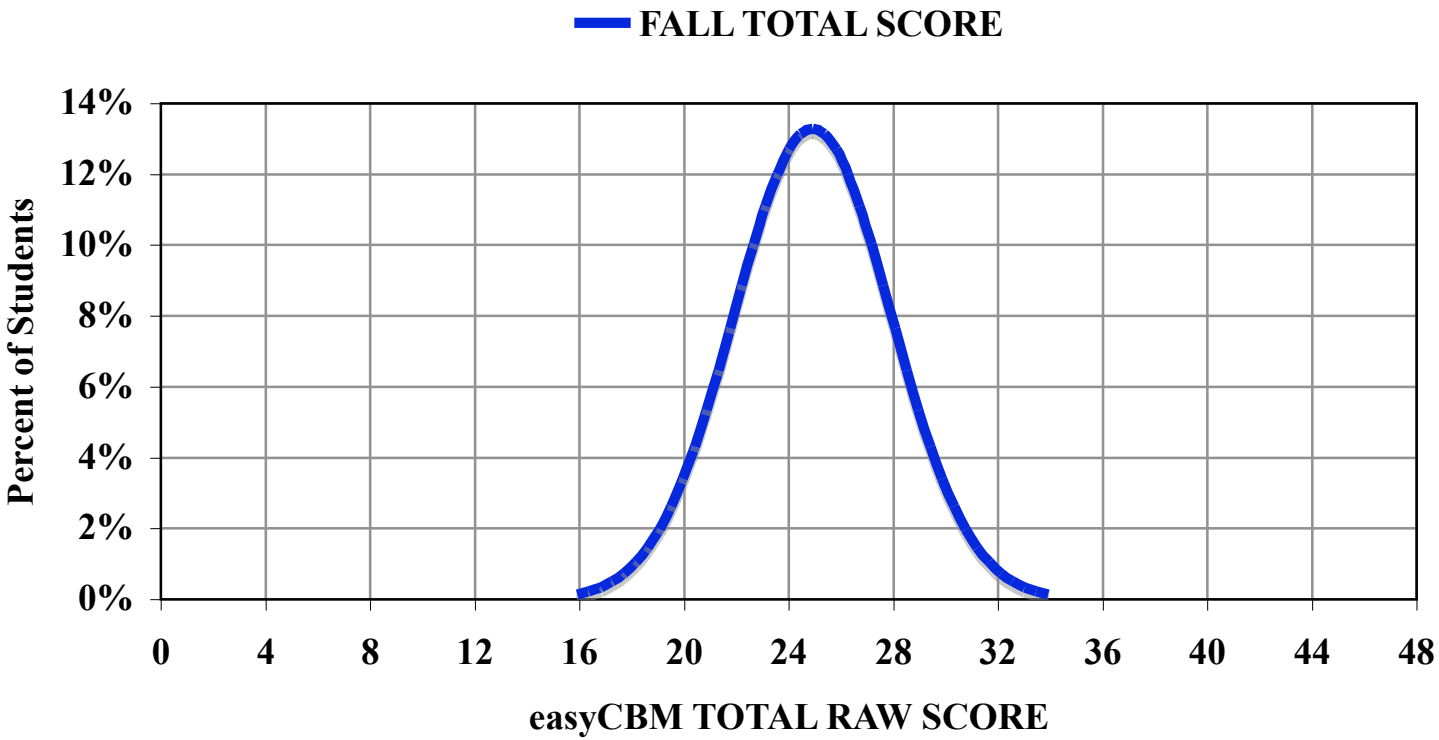
Research on growth, and the notion of growth as “effect size”

Hattie, J. A. C. (2009). *Visible learning*. New York, NY: Routledge.



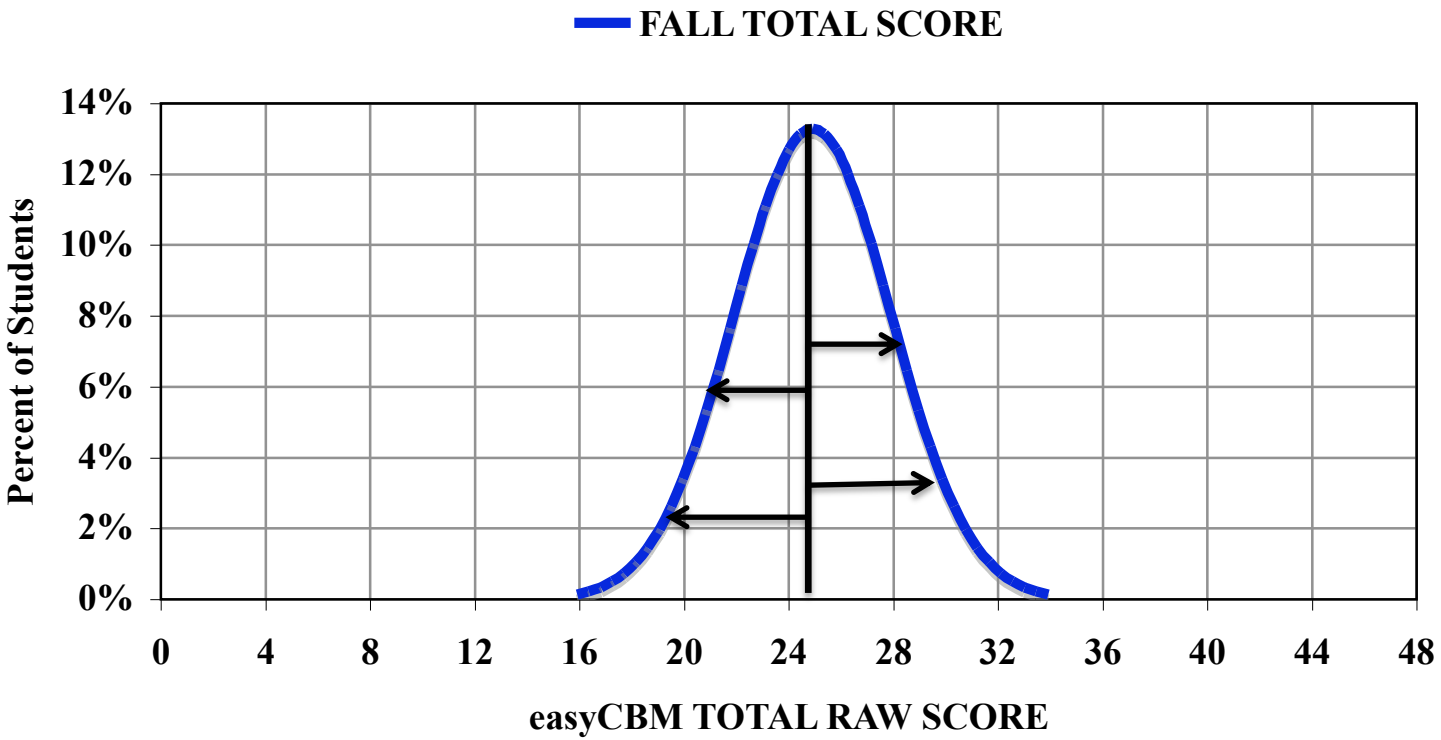


The pretest (or baseline)



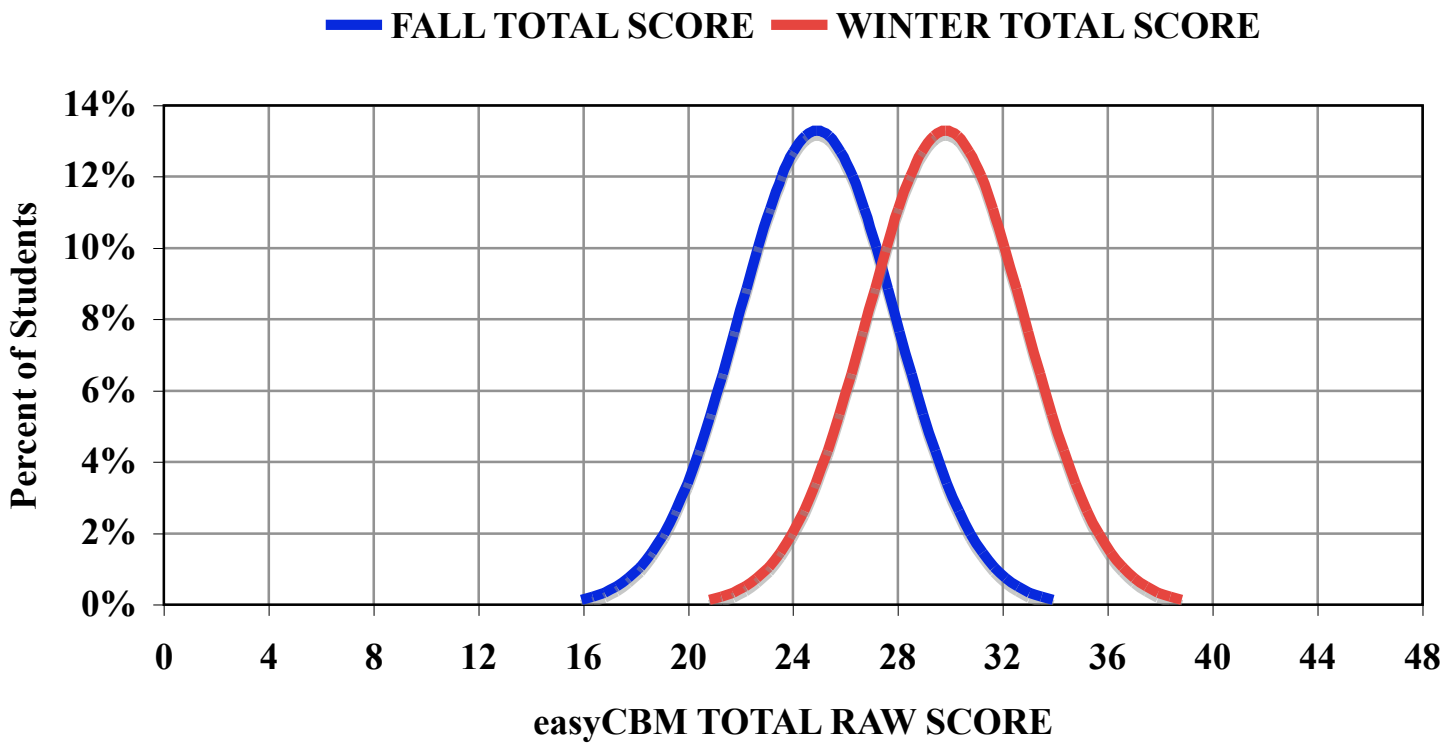
Students take a pretest. Their scores vary, but the mean is 25.

About the normal distribution: Central tendency and dispersion



- The **mean** is a single value that summarizes the group
- The **standard deviation** is a single value that summarizes the *dispersion* – of how far the average student deviates from the mean.

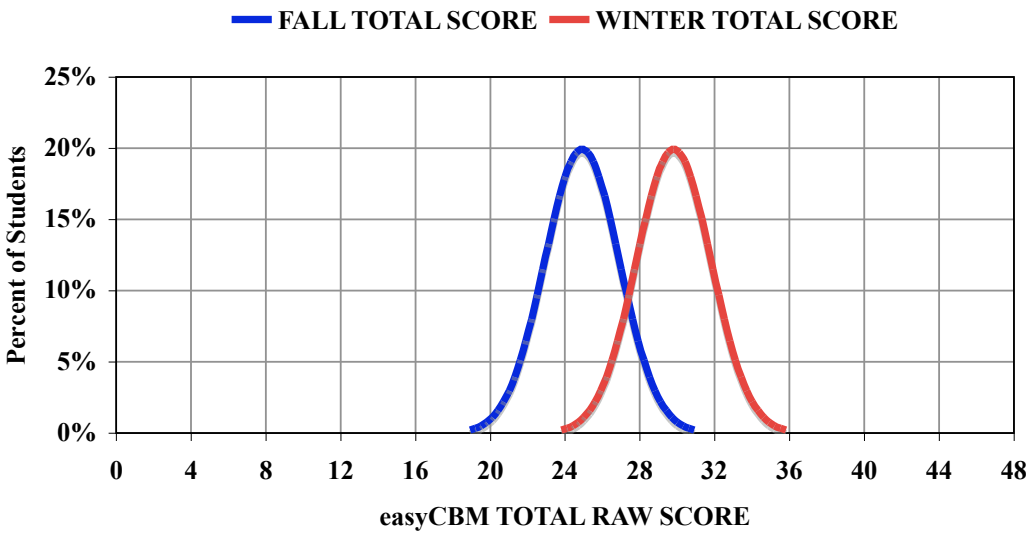
The posttest: Outcome of learning



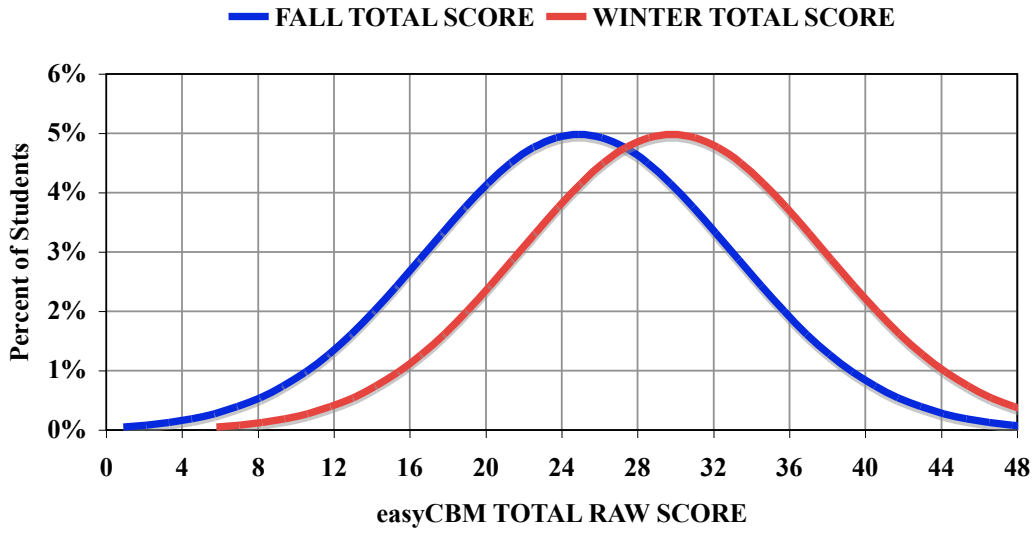
The same students take the winter test (or other posttest). Every student has learned, and scored higher. The **winter mean is 30** points. The average increased by **5 points**.

Decision Point: Is this a good thing? Did every child grow?

Which is the more significant growth?



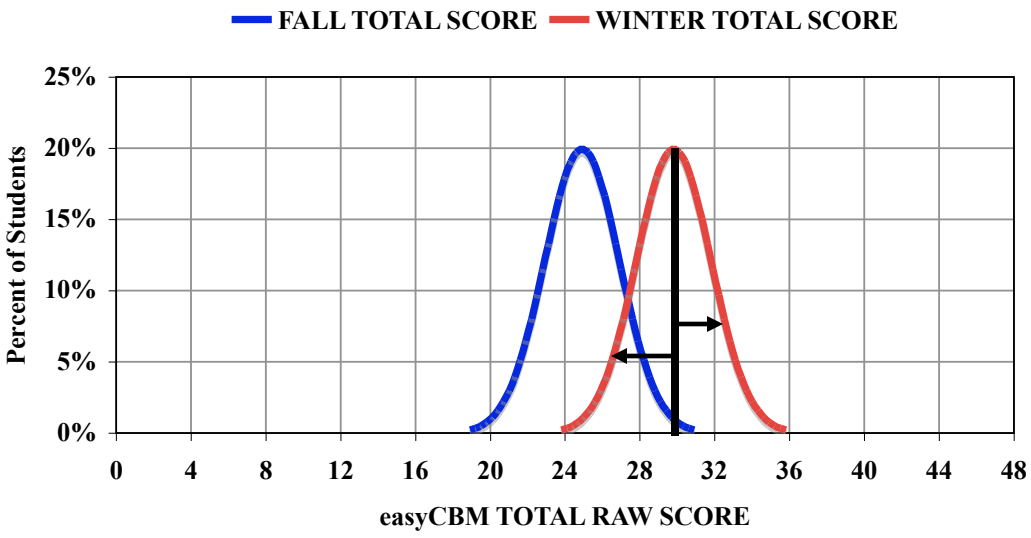
Mean		
Fall	Spring	Change
25	30	+5.0



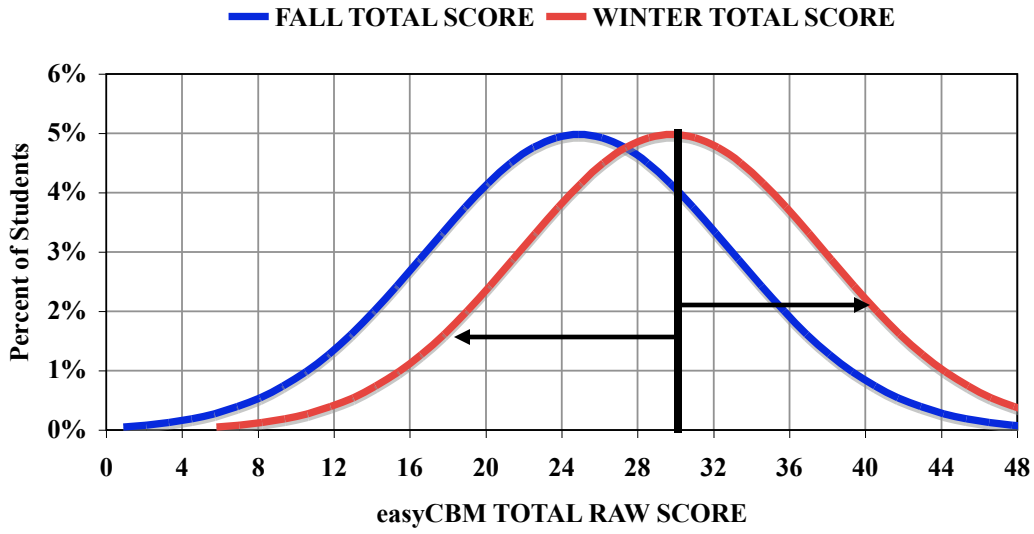
Mean		
Fall	Spring	Change
25	30	+5.0

Imagine now that the scores varied more (were more spread out), but still an average gain of 5 points. Which is the more significant gain?

Which is the more significant growth?

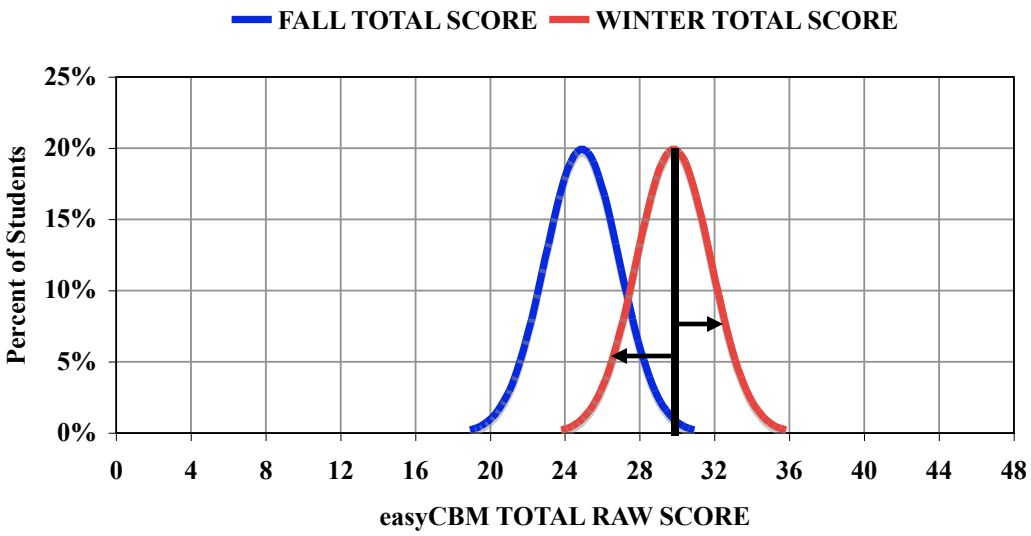


SD	Mean		
	Fall	Spring	Change
3	24.9	29.9	+5.0



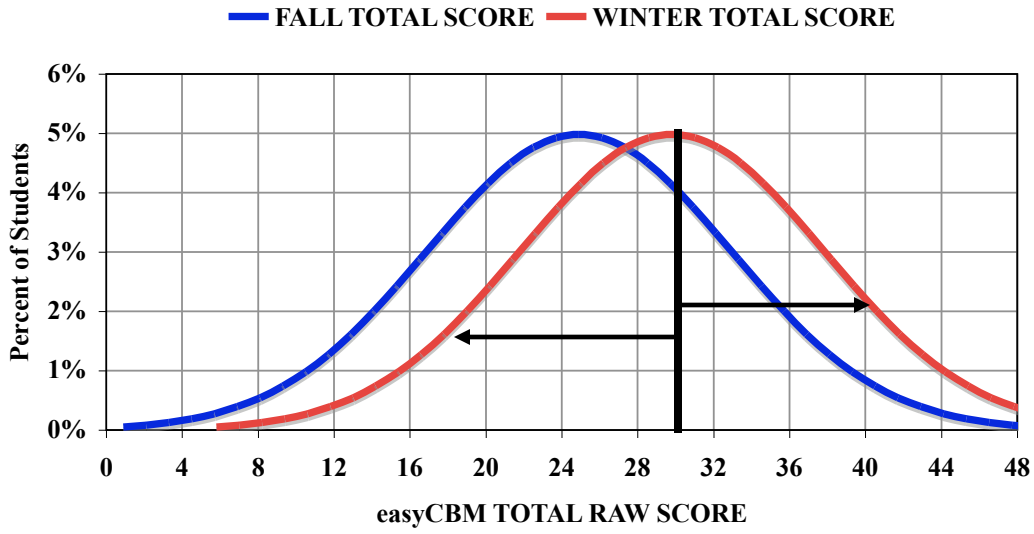
SD	Mean		
	Fall	Spring	Change
8	24.9	29.9	+5.0

Which is the more significant growth?



SD	Mean		
	Fall	Spring	Change
3	24.9	29.9	+5.0

$$\frac{\text{Change}}{\text{SD}} = \frac{5}{3} = 1.68$$



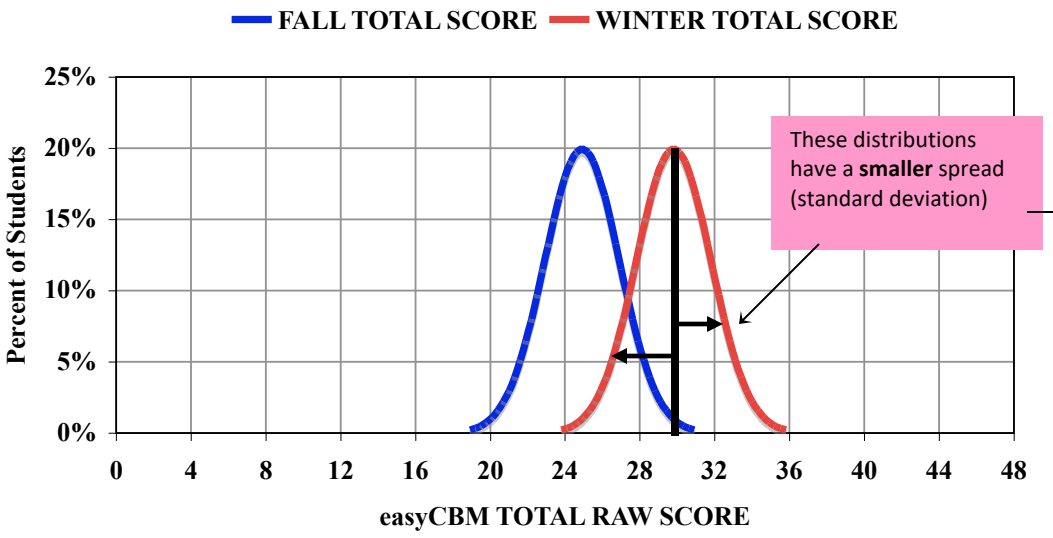
SD	Mean		
	Fall	Spring	Change
8	24.9	29.9	+5.0

$$\frac{\text{Change}}{\text{SD}} = \frac{5}{8} = .625$$

When we consider the mean gain relative to the amount of spread (standard deviation), the more significant gain is one where the mean difference is large relative to the spread of scores.

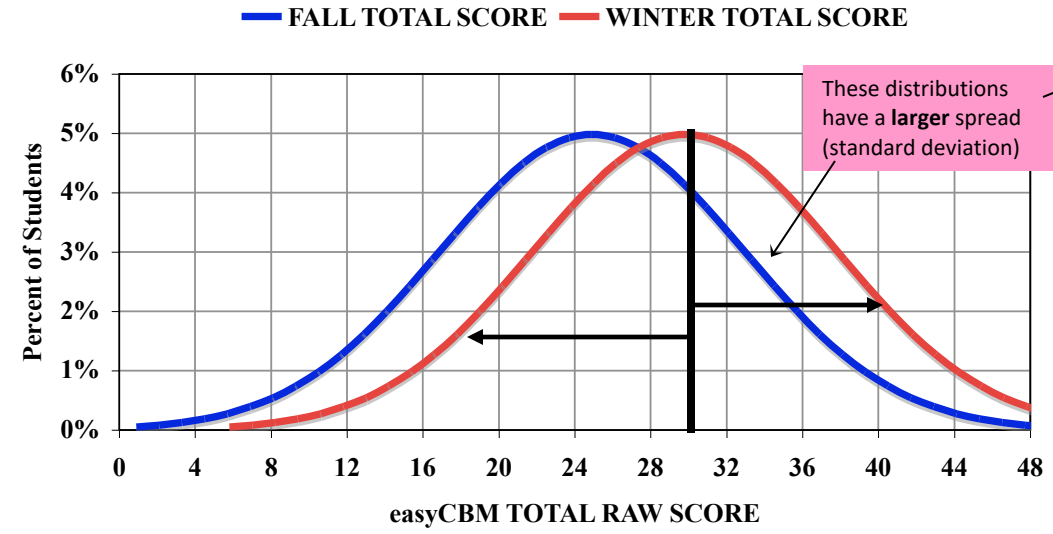
This gain is called an standardized effect or "effect size" and is expressed in standard deviation units. Research (Hattie, 2009) suggests that educators should aim for a gain of at least 0.4 SD.

Which is the more significant growth?



SD	Mean Fall	Mean Spring	Change
3	24.9	29.9	+5.0

$$\frac{\text{Change}}{\text{SD}} = \frac{5}{3} = 1.68$$



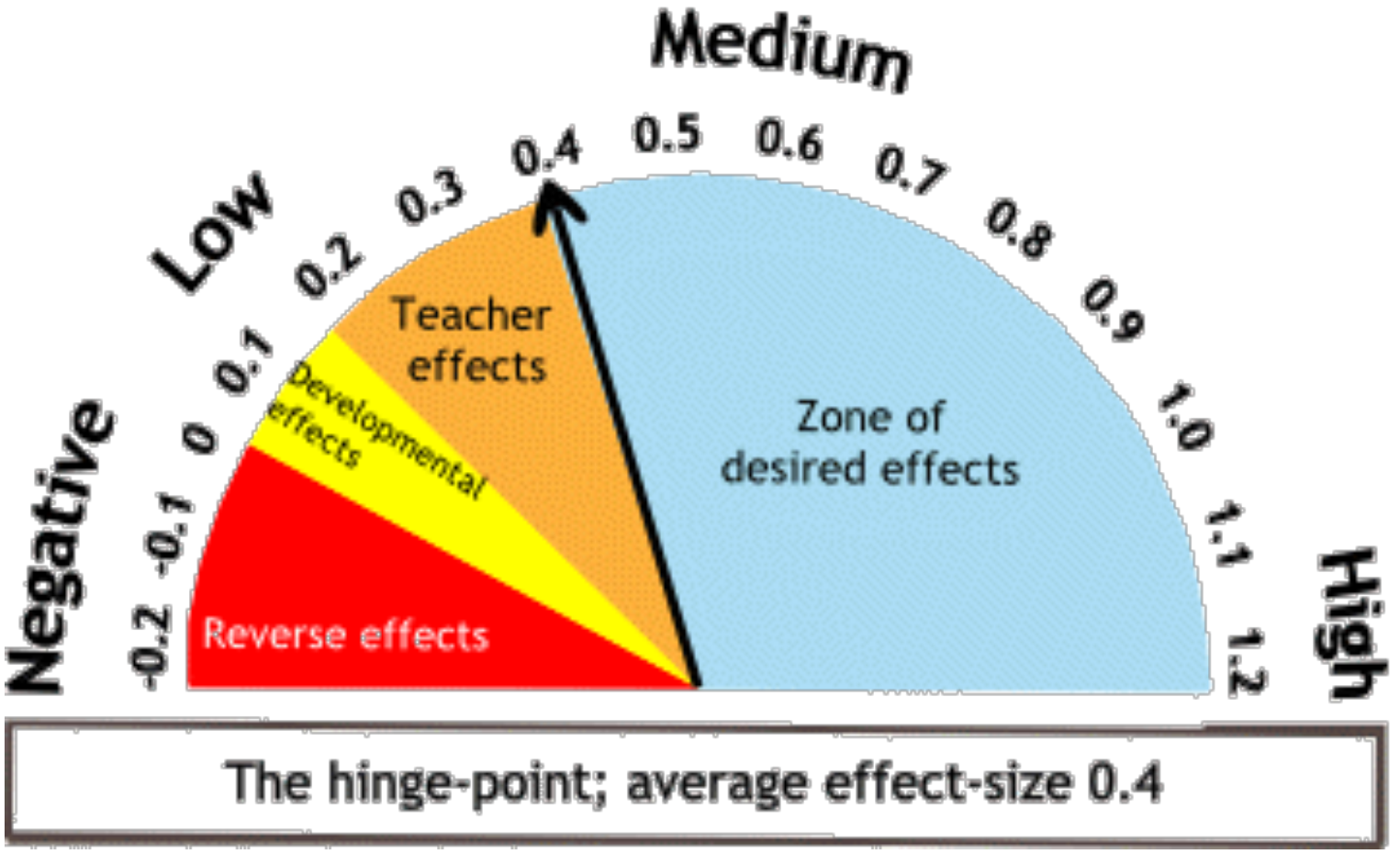
SD	Mean Fall	Mean Spring	Change
8	24.9	29.9	+5.0

$$\frac{\text{Change}}{\text{SD}} = \frac{5}{8} = .625$$

When we consider the mean gain relative to the amount of spread (standard deviation), the more significant gain is one where the mean difference is large relative to the spread of scores.

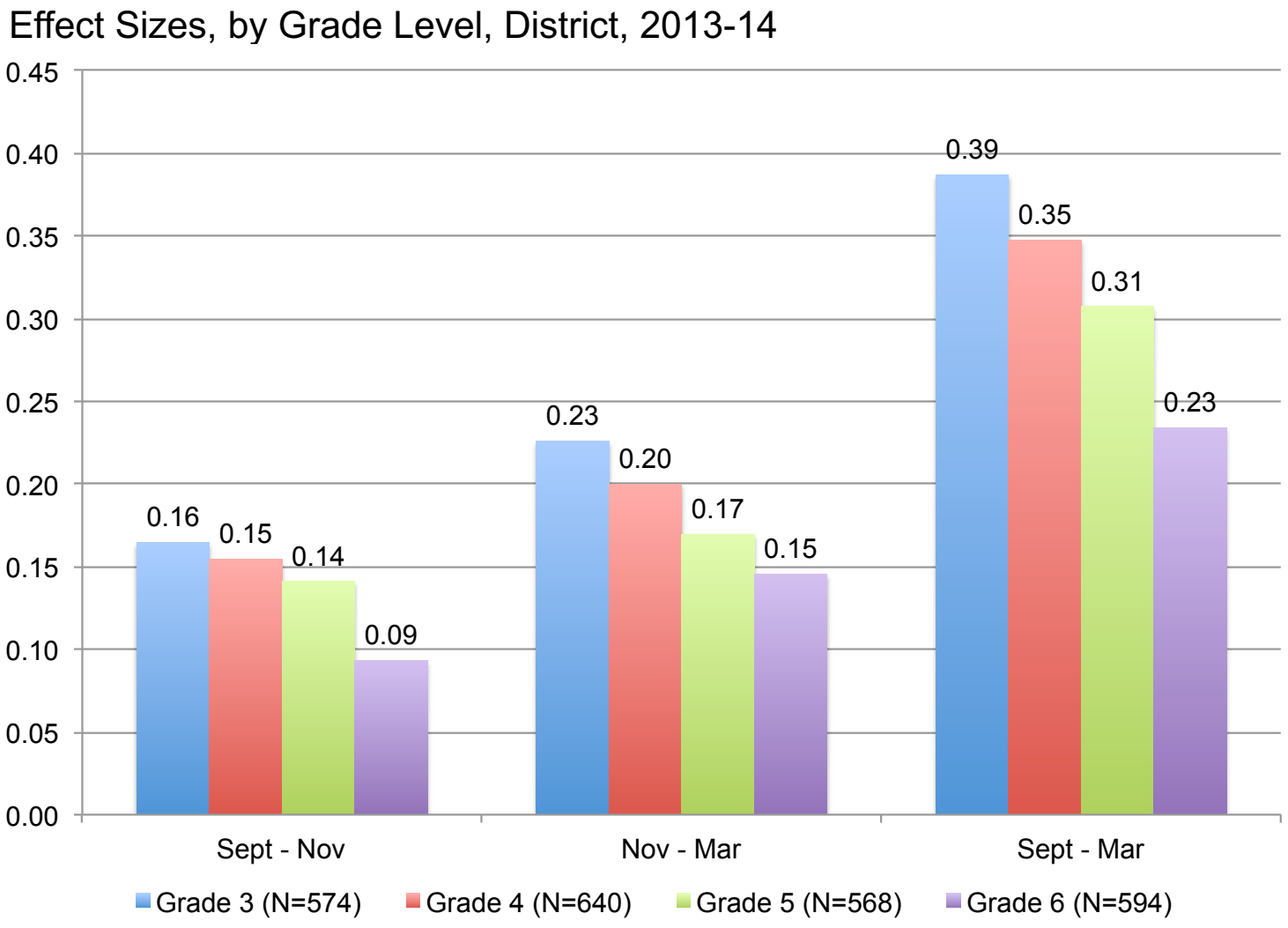
This gain is called an standardized effect or "effect size" and is expressed in standard deviation units. Research (Hattie, 2009) suggests that educators should aim for a gain of at least 0.4 SD.

Research on growth, and the notion of growth as “effect size”



Hattie, J. A. C. (2009). *Visible learning*. New York, NY: Routledge.

What's "normal growth" in Lexile comprehension from September to March?



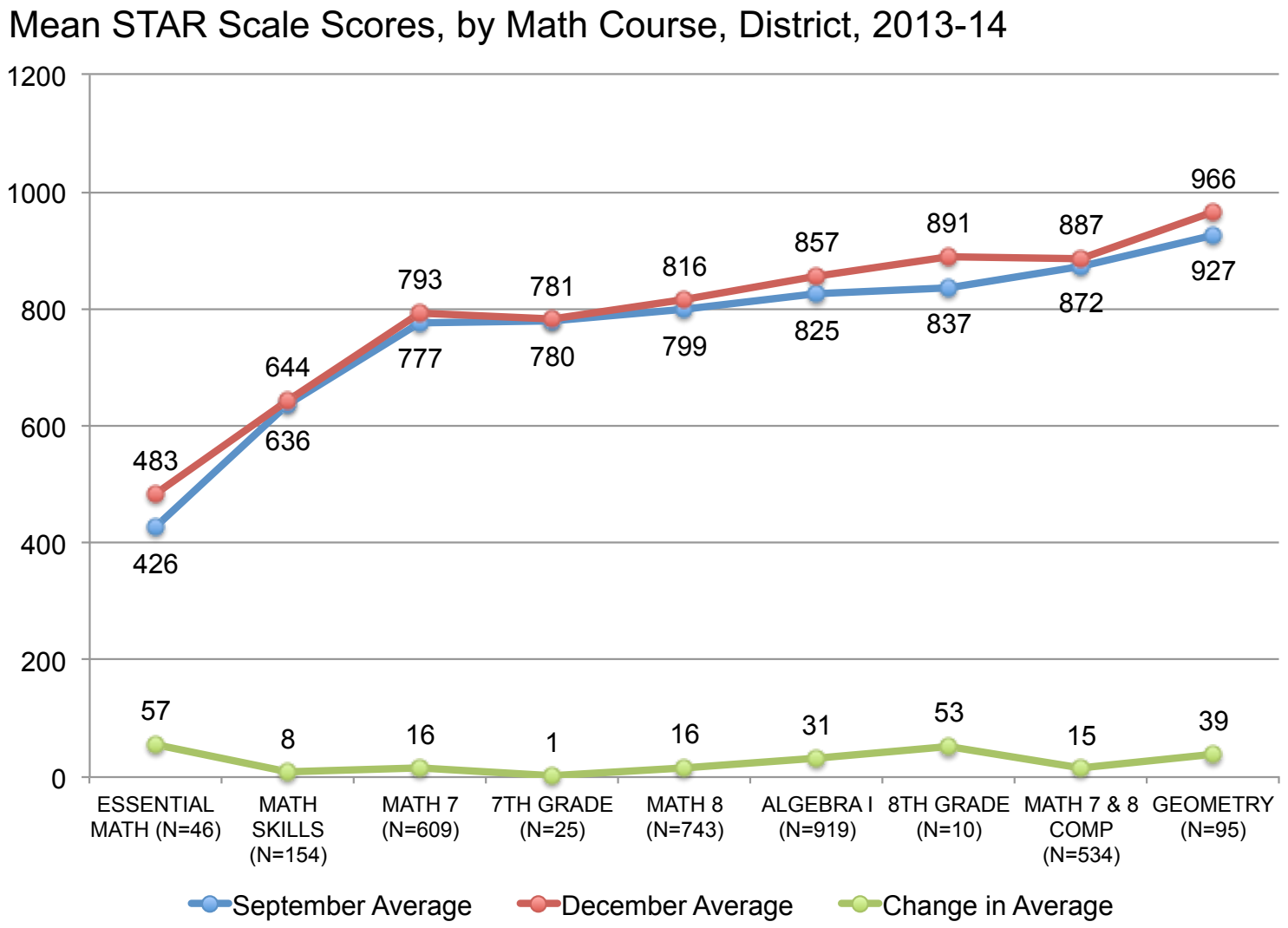
About STAR Math

STAR Math Enterprise CCSS Functional Grade Level Cut Scores Based on an Average of 70% Mastery Across Standards

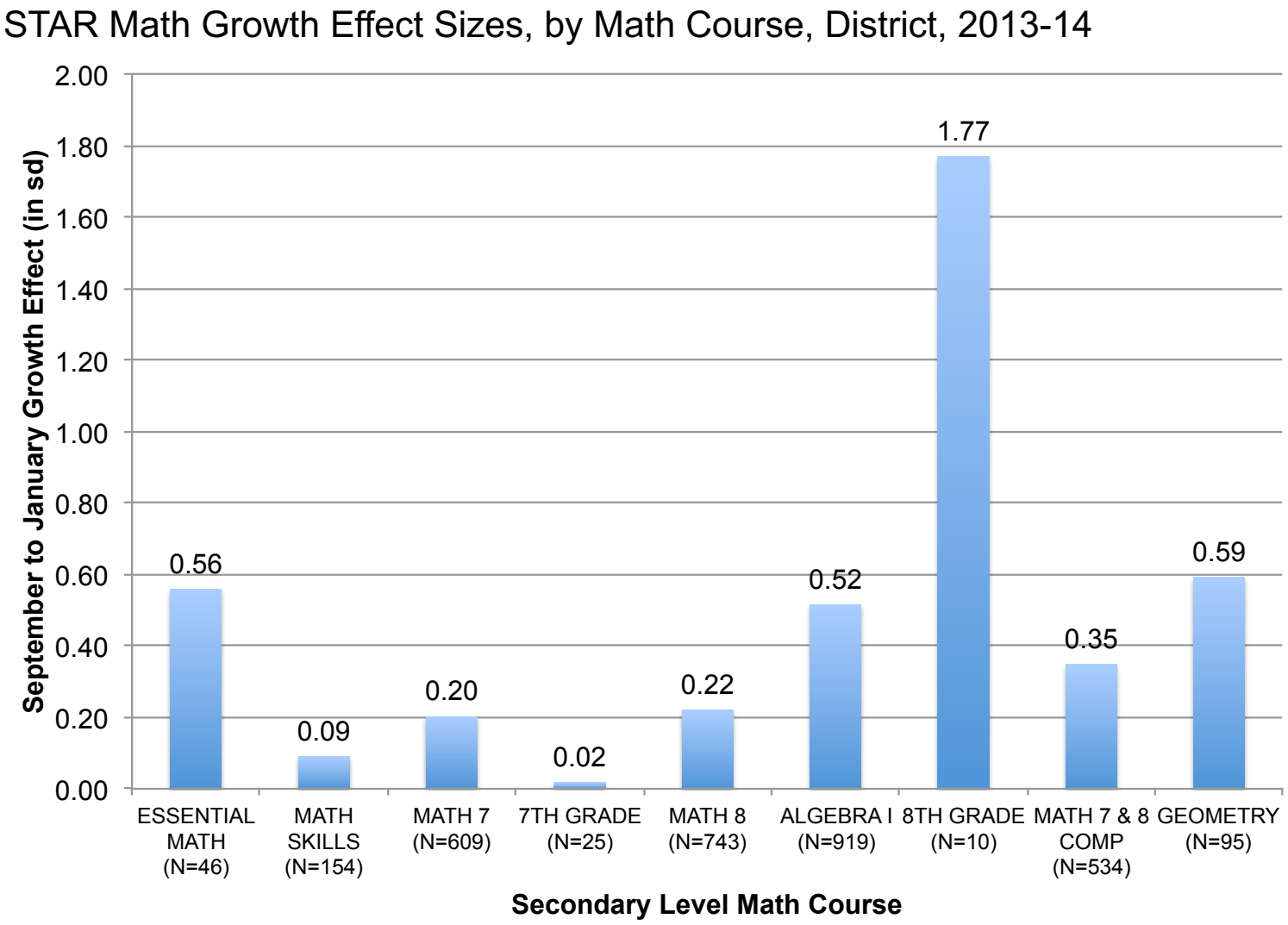
Grade	Below Grade Level		On Grade Level						Above Grade Level	
			On Pace, Fall		On Pace, Winter		On Pace, Spring			
	Low	High	Low	High	Low	High	Low	High	Low	High
K	n/a	n/a	0	127	128	254	255	381	382	1400
1	0	381	382	398	399	414	415	430	431	1400
2	0	430	431	455	456	478	479	502	503	1400
3	0	502	503	544	545	586	587	627	628	1400
4	0	627	628	655	656	683	684	710	711	1400
5	0	710	711	739	740	768	769	796	797	1400
6	0	796	797	813	814	830	831	846	847	1400
7	0	846	847	847	848	848	849	849	850	1400
8	0	849	850	862	863	875	876	888	889	1400
9	0	888	889	892	893	895	896	898	899	1400
10	0	898	899	901	902	902	903	903	904	1400
11	0	903	904	905	906	906	907	907	908	1400
12	0	907	908	909	910	910	911	911	912	1400

SOURCE: STAR Math Technical Manual

What's "normal growth" in STAR Math achievement from September to January?



What STAR math growth (effects) do we see?



How to calculate effect size

Student scores

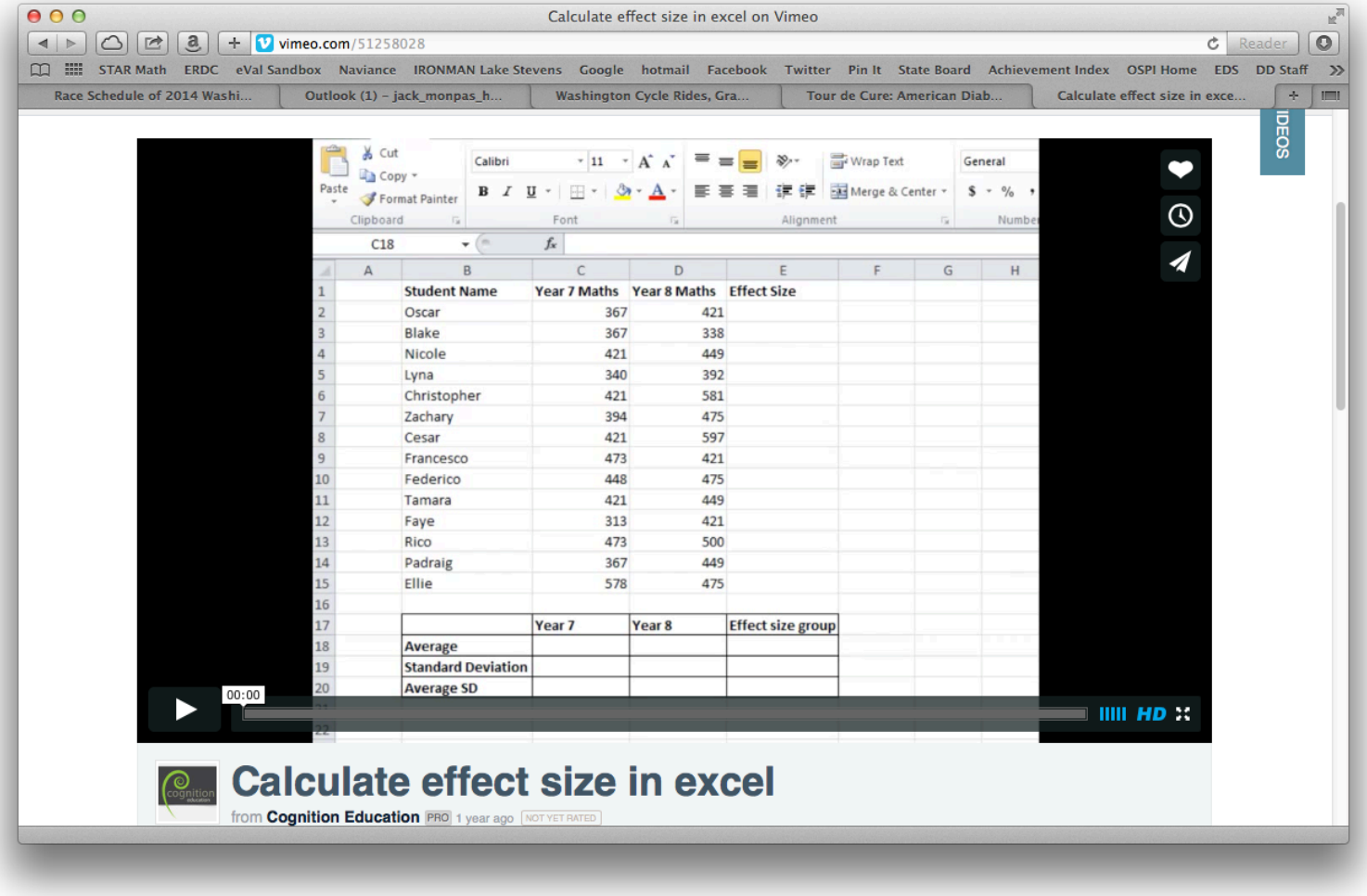
Student	Pretest	Posttest	Change (Post-Pre)
Jack	25	28	3
John	32	33	1
Seth	20	25	5
Kim	35	36	1

Summary statistics

	Pretest	Posttest	Change (Post-Pre)	Effect (Mean Change / SD)
Average (=AVERAGE())	28	31	3	0.37
SD (=STDEV())	7	5		

How teachers can calculate effect size

http://vimeo.com/51258028



Advantages of effect size

1. It offers a sense of low, medium and high
2. It does not depend on sample size
3. It offers a growth metric that is standard across tests and content areas
4. It takes full advantage of our common district assessments designed with scale scores
5. It gives credit for growth all along the scale, not just a jump over a proficiency bar
6. It enables us to compare our work to research on interventions in the broader profession
7. It is easy to calculate in Excel or Google Spreadsheet with student scores
8. Teachers / PLCs can use it with common pre- and post assessments
9. We can use it to look at group gaps as well as growth

Challenges of effect size

1. It's statistical. It's technical. Requires some training.
2. It ignores proficiency benchmarks.

Going forward

1. Propose that we add effect size to our box of tools for looking at growth
2. Director of Assessment can provide data and/or training
    - Working with assessment data in Excel
    - Pre-post gain scores
    - Effect sizes