Data Visualization using R

Damian W. Betebenner

National Center for the Improvement of Educational Assessment
Dover, NH

NERA Training Session
Trumbull, Connecticut October 27th, 2016

Purpose of today's session

- To learn about R and its capabilities for visualizing data.
- To provide participants with a comprehensive list of resources for producing their own data visualizations.
- To demonstrate the power of programatic drawing over WYSIWYG drawing.
- To inspire participants to develop their own customized visualizations and push the envelope for what is possible

About Me

Dr. Damian Betebenner, PhD

- Senior Associate at the Center for Assessment (NCIEA).
- Developed student growth percentiles and percentile growth trajectories to help states and educational associations employ student growth in decision making [Betebenner, 2008, Betebenner, 2009].
- In the process of refining and sharing these techniques with other states including Colorado, Massachusetts, Arizona, Indiana, and 15 other states in various stages of investigation/adoption.
- Interested in the rise of the data sciences/scientist: Data analysis and data visualization and their use within education. Began using R in 1998.

About the R Software

- R is an GNU open source, free, statistical software environment (package/language) that is available for source compilation or in pre-compiled binary form for numerous operating systems.
- R can be downloaded from CRAN (Comprehensive R Archive Network) http://cran.r-project.org/
- The R language has become a *de facto* standard among statisticians for the development of statistical software.
- The philosophy behind R/S (John Chambers): "To Turn Ideas into Software Quickly and Faithfully" [Chambers, 2000, p. v]

About the R Software

Daryl Pregibon, Google

R is really important to the point that it's hard to overvalue it.

- R had a recent New York Times article written about it http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html.
- A particular strength of R is its data visualization capabilities
- The greatest strength of R (in my humble opinion) is the IMMENSE amount of code available online to learn from. Learn from the masters.
- This training session will introduce users to data visualization using R as well as the resources available to continue their explorations.

About Data

Rutherford D. Roger

We are drowning in information [data] and starving for knowledge [information].

- The Economist, February 27th, 2010, devoted a special issue to "The Data Deluge".
- In 2008, 1,200 exabytes (1.2 ZB) of digital data was created (n.b., giga, tera, peta, exa, zetta, yotta).
- Compound annual increase of data at 60%.
- Petabyte level computing is reaching commodity levels with Amazon EC2/S3 offering specials for petabyte scale projects.

About Data

Hal Varian, Chief Economist, Google

A new kind of professional has emerged, the data scientist, who combines the skills of a software programmer, statistician, and story-teller/artist to extract nuggets of gold hidden under mountains of data. The job of the data scientist will become the sexiest around. Data are widely available; what is scarce is the ability to extract wisdom from them.

- The Data doesn't speak for itself. The data can tell a thousand stories.
- Transforming data into information and ultimately into knowledge requires a broad range of expertise (i.e., the emerging data scientist).
- Goal: The right data to the right people at the right time in the right format!
- To reach this goal subject matter experts MUST work more closely with IT specialists, ensuring the right stories are told.

About Data Visualization

- Data visualization leverages the immense capacity of the human eye to extract patterns from visual stimuli.
- Communicating with data is a form of storytelling. Visualization allows the storyteller to communicate more effectively and engage the user in more complicated stories.
- Color, pattern, and symbols can be combined to communicate highly complicated stories that, without pictures, are difficult even for experts to follow.
- As such, data visualization, as communication, is a form of teaching.

About Programmatic Drawing

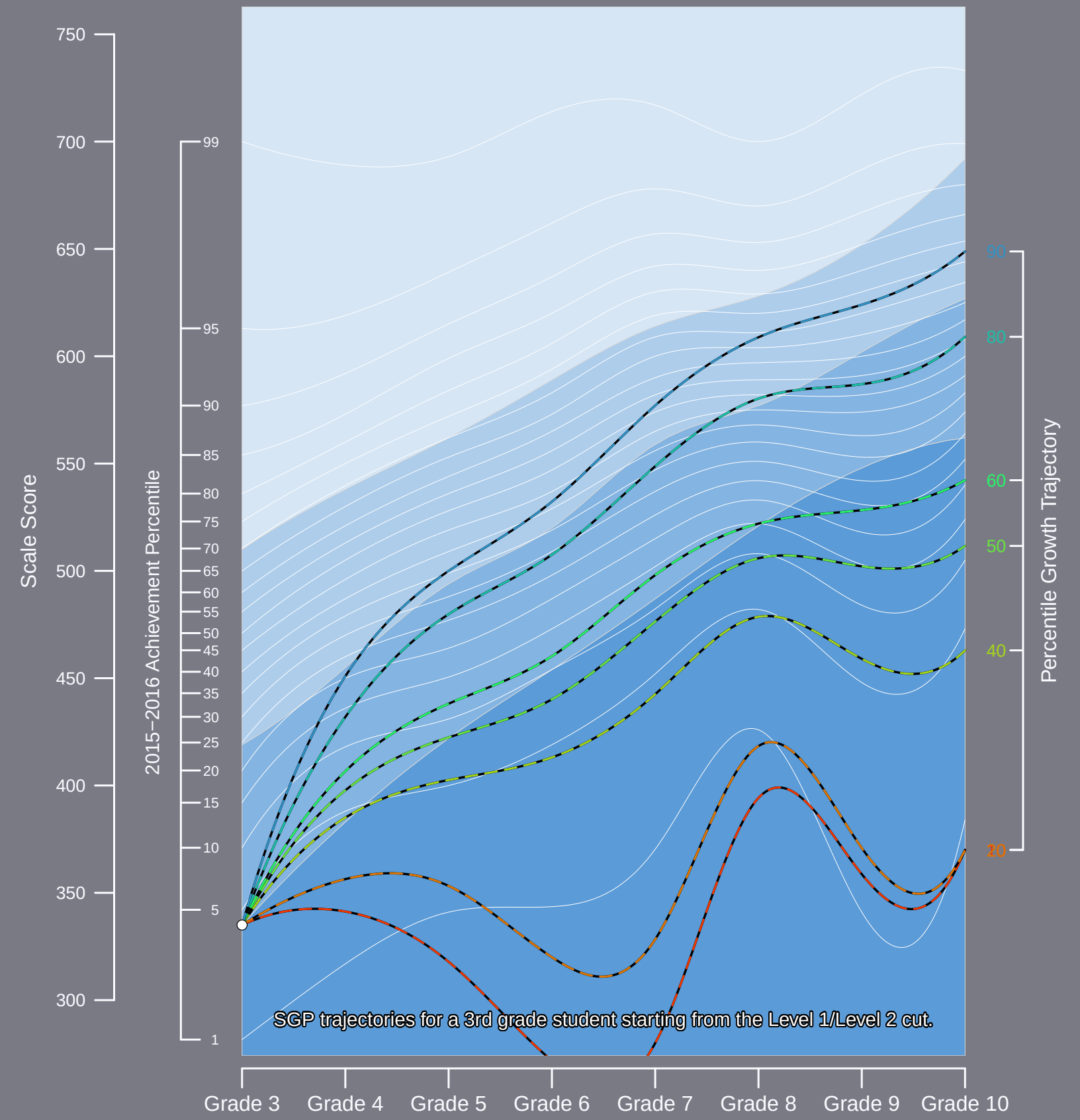
Byron Ellis via Twitter

[R] is for making new things. Point and click is for redoing old things.

- Almost all graphical programs (especially those incorporated in statistical analysis packages) rely upon WYSIWYG interface.
- WYSIWYG pros: Easy to learn. WYSIWYG cons: Limited ability to customize.
- Because of constrained options, WYSIWYG interfaces both dictate and limit the stories you can tell with your data.
- Programming your drawings (i.e., programatic drawing) allows for unlimited possibilities, constrained only by the creator's imagination.

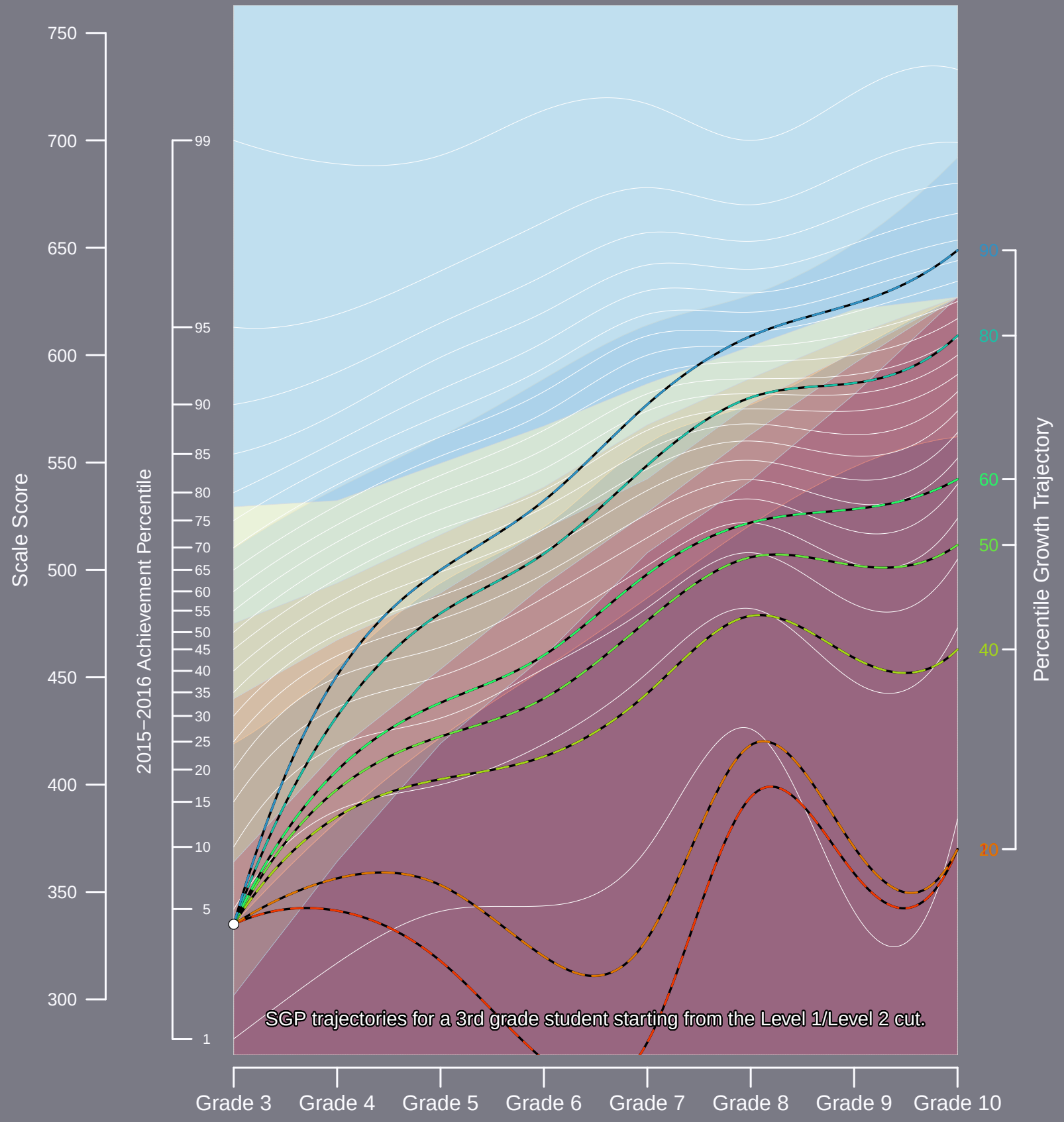
DEMONSTRATION: 2015−2016 Mathematic

Norm & Criterion Referenced Growth & Achievement

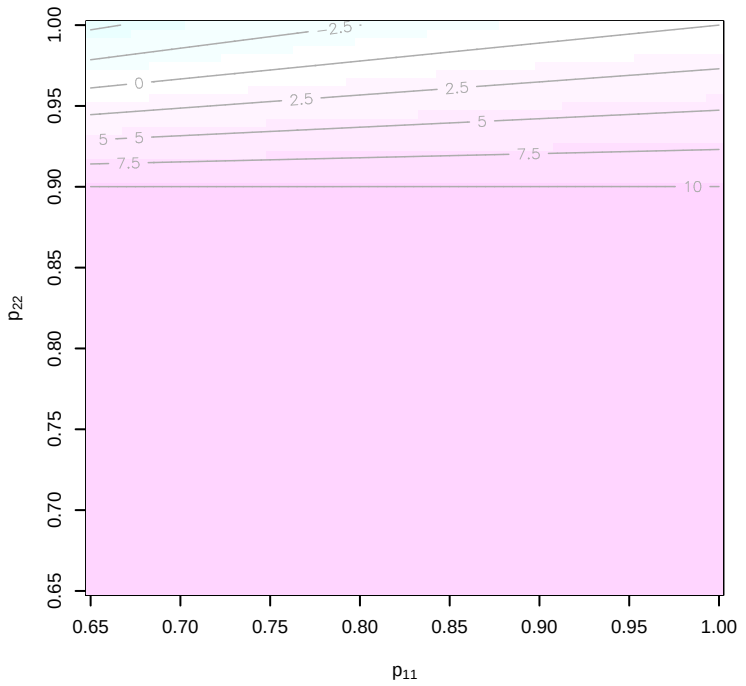


DEMONSTRATION: 2015–2016 Mathematic

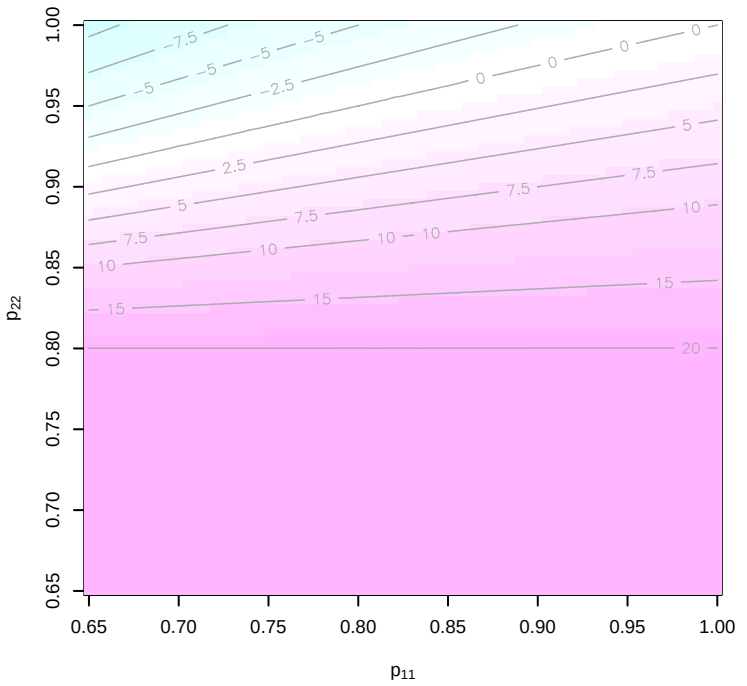
Norm & Criterion Referenced Growth & Achievement



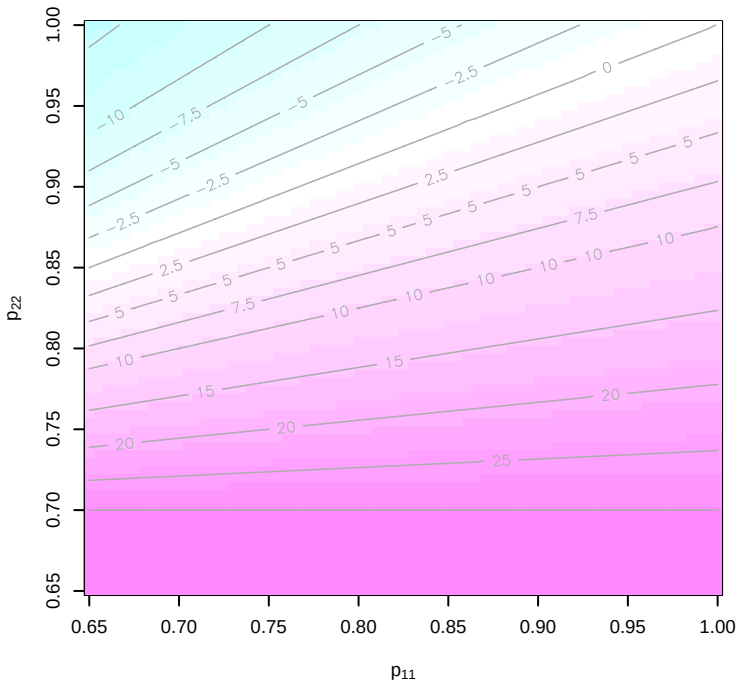
Observed Percent Proficient = 10 %



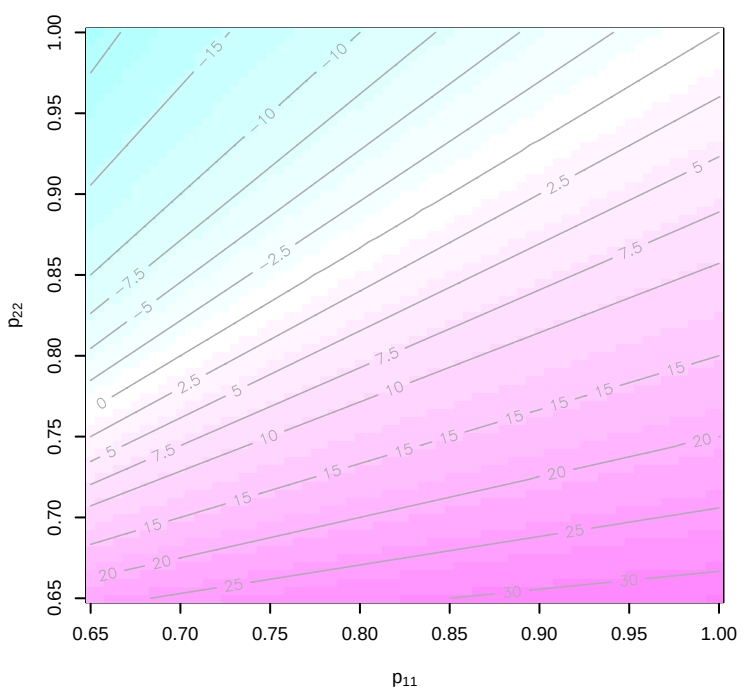
Observed Percent Proficient = 20 %



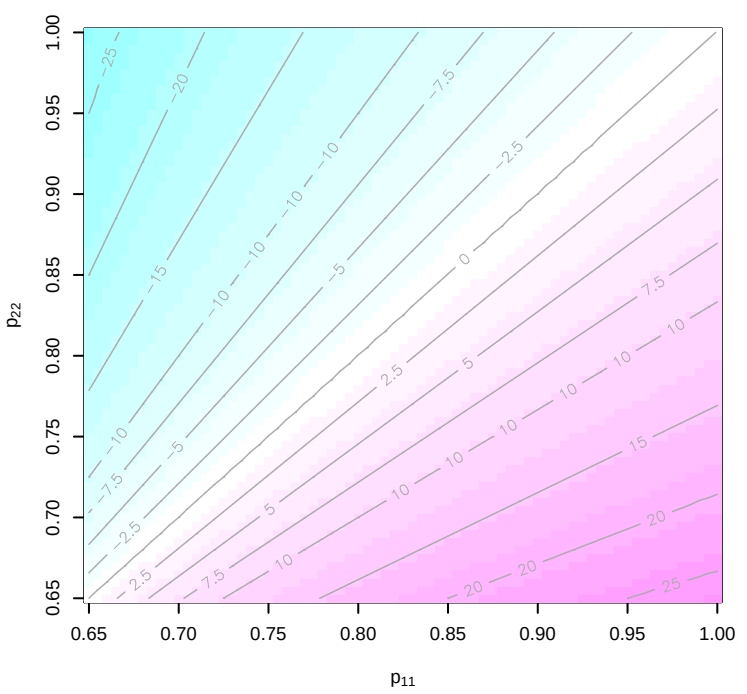
Observed Percent Proficient = 30 %



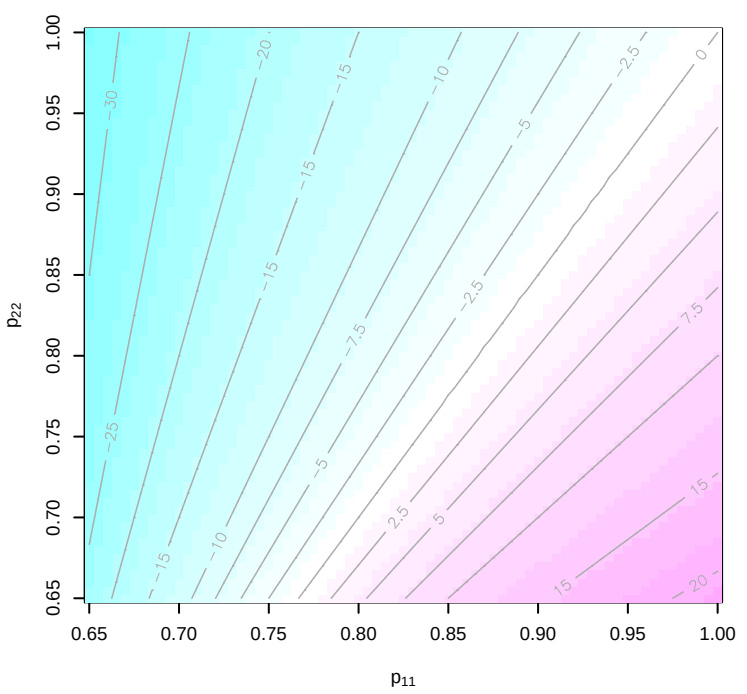
Observed Percent Proficient = 40 %



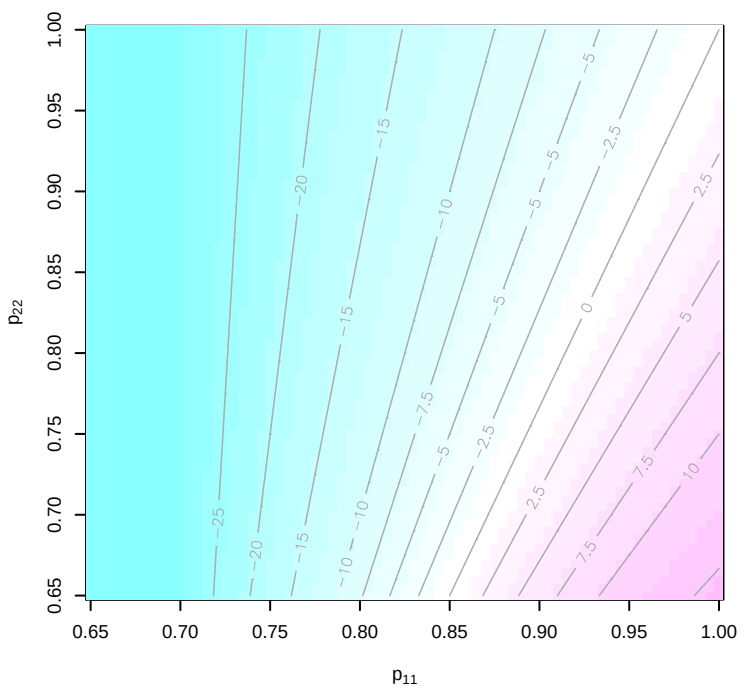
Observed Percent Proficient = 50 %



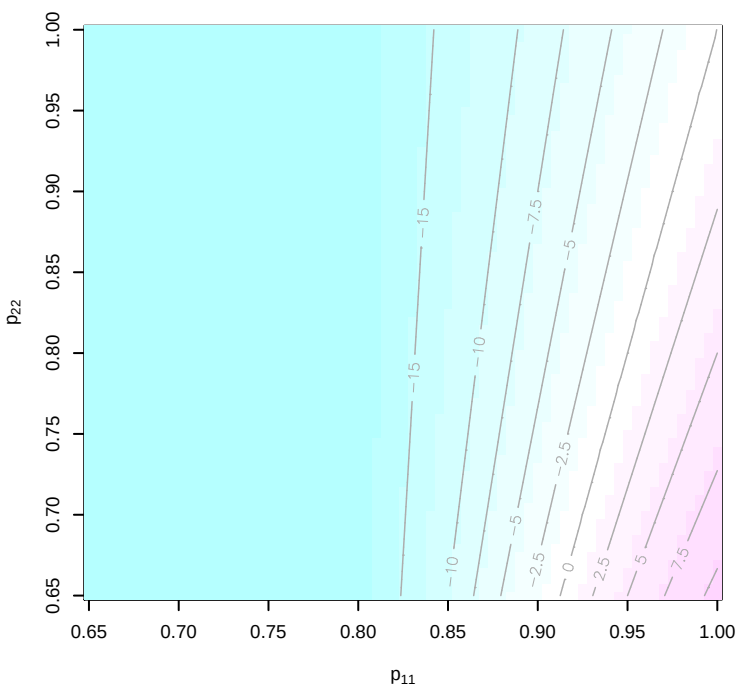
Observed Percent Proficient = 60 %



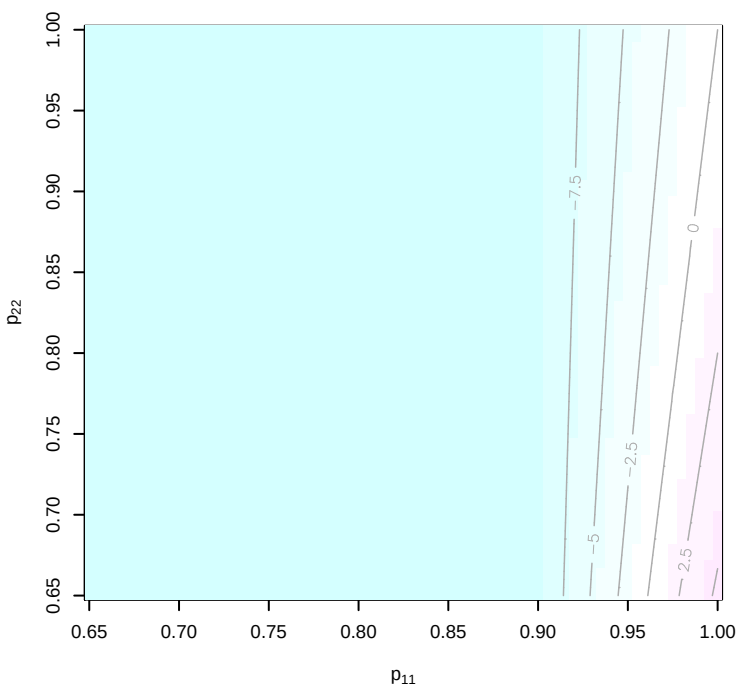
Observed Percent Proficient = 70 %



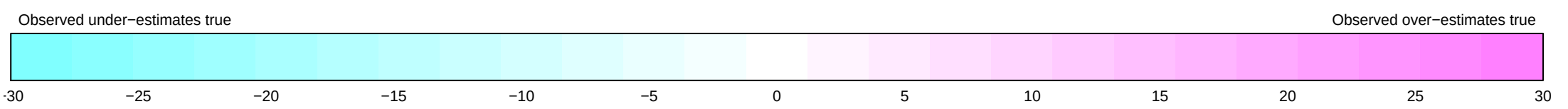
Observed Percent Proficient = 80 %

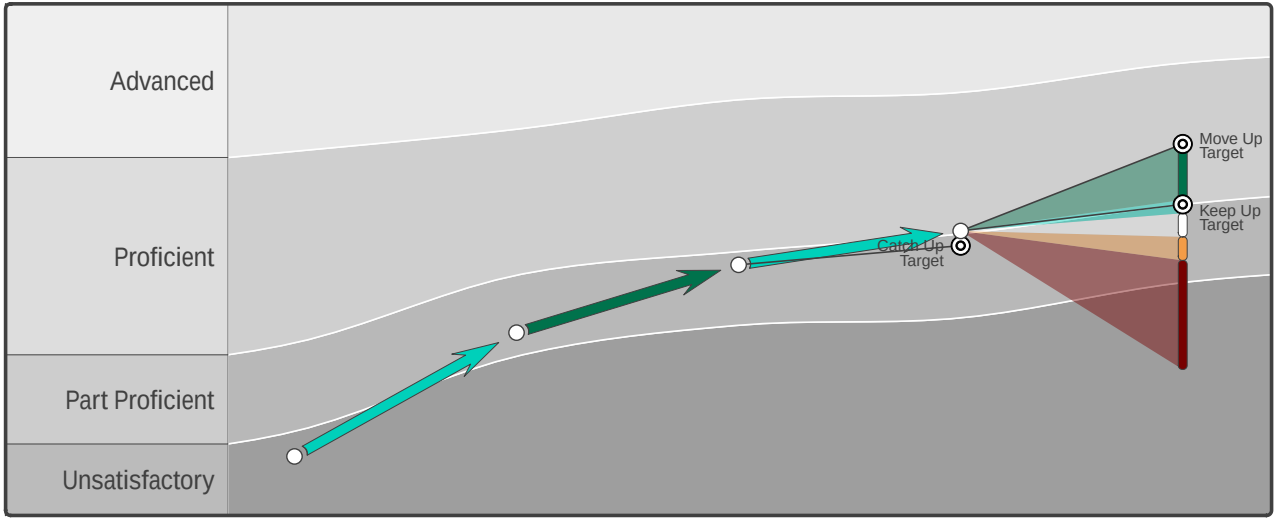


Observed Percent Proficient = 90 %



Bias in Observed Percent Proficient (Observed − True)





Reading

Achievement

DEMO Reading Scale Score

Growth

Level	Percentiles
Very High	81st – 99th
High	61st – 80th
Typical	40th – 60th
Low	20th – 39th
Very Low	1st – 19th

	Grade 3 2009–2010	Grade 4 2010–2011	Grade 5 2011–2012	Grade 6 2012–2013	Grade 7 2013–2014	
Scale Score	449	533	579	602		Achievement
Achievement Level	Unsatisfactory	Part Proficient	Part Proficient	Proficient		
Achievement Target				Proficient (592)	Proficient (620)/Advanced (661)	
Growth Percentile		69	85	74		Growth
Growth Level		High	Very High	High		
Growth Target				Catch Up (60)	Keep Up (76)/Move Up (99)	

Billy Gonzales

North End Middle School

How to interpret this student growth & achievement report

DEMO Scale Score

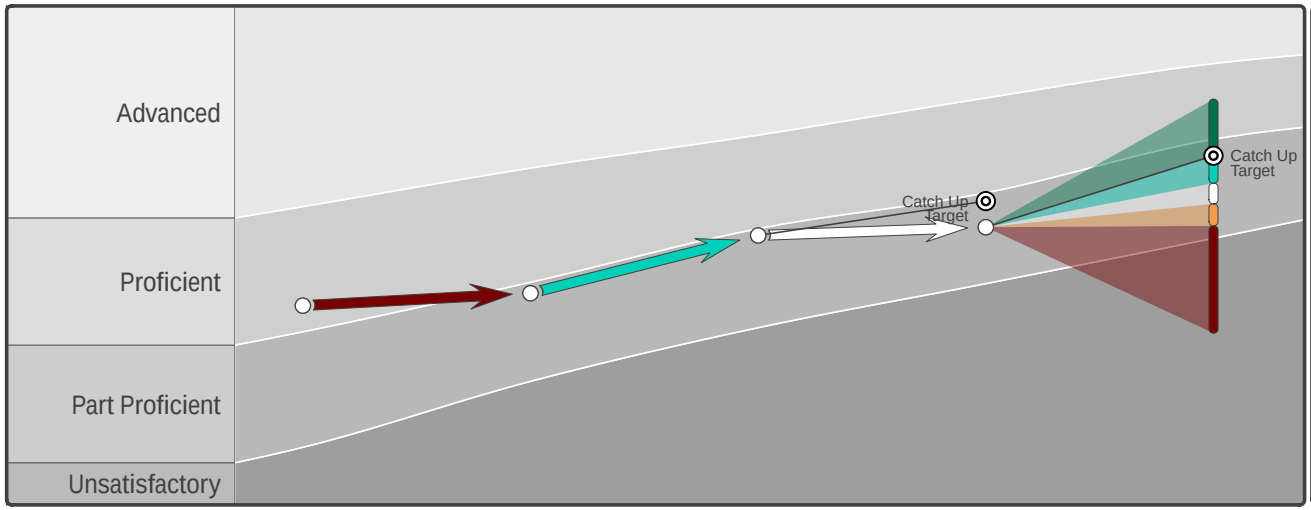
DEMO Achievement Levels

Student Growth Percentile

Catch Up/Keep Up
Move Up/Stay Up
Targets

Suggested Uses

- Review past growth to assess student academic progress toward DEMO achievement goals.
- Develop remediation or enrich−ment plans based on rate of growth needed to reach higher DEMO achievement levels.
- Identify the rate of progress needed in order to reach or maintain proficient status on the DEMO next year.



Math

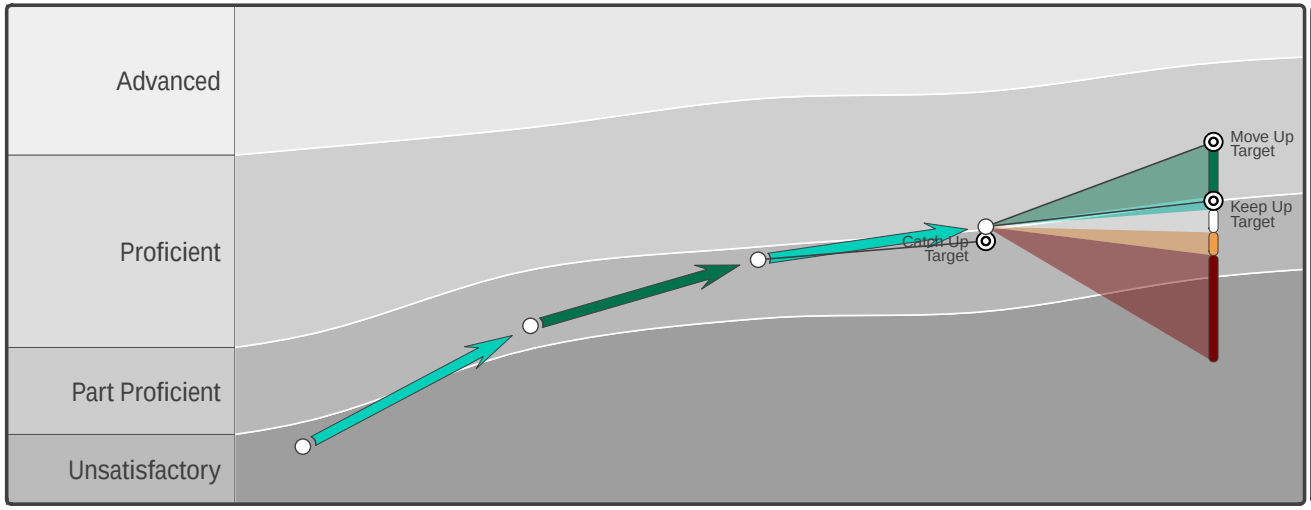
Achievement

DEMO Math Scale Score

Growth

Level	Percentiles
Very High	81st − 99th
High	61st − 80th
Typical	40th − 60th
Low	20th − 39th
Very Low	1st − 19th

	Grade 3 2009−2010	Grade 4 2010−2011	Grade 5 2011−2012	Grade 6 2012−2013	Grade 7 2013−2014	
Scale Score Achievement Level Achievement Target	438 Proficient	447 Part Proficient	489 Part Proficient	495 Part Proficient Proficient (514)	Proficient (547)	Achievement
Growth Percentile Growth Level Growth Target		18 Very Low	64 High	44 Typical Catch Up (70)	Catch Up (82)	Growth



Reading

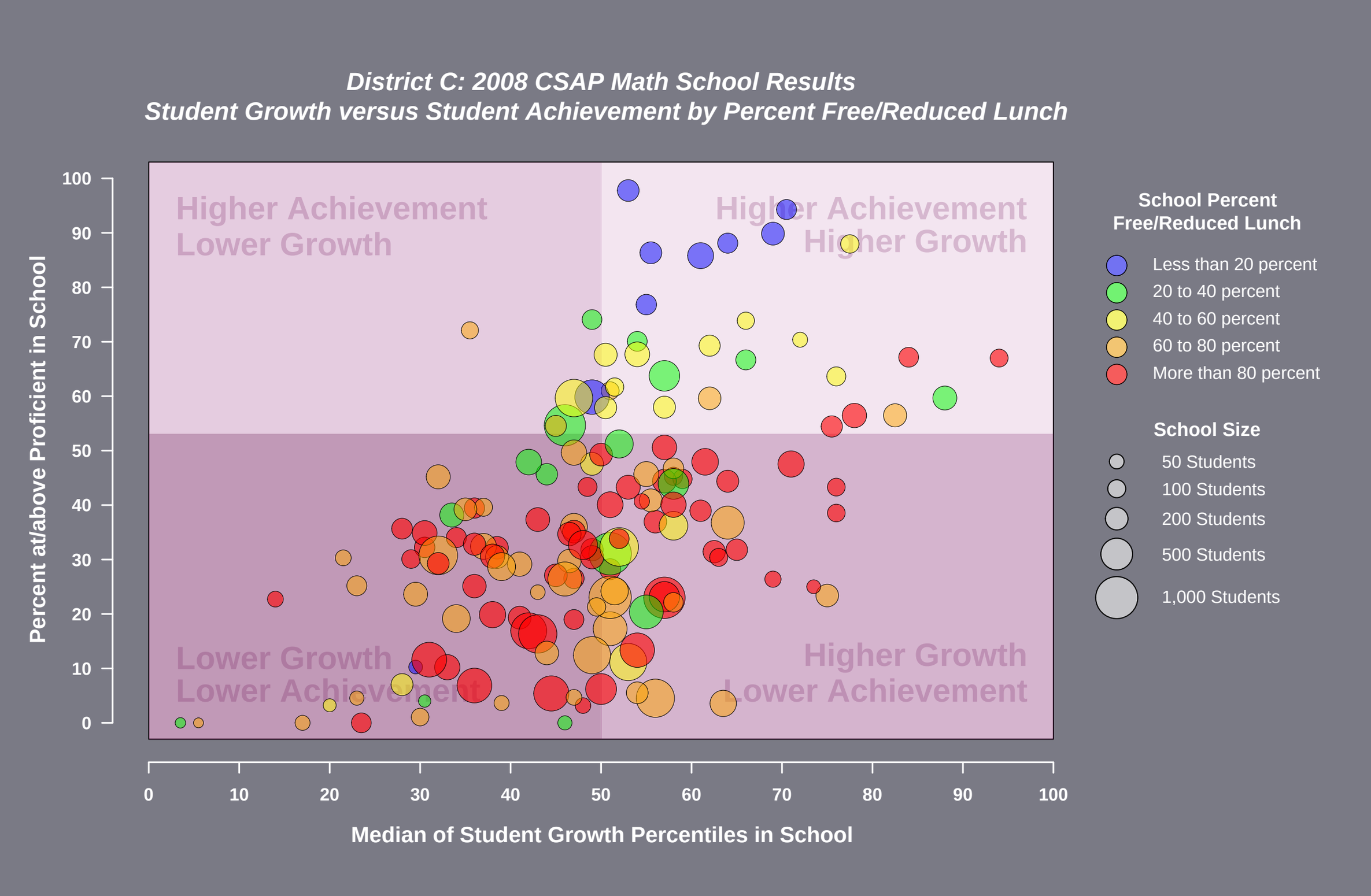
Achievement

DEMO Reading Scale Score

Growth

Level	Percentiles
Very High	81st − 99th
High	61st − 80th
Typical	40th − 60th
Low	20th − 39th
Very Low	1st − 19th

	Grade 3 2009−2010	Grade 4 2010−2011	Grade 5 2011−2012	Grade 6 2012−2013	Grade 7 2013−2014	
Scale Score Achievement Level Achievement Target	449 Unsatisfactory	533 Part Proficient	579 Part Proficient	602 Proficient Proficient (592)	Proficient (620)/Advanced (661)	Achievement
Growth Percentile Growth Level Growth Target		69 High	85 Very High	74 High Catch Up (60)	Keep Up (76)/Move Up (99)	Growth


District C: 2008 CSAP Math School Results
Student Growth versus Student Achievement by Percent Free/Reduced Lunch
Higher Achievement
Lower Growth
Higher Achievement
Higher Growth
Lower Growth
Lower Achievement
Higher Growth
Lower Achievement
Percent at/above Proficient in School
0
10
20
30
40
50
60
70
80
90
100
Median of Student Growth Percentiles in School
0
10
20
30
40
50
60
70
80
90
100
School Percent Free/Reduced Lunch
Less than 20 percent
20 to 40 percent
40 to 60 percent
60 to 80 percent
More than 80 percent
School Size
50 Students
100 Students
200 Students
500 Students
1,000 Students

So Many Amazing Resources: Free

- Numerous help manuals installed when R is installed (Under the Help dropdown menu in Windows)
- One can download R and numerous resources at the CRAN (Comprehensive R Archive Network) http://cran.r-project.org/.
- Click on the "Contributed" hyperlink at CRAN to access dozens of free resources in many languages.
- Most resources provide many nuggets of wisdom that proves to be useful. Good places to start include:
  - John Maindonald's *Using R for Data Analysis and Graphics–Introduction, Examples and Commentary*
  - John Verzani's *Simple R*
- R has a very active listserve for R (R-help) that is easy to search using Google and keywords.

So Many Amazing Resources: Free

- The R Graph Gallery. Great examples with source code: http://addictedtor.free.fr/graphiques/.
- Spatial Data and R. Great examples with source code http://r-spatial.sourceforge.net/.
- A four part online video tutorial for using ggplot http://blog.revolution-computing.com/2010/03/video-hadley-wickham-gives-a-short-course-on-graphics-with-html
- Revolution Computing (a company that produces a version of R that cost $$$) has a blog that is REALLY cool: http://blog.revolution-computing.com
- Data Mashups using R. An amazing tutorial ($4.99 from O'Reilly).

So Many Amazing Resources: Published Books

- The Grid Graphics Package (and traditional graphics overview) [Murrell, 2006]
- Lattice Graphics Package: Depyan Sarkar [Sarkar, 2008].
- ggplot Graphics Package: Hadley Wickham [Wickham, 2009].
- Spatial Data Analysis and Visualization [Bivand et al., 2008].

Inspiration

"A Thing of Beauty is a Joy Forever"

John Keats

References

Betebenner, D. W. (2008).
Toward a normative understanding of student growth.
In Ryan, K. E. and Shepard, L. A., editors, *The Future of Test-Based Educational Accountability*, pages 155–170. Taylor & Francis, New York.

Betebenner, D. W. (2009).
Norm- and criterion-referenced student growth.
Educational Measurement: Issues and Practice, 28(4):42–51.

Bivand, R. S., Pebesma, E. J., and Gòmez-Rubio, V. (2008).
Applied Spatial Data Analysis with R.
Springer, New York.

Chambers, J. M. (2000).
Programming with Data.
Springer, New York.

Murrell, P. (2006).
R Graphics.
Springer, New York.

Sarkar, D. (2008).
Lattice: Multivariate Data Visualization with R.
Springer, New York.

Wickham, H. (2009).
ggplot2: Elegant Graphics for Data Analysis.
Springer, New York.